Information-Directed Offline-to-Online Reinforcement Learning

Keru Chen
School of Electrical, Computer and Energy Engineering, Arizona State University
kchen234@asu.edu

Abstract

Decision-making from offline datasets typically warm-starts a policy or score model from fixed offline data and then refines it with limited online interaction. Offline data reduces uncertainty, but it does not remove the need for exploration; it changes what remains to be explored. We formalise this residual uncertainty by the conditional mutual information $I(\chi; \tau_{1:T} \mid \mathcal{D}_N)$ between a learning target χ and the online trajectories after conditioning on the offline dataset. This view leads naturally to information-directed sampling (IDS), a family parameterised by $\eta \geq 0$ that selects actions by trading off instantaneous regret against information gain. We prove a generic offline-to-online Bayesian regret bound for IDS through a ratio certificate: any information-ratio bound satisfied by a reference Thompson-sampling policy over the same randomised policy class is inherited by IDS. In a known-dynamics Bayesian linear-reward model, the conditional mutual information has a log-determinant form, and vanilla IDS ($\eta = 0$) satisfies $\widetilde{O}\left(Hd \min\left\{\sqrt{T},\, T\sqrt{C^{\dagger}_{\beta,\mathrm{IDS}_0}(N,T)/N}\right\}\right)$, where the coverage coefficient is tied to the visitation distribution induced by vanilla IDS itself. We also identify a warm-start regime with a dominated but informative probe in which vanilla IDS selects the probe while Thompson sampling never does, giving a constant-factor Bayesian regret separation. Controlled bandit experiments and D4RL offline-to-online RL experiments validate this mechanism: IDS is most beneficial when offline data is informative but leaves biased or low-probability residual uncertainty that targeted online actions can resolve, a regime shared by offline RL, offline black-box optimization, and Bayesian optimization.

1 Introduction

Modern data-driven decision-making systems rarely learn entirely online. Whether the deployed pipeline is offline reinforcement learning, offline model-based black-box optimization, Bayesian optimization, or a contextual bandit, a model or policy is first warm-started from a fixed offline dataset and then adapted with a limited amount of online interaction [Levine et al., 2020, Kumar et al., 2020, Kostrikov et al., 2022, Nair et al., 2020, Nakamoto et al., 2023, Yao et al., 2025b, Lee et al., 2022, Li et al., 2023, Yao et al., 2026, Ball et al., 2023, Zhao et al., 2025, Liu et al., 2026]. This offline-to-online pipeline is appealing because offline data makes online learning safer and cheaper [Chen et al., 2024, Yao et al., 2025a]. However, offline data does not remove the need for exploration. It only changes what remains to be explored. After conditioning on the offline dataset, the learner no longer needs to explore broadly over the original state-action (or design) space, but must instead resolve the residual uncertainty left by the warm start. We focus on the offline-to-online RL setting, but the principle is community-agnostic.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

Most existing offline-to-online methods treat the two phases through different principles. Offline algorithms are conservative: they avoid actions that are poorly covered by the behaviour policy because extrapolation error is the dominant failure mode [Fujimoto et al., 2019, Kumar et al., 2020, Jin et al., 2021, Xie et al., 2021, Rashidinejad et al., 2021, Uehara and Sun, 2022]. Online algorithms are exploratory: they deliberately visit uncertain parts of the state-action space to improve future decisions [Abbasi-Yadkori et al., 2011, Osband et al., 2013, Jin et al., 2020]. The transition between these principles is usually handled by algorithmic devices such as replay balancing, policy constraints, calibrated pessimism, or exploration bonuses [Nakamoto et al., 2023, Li et al., 2023, Lee et al., 2022, Song et al., 2023]. These methods can work well, but they do not isolate the basic statistical question: after observing $\mathcal{D}_N$, which online actions are worth taking because they cheaply resolve the residual uncertainty left by the warm start?

We study offline-to-online learning from this residual-uncertainty perspective. Let χ denote a learning target, such as an unknown value parameter, model parameter, or latent task mode. The conditional mutual information $I(\chi; \tau_{1:T} \mid \mathcal{D}_N)$ measures how much information the online trajectories can still extract about χ after the offline dataset has already been observed. This quantity is not a penalty for poor offline coverage; it is the information still missing after the posterior has been conditioned on $\mathcal{D}_N$. The central claim of this paper is that offline-to-online exploration should be directed at this residual information rather than at uncertainty under the original prior.

This view points to the information-directed sampling family [Russo and Van Roy, 2016, 2018], which we write as IDS_η for the regularised rule and IDS_0 for vanilla IDS. The family evaluates actions by a regret-information tradeoff

$$\Psi^\eta = \Delta^2/(g+\eta),$$

where Δ is instantaneous expected regret and g is information gain. Thompson sampling also admits information-ratio analyses [Thompson, 1933, Osband et al., 2013, Russo and Van Roy, 2014], but it explores implicitly by probability matching: it samples a model from the posterior and acts optimally for that model. After a strong offline warm start, this distinction matters. After offline warm start, the posterior often concentrates on a dominant mode while leaving low-probability but consequential modes unresolved. Thompson sampling tests that mode only when it samples it. The IDS family can instead choose a deliberately suboptimal diagnostic action if that action resolves the remaining uncertainty at sufficiently low regret.

Our Bayesian viewpoint is closest in spirit to Hu et al. [2024], who use probability matching to navigate the pessimism-versus-optimism dilemma in off-to-on fine-tuning. We share the residual-posterior view but study an explicit regret–information tradeoff: a diagnostic action can be informative even when it is not optimal under any sampled model, so it is invisible to probability matching but selectable by IDS.

We make three contributions. (i) A generic off-to-on Bayesian regret bound for IDS_η via a ratio certificate: a reference Thompson policy is used only as a witness, and the conditional mutual-information chain rule converts the sum of online information gains into the single residual-information term $I(\chi; \tau_{1:T} \mid \mathcal{D}_N)$. (ii) A known-dynamics linear-reward instantiation in which residual information has a closed-form log-determinant, yielding

$$\mathrm{BayesRegret}_T^{\mathrm{off}}(\pi^{\mathrm{IDS}_0,w}) \le \widetilde{O}\left(Hd \min\left\{ \sqrt{T},\, T\sqrt{C^\dagger_{\beta,\mathrm{IDS}_0}(N,T)/N} \right\} \right),$$

where the warm-start branch tracks a coverage coefficient pinned to the visitation distribution induced by IDS_0 itself, rather than a static object pinned to an optimal policy. (iii) A dominated-probe regime in which Thompson sampling pays at least $(1-p)(1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)$ over a horizon of order $1/p$ while IDS_0 pays at most $(1-p)c_0 + pc_1$, giving a constant-factor separation that is structural rather than asymptotic.

Experiments test the same mechanism. A hidden-mode bandit exhibits the predicted phase transition as the offline posterior concentrates; a biased linear contextual bandit shows the IDS_η ratio improving over standard baselines under informative-but-biased warm start; and **ROID** (*Residual-Optimized Information-Directed Sampling*), a deep IDS_η selector with a TD3+BC backbone [Fujimoto and Gu, 2021] and bootstrapped Q-ensemble, achieves state-of-the-art on D4RL [Fu et al., 2020] continuous-control off-to-on tasks.

2 Related Work

Offline RL and pessimism. Conservative or behaviour-constrained value estimation underlies most offline-only methods, both algorithmically (BCQ [Fujimoto et al., 2019], CQL [Kumar et al., 2020], IQL [Kostrikov et al., 2022], AWAC [Nair et al., 2020], TD3+BC [Fujimoto and Gu, 2021]) and analytically (pessimism in linear MDPs and Bellman-consistent function classes [Jin et al., 2021, Xie et al., 2021, Uehara and Sun, 2022, Rashidinejad et al., 2021]). These methods penalise extrapolation but do not directly extend to the online phase.

Off-to-on RL. Recent work on the two-phase pipeline includes balanced replay and pessimistic ensembles [Lee et al., 2022], calibrated offline pretraining [Nakamoto et al., 2023], iterative policy regularisation [Li et al., 2023], hybrid actor-critic with offline data [Ball et al., 2023, Song et al., 2023], adaptive online policies [Zheng et al., 2023, Guo et al., 2023], and analyses of how offline data should accelerate online learning [Wagenmaker and Pacchiano, 2023, Xie et al., 2023]. Most closely related is the Bayesian design-principles framework of Hu et al. [2024], which argues that probability matching is a principled way to avoid the pessimism-versus-optimism dilemma in offline-to-online fine-tuning and introduces BOORL. Our work shares the posterior-conditioning viewpoint but replaces probability matching with an explicit information-ratio criterion. This lets us analyse residual information through $I(\chi; \tau_{1:T} \mid \mathcal{D}_N)$, track a coverage coefficient tied to the visitation induced by IDS, and identify a dominated-probe regime in which IDS and Thompson sampling separate.

Information-directed sampling. The information-ratio framework was introduced for posterior sampling by Russo and Van Roy [2014, 2016]; IDS was introduced for bandits by Russo and Van Roy [2018], extended to heteroscedastic noise by Kirschner and Krause [2018], and lifted to RL by Hao and Lattimore [2022]; broader information-theoretic foundations appear in Lu et al. [2023]; a closely related decision-estimation viewpoint appears in Foster et al. [2021]. Compared to Hao and Lattimore [2022], who establish IDS regret bounds in the purely online linear-mixture and tabular regimes, our analysis (a) operates in the off-to-on regime where the prior is itself the offline-induced posterior, (b) tracks a coverage object pinned to IDS-induced visitation, and (c) identifies a structural regime in which IDS strictly beats TS, which is not visible from any TS-symmetric analysis.

Cross-community connection. Decision-making from offline datasets is also studied in offline model-based black-box optimization, Bayesian optimization, and contextual bandits, where a surrogate (or reward predictor) fit on logged data is later refined by online queries. The same structural question recurs: after conditioning on the offline dataset, where does the residual uncertainty live, and which online queries cheapest resolve it? The conditional-mutual-information view and the regret–information ratio are community-agnostic, and our bandit and D4RL experiments speak to both the bandits/BO and offline-RL audiences this workshop convenes.

3 Setup

A finite-horizon episodic MDP $M = (\mathcal{S}, \mathcal{A}, H, P, r)$ has horizon H and rewards in $[0, 1]$. A policy is $\pi = (\pi_1, \ldots, \pi_H)$ with $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$. For parameter w and policy π, $V_h^\pi(s; w)$ and $Q_h^\pi(s, a; w)$ are the standard value functions and π_w^* denotes an optimiser. For the regret analysis, we work in a known-dynamics Bayesian linear-reward model. This setting retains the offline-to-online exploration problem while isolating uncertainty in the stage-local reward/observation channel: a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ satisfies $\|\phi\|_2 \le 1$, and for each stage $h \in [H]$ an unknown $w_h \in \mathbb{R}^d$ gives $Q_h^\pi(s, a) = \langle \phi(s, a), w_h^\pi \rangle$. The prior is $w_h \sim \mathcal{N}(0, \lambda^{-1} I)$ with fixed precision $\lambda \ge 1$, independent across stages, and observations carry unit-variance Gaussian noise; we write $w = (w_1, \ldots, w_H)$.

The offline dataset is $\mathcal{D}_N = \bigcup_h \mathcal{D}_{N,h}$ with $n_h \ge 1$ samples at stage h and $N = \sum_h n_h$. The stage-h Gram matrix is $\Lambda_{h,N} := \lambda I + \sum_{(s,a) \in \mathcal{D}_{N,h}} \phi(s, a)\phi(s, a)^\top$. The online phase consists of T episodes indexed $t = 1, \ldots, T$; with $\phi_{t,h} := \phi(s_{t,h}, a_{t,h})$ the rank-1 recursion $\Lambda_{h,N+t} = \Lambda_{h,N+t-1} + \phi_{t,h}\phi_{t,h}^\top$ defines $\Lambda_{h,N+T}$. Let τ_t be the episode-t trajectory and $\mathcal{H}_t := \sigma(\mathcal{D}_N, \tau_{1:t-1})$, with posterior $\beta_t := \mathbb{P}(w \in \cdot \mid \mathcal{H}_t)$. For a candidate policy π and learning target χ, the per-episode regret and information gain are

$$\Delta_t(\pi) := \mathbb{E}_{t,\pi}[V_1^{\pi_w^*}(s_{t,1}; w) - V_1^\pi(s_{t,1}; w)], \qquad I_t^\chi(\pi) := I(\chi; \tau_t \mid \mathcal{H}_t, \pi_t = \pi).$$

Throughout the information-ratio analysis, Π denotes the randomised closure of the candidate policy class; in particular, the episode-level posterior-sampling mixture used as the reference comparator is feasible. For $\eta \geq 0$, the regularised IDS policy is

$$\pi_t^{\mathrm{IDS}_\eta,\chi} \in \arg\min_{\pi\in\Pi} \Psi_t^\eta(\pi;\chi), \qquad \Psi_t^\eta(\pi;\chi) := \frac{\Delta_t(\pi)^2}{I_t^\chi(\pi)+\eta}.$$

We write IDS_η for the rule with score $\Delta^2/(g+\eta)$ and call IDS_0 *vanilla* IDS; "IDS" without a subscript refers to this family. Vanilla IDS_0 uses the convention $\Psi = +\infty$ when $I_t^\chi = 0 < \Delta_t$ and $\Psi = 0$ when $\Delta_t = I_t^\chi = 0$. Bayesian regret over T online episodes conditional on $\mathcal{D}_N$ is $\mathrm{BayesRegret}_T^{\mathrm{off}}(\pi) := \mathbb{E}^{\mathrm{off}}[\sum_{t=1}^T \Delta_t(\pi_t)]$. The notation $\widetilde{O}(\cdot)$ hides factors that are polylogarithmic in T, H, d, N.

Scope of the regret-bound instantiation. For the regret-bound results in Section 4, we work in the stage-local linear-Gaussian observation model: the transition kernel P is known and does not depend on w, and the mean rewards used to define regret remain bounded in $[0,1]$. The component of the stage-h observation $o_{t,h}$ that carries information about w_h is a scalar linear-Gaussian regression target

$$y_{t,h} = \phi(s_{t,h}, a_{t,h})^\top w_h + \epsilon_{t,h}, \qquad \epsilon_{t,h} \sim \mathcal{N}(0,1) \text{ independent},$$

which we treat as the Bayesian observation channel for posterior updates on w_h. The next state $s_{t,h+1}$ is conditionally independent of w given $(\mathcal{H}_{t,h}, a_{t,h}, y_{t,h})$. The structural separation in Section 5 is a separate finite construction and does not rely on this Gaussian observation model.

Linear-reward value-difference inequality. Under the linear-reward and known-dynamics assumptions just stated, a direct telescoping argument yields the following value-difference inequality, which is the only structural property of the linear-Q parametrisation that the regret-ratio proof of Section 4 uses: for any two parameter vectors $u, v \in \mathbb{R}^{Hd}$ and any policy π,

$$\left|V_1^\pi(s_1;u) - V_1^\pi(s_1;v)\right| \leq \sum_{h=1}^H \mathbb{E}_{\pi,P}\big[\big|\phi(s_h,a_h)^\top(u_h - v_h)\big|\big], \tag{1}$$

where the expectation is taken over the known transition kernel P and the policy randomness. We refer to (1) as the *linear-reward value-difference inequality*.

4 An Off-to-On Bayesian Regret Bound for IDS

The argument has three pieces. An information-ratio bound on a reference policy passes to IDS automatically because IDS optimises the same ratio. Cauchy–Schwarz turns a per-episode ratio bound into a sum of square-rooted information gains. The conditional-MI chain rule turns that sum into a single number, $I(\chi;\tau_{1:T} \mid \mathcal{D}_N)$, into which the offline data enters only through the conditioning. The first two pieces are standard for posterior-sampling analysis [Russo and Van Roy, 2014, 2016]; the change of executed policy and the routing through conditional MI are what make the argument work in the off-to-on regime. Full proofs are in Appendix A.

For each t draw $\widetilde{w}_t \sim \beta_t$ independently of w and set $\pi_t^{\text{ref-TS}} := \pi^*(\widetilde{w}_t)$. In the known-dynamics linear-reward instantiation of Section 3, the standard probability-matching identity applies to state-action functionals under the reference posterior sample (Lemma 9 in Appendix A). We require an information-ratio condition on this reference policy.

Assumption 1 (Reference information-ratio condition)**.** There exist a measurable event G in the augmented probability space that includes the auxiliary reference samples $(\widetilde{w}_t)_{t\leq T}$, and a constant $C_\chi \geq 0$, such that, on G and for every $t \leq T$,

$$\Delta_t(\pi_t^{\text{ref-TS}})^2 \leq C_\chi\, I_t^\chi(\pi_t^{\text{ref-TS}}).$$

In the linear-Q model we verify Assumption 1 with $C_w = O(Hd)$ via stage-local decomposition (Section 4.1; proof in Appendix A).

Lemma 2 (Ratio certificate)**.** *On G and for every $\eta \geq 0$, $\Psi_t^\eta(\pi_t^{\mathrm{IDS}_\eta,\chi};\chi) \leq \Psi_t^\eta(\pi_t^{\text{ref-TS}};\chi) \leq C_\chi$.*

The first inequality holds because IDS minimises Ψ_t^η; the second is Assumption 1 after dividing by $I_t^\chi + \eta$. The lemma is one line, but the role it plays is specific: once the executed policy is no longer probability-matching, the reference policy serves only as a witness, and the executed policy inherits its ratio bound over the same policy class. This is the only point in the argument where the choice of executed policy matters.

Proposition 3 (Master inequality)**.** *Under Assumption 1, for every $\eta \geq 0$,*

$$\mathrm{BayesRegret}_T^{\mathrm{off}}(\pi^{\mathrm{IDS}_\eta,\chi}) \leq \sqrt{TC_\chi\big(I(\chi;\tau_{1:T} \mid \mathcal{D}_N) + T\eta\big)} + HT\mathbb{P}^{\mathrm{off}}(G^c).$$

The proof square-roots Lemma 2, applies Cauchy–Schwarz to the sum over episodes, takes $\mathbb{E}^{\mathrm{off}}[\cdot]$ with Jensen on the right, and uses that $\pi_t^{\mathrm{IDS}_\eta,\chi}$ is $\mathcal{H}_t$-measurable to invoke the conditional-MI chain rule [Cover and Thomas, 2006, Thm. 2.5.2]: $\sum_t \mathbb{E}^{\mathrm{off}}[I_t^\chi] = I(\chi;\tau_{1:T} \mid \mathcal{D}_N)$. The full argument is in Appendix A. The proposition is the conceptual core of the paper: the offline dataset enters only through the conditioning of one mutual information, and the two phases are unified at the level of the proof.

4.1 Linear-Q Instantiation

Two ingredients turn Proposition 3 into a closed-form bound: a closed form for $I(w;\tau_{1:T} \mid \mathcal{D}_N)$, and a bound on C_w. Both follow from the standard stage-local decomposition $I(w;\tau_t \mid \mathcal{H}_t,\pi) = \sum_h I(w_h; o_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}, \pi)$ (Lemma 11).

Lemma 4 (Closed-form per-stage information)**.** *Under the known-dynamics linear-reward instantiation, $I(w_h; o_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}) = \frac{1}{2}\log(1 + \phi_{t,h}^\top \Lambda_{h,N+t-1}^{-1}\phi_{t,h})$.*

Lemma 5 (Reference ratio constant)**.** *Assumption 1 holds (with the trivial event $G = \Omega$, i.e. $\mathbb{P}^{\mathrm{off}}(G^c) = 0$) with $C_w = H\Gamma^2 = O(Hd)$, where $\Gamma = \sqrt{4d/\log 2}$.*

The proof of Lemma 5 combines the probability-matching identity, the linear-reward value-difference inequality (1), two applications of Cauchy–Schwarz, and the Bayesian second-moment identity $\mathbb{E}[\|\widetilde{w}_{t,h} - w_h\|^2_{\Lambda_{h,N+t-1}} \mid \mathcal{H}_t] = 2d$, which holds because w_h and $\widetilde{w}_{t,h}$ are independent posterior draws given $\mathcal{H}_t$. Combining Lemmas 4–5 with the rank-1 log-determinant identity [Abbasi-Yadkori et al., 2011, Lem. 11] yields

$$I(w;\tau_{1:T} \mid \mathcal{D}_N) = \tfrac{1}{2}\sum_{h=1}^{H} \mathbb{E}^{\mathrm{off}}\big[\log\det(\Lambda_{h,N+T}) - \log\det(\Lambda_{h,N})\big]. \tag{2}$$

4.2 Two Log-Determinant Bounds and an Interpolation Theorem

Two complementary bounds control the log-determinant. The elliptical-potential bound [Lattimore and Szepesvári, 2020, Lem. 19.4] gives $\log(\det\Lambda_{h,N+T}/\det\Lambda_{h,N}) \leq d\log(1+T/(d\lambda_{\min}(\Lambda_{h,N})))$, which yields the standard $\sqrt{T}$ regret rate. The trace bound, obtained via $\log\det(I+M) \leq \mathrm{Tr}(M)$ [Horn and Johnson, 2012], gives $\log(\det\Lambda_{h,N+T}/\det\Lambda_{h,N}) \leq \sum_{t=1}^T \phi_{t,h}^\top\Lambda_{h,N}^{-1}\phi_{t,h}$, which we control by a coverage coefficient pinned to IDS's own visitation.

Definition 6 (IDS_η-induced coverage)**.** Assume $n_h \geq 1$ for every $h \in [H]$. With

$$\overline{\Sigma}_{h,T}^{\mathrm{IDS}_\eta}(\mathcal{D}_N) := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\phi_{t,h}\phi_{t,h}^\top \;\big|\; \mathcal{D}_N\big]$$

the time-averaged online visitation covariance under IDS_η, define

$$C_{\beta,h}^{\mathrm{IDS}_\eta}(N,T) := \lambda_{\max}\Big((\Lambda_{h,N}/n_h)^{-1/2}\,\overline{\Sigma}_{h,T}^{\mathrm{IDS}_\eta}\,(\Lambda_{h,N}/n_h)^{-1/2}\Big),\;\; C_{\beta,\mathrm{IDS}_\eta}^{\dagger}(N,T) := \frac{N}{H}\sum_{h=1}^{H}\frac{C_{\beta,h}^{\mathrm{IDS}_\eta}(N,T)}{n_h}.$$

We write $C_{\beta,\mathrm{IDS}_0}^{\dagger}(N,T)$ for the same quantity under vanilla IDS_0 ($\eta = 0$).

This coverage object differs from the standard off-to-on coverage coefficient [Rashidinejad et al., 2021, Uehara and Sun, 2022] in two ways: it is pinned to the time-averaged visitation under the algorithm rather than under π^*, and it depends on T and on the algorithm itself. Both differences are necessary in the off-to-on regime, because the algorithm's online visitation is what actually multiplies the inverse offline Gram matrix in the trace bound.

Theorem 7 (Off-to-on Bayesian regret bound)**.** *Under Assumption 1 (verified at $C_w = O(Hd)$), vanilla* IDS_0 *satisfies*

$$\mathrm{BayesRegret}_T^{\mathrm{off}}(\pi^{\mathrm{IDS}_0,w}) \le \widetilde{O}\Big(Hd\,\min\{\sqrt{TL_N(T)},\ T\sqrt{C^{\dagger}_{\beta,\mathrm{IDS}_0}(N,T)/N}\}\Big),$$

with $L_N(T) := \frac{1}{H}\sum_h \log(1 + T/(d\lambda_{\min}(\Lambda_{h,N})))$.

For regularised IDS_η with $\eta > 0$, the analogous two-branch bound holds with $C^{\dagger}_{\beta,\mathrm{IDS}_\eta}$ in place of $C^{\dagger}_{\beta,\mathrm{IDS}_0}$ and an additive slack $T\sqrt{C_w\eta}$; the precise statement and proof are deferred to Corollary 12 in Appendix A.

Theorem 7 follows by combining Proposition 3, Lemma 5, equation (2), and the two log-determinant bounds, with a conditional-trace identity bounding the trace branch by $C^{\dagger}_{\beta,\mathrm{IDS}_0}/N$; see Appendix A. As $T \to \infty$ the bound recovers the Bayesian linear-bandit rate $\widetilde{O}(Hd\sqrt{T})$, the natural extension of $\widetilde{O}(d\sqrt{T})$ [Russo and Van Roy, 2014] to horizon H; as $N \to \infty$ at fixed T the warm-start branch shrinks at rate $\sqrt{C^{\dagger}_{\beta,\mathrm{IDS}_0}/N}$. What is new is that the executed policy is IDS, that the inequality is stated for an arbitrary target χ, and that the warm-start branch tracks a coverage coefficient pinned to the algorithm's own visitation.

5 Structural Separation: When IDS Strictly Beats TS

This section is separate from the randomized-closure guarantee: over a finite candidate set, IDS can select an available diagnostic probe that is never posterior-optimal, whereas TS cannot. Theorem 7 controls worst-case Bayesian regret but does not, by itself, distinguish IDS from any other algorithm whose ratio is at most C_χ, because TS satisfies the same bound. We now exhibit a structurally identifiable subclass of off-to-on problems in which IDS strictly outperforms TS by a constant factor. The mechanism is simple. Once warm-start has concentrated the posterior on a high-probability mode but a low-probability mode remains, the optimal action is almost always one thing, but the algorithm still does not know which thing. Posterior sampling has to commit to whichever mode its sample selects, so it can only resolve the residual mode by playing it, paying positive regret each time it does. An information-aware algorithm can instead pay the regret of a single suboptimal action whose only purpose is to disambiguate.

Let $\theta \in \{0,1\}$ index the residual mode after offline pretraining and let $p := \mathbb{P}(\theta = 1 \mid \mathcal{D}_N) \in (0,1)$. Take $\chi := \theta$ and restrict the candidate class to $\Pi^\dagger := \{\pi_0, \pi_1, \pi_P\}$. The four assumptions, in compact form (formal statements in Appendix B), are: cell-optimality **(B1)**, in which π_0 is uniquely optimal on $\theta = 0$, π_1 is uniquely optimal on $\theta = 1$, and π_P is suboptimal on both; conditional gap constants $\underline{\Delta}_\pm, c_0, c_1 > 0$ **(B2)**; an information structure **(B3)** in which π_0 is uninformative about θ and π_1, π_P are perfectly informative; and the IDS_0 convention **(B4)** that $g = 0 \wedge \Delta > 0 \Rightarrow \Psi = +\infty$.

Theorem 8 (Structural separation)**.** *Under* **(B1)–(B4)** *and* $\Pi = \Pi^\dagger$*, TS never selects* π_P *and the first time it plays* π_1 *is* $T^* \sim \mathrm{Geom}(p)$*. If* $(1-p)c_0 + pc_1 < (1-p)\underline{\Delta}_-$*, vanilla* IDS_0 *selects* π_P *at episode* 1 *and* $\mathrm{BayesRegret}_T^{\mathrm{IDS}_0}(\mathcal{D}_N) \le (1-p)c_0 + pc_1$ *for every* $T \ge 1$*. For every* $T \ge \lceil 1/p \rceil$, $\mathrm{BayesRegret}_T^{\mathrm{TS}}(\mathcal{D}_N) \ge (1-p)(1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)$*. If additionally* $(1-p)c_0 + pc_1 < (1-p)(1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)$*, then* $\mathrm{BayesRegret}_T^{\mathrm{IDS}_0}(\mathcal{D}_N) < \mathrm{BayesRegret}_T^{\mathrm{TS}}(\mathcal{D}_N)$ *for every* $T \ge \lceil 1/p \rceil$.

The proof (Appendix B) computes the three values of Ψ at episode 1 to establish IDS's choice, uses a geometric-hitting-time argument to lower-bound TS regret, and combines the two. The TS lower bound holds because TS by construction never plays π_P, regardless of how the posterior-sampling step is implemented: if $\widetilde{\theta} \in \{0,1\}$, then $\pi^*(\widetilde{w}) \in \{\pi_0, \pi_1\}$ by cell-optimality, so the dominated probe is structurally invisible to TS. The result is a structural separation, not a universal lower bound:

it isolates a concrete off-to-on regime in which probability matching provably misses an informative dominated probe.

A warm-start corollary (Corollary 17) makes the asymptotic gap explicit: if $c_0 < \underline{\Delta}_-$ and $c_0 < (1 - e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)$, the assumptions hold along any sequence of offline datasets with $p_N \downarrow 0$, and with $T_N := \lceil 1/p_N \rceil$ the asymptotic regret ratio satisfies $\liminf_N \mathrm{BayesRegret}_{T_N}^{\mathrm{TS}}/\mathrm{BayesRegret}_{T_N}^{\mathrm{IDS}_0} \geq (1 - e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)/c_0 > 1$. An explicit explicit finite separation instance satisfying **(B1)–(B4)** is constructed in Appendix C.

6 ROID: A Practical Algorithm for Off-to-On Deep RL

The analysis of Sections 4–5 is in the known-dynamics Bayesian linear-reward model with a closed-form posterior. To run the same selection rule on continuous control, we need a finite candidate set and practical surrogates for Δ and g computed from deep critics. We instantiate the regret–information selection rule as a concrete algorithm, **ROID** (*Residual-Optimized Information-Directed Sampling*), and separate the theory-faithful linear implementation from the deep implementation, where ensemble uncertainty provides a practical posterior proxy.

Path A (linear contextual bandit, used for bandit experiments). We use a linear-Gaussian contextual bandit matching the analysis: $r(s, a) = \phi(s, a)^\top w^\star + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and a fixed feature map ϕ. The offline data, collected by a mismatched behaviour policy, is used to warm-start a Bayesian linear regression posterior, which is updated online in closed form. All methods share the same posterior. IDS is implemented with a sampling-based approximation: we draw posterior samples to estimate $\Delta(a) = \mathbb{E}_w[\max_{a'} \phi(s, a')^\top w - \phi(s, a)^\top w]$, and compute $g(a) = \frac{1}{2} \log(1 + \phi(s, a)^\top \Lambda^{-1} \phi(s, a)/\sigma^2)$ in closed form. Actions are selected by minimising $\Delta(a)^2/(g(a) + \eta)$. This setting matches the assumptions of Theorem 7, so the bandit experiments directly test the theory.

Path B (deep ensemble, used for D4RL). For continuous control, we use a Bayesian ensemble backbone [Hu et al., 2024] and modify only the online action selector. The offline stage trains a bootstrapped TD3+BC ensemble; online, this ensemble is used as a posterior surrogate over Q-functions. For each state s, we sample a finite actor-anchored candidate set A_t and choose actions using an IDS-inspired score.

Specifically, for each candidate $a \in A_t$, we estimate

$$\Delta(a) \approx \frac{1}{K} \sum_{k=1}^{K} \Big(V_k^\star(s) - \widehat{Q}_k(s, a) \Big), \qquad g(a) \approx \frac{1}{2} \log \Big(1 + \alpha \, \mathrm{Var}_k[\widehat{Q}_k(s, a)]/\sigma^2 \Big),$$

where $\widehat{Q}_k$ denotes a clipped ensemble critic and $V_k^\star(s)$ is computed over an enlarged candidate set. The selected action is

$$a_t = \arg\min_{a \in A_t} \frac{\Delta(a)^2}{g(a) + \eta}.$$

Thus ROID preserves the regret–information form of IDS, but replaces the exact Bayesian posterior and closed-form information gain by ensemble-based surrogates. The D4RL experiment evaluates whether the ROID selection principle remains stable and effective when implemented with deep posterior surrogates, while the formal guarantees remain those of the linear-Gaussian setting. Full architectural details, clipping, calibration, replay mixing, and hyperparameters are given in Appendix E.

7 Experiments

The experiments test the predicted mechanism rather than establish a new offline-RL backbone. We check three claims: that an information-aware rule probes a cheap diagnostic action when the warm-start posterior leaves a low-probability mode unresolved (hidden-mode bandit); that the regret-information ratio improves over greedy and probability matching under biased warm starts (linear contextual bandit); and that the same rule integrates into a deep offline-to-online pipeline without instability (D4RL continuous control).

(a) Hidden-mode bandit ($T = 500$).

N	p_N	greedy	UCB	TS	IDS_0
100	0.377	0.15	0.50	1.06	**0.15**
200	0.268	0.15	0.15	1.28	**0.15**
300	0.182	0.15	0.15	1.29	**0.15**
1000	0.0066	3.31	3.31	1.79	**0.15**

(b) Biased contextual bandit ($T = 200$).

N	greedy	UCB	TS	$\mathrm{IDS}_{0.5}$
20	42.85	4.80	10.64	**3.57**
50	18.90	1.94	4.73	**1.67**
100	0.88	**0.113**	0.445	0.641

Table 1: Bandit experiments validating the residual-information mechanism. Left: hidden-mode bandit showing the dominated-probe effect. Right: biased contextual bandit showing robustness under warm-start bias. Full η sweeps are in Appendix D.

7.1 Hidden-Mode Bandit

The first experiment isolates the separation mechanism of Theorem 8. A binary latent variable $\theta \in \{0, 1\}$ selects one of two modes. The agent has three actions: $a_0 \equiv \pi_0$ (default), $a_1 \equiv \pi_1$ (rare-mode), and $a_P \equiv \pi_P$ (probe). The rewards are $r(a_0) = 1$ in both modes, $r(a_1) = 0.2/2.0$, and $r(a_P) = 0.85/1.85$, so the probe is suboptimal by 0.15 under both modes but reveals θ in one observation. The offline dataset induces a posterior residual probability $p_N := \mathbb{P}(\theta = 1 \mid \mathcal{D}_N)$ that decreases with N. We run $T = 500$ online steps and average over 10 seeds. Since Theorem 8 concerns vanilla IDS_0, the main text reports IDS_0; the full η sweep is in Appendix D.

Panel (a) of Table 1 shows the predicted phase transition as the residual mode probability p_N shrinks. For moderate residual uncertainty ($N \in \{100, 200, 300\}$), the probe has a favourable regret-information tradeoff, and IDS_0 pays essentially only the one-step probe cost. TS behaves differently: because it explores by probability matching, it resolves the rare mode only when that mode is sampled, which produces consistently larger regret.

The most diagnostic case is $N = 1000$, where $p_N = 0.0066$. Greedy commits to the default action and never resolves the rare mode. UCB also fails because its uncertainty bonus scales with the posterior standard deviation, which shrinks as $\sqrt{p_N(1 - p_N)}$. Vanilla IDS_0 still probes because the boundary convention $g = 0 < \Delta \Rightarrow \Psi = +\infty$ rules out actions that incur positive regret while providing no information. This is the empirical signature of Theorem 8: the advantage of IDS_0 comes from selecting a dominated but informative action, not from generic optimism.

7.2 Linear Contextual Bandit with Biased Warm-Start

The second experiment tests a less discrete version of the same mechanism: the offline posterior is informative but biased. We use a linear contextual bandit

$$r(s, a) = \phi(s, a)^\top w^* + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

with a nonlinear random feature map. Offline data is generated by a biased behaviour policy $w_{\mathrm{behav}} = w^* + \beta\delta$ with $\beta = 6$, so the warm-start posterior is shifted toward the behaviour policy. A candidate set of size $M = 256$ is sampled once and held fixed across online steps, matching the actor-anchored candidate set used later in Section 6. State and action dimensions are 16 and 4, the feature dimension is 128, $\sigma = 0.05$, the online horizon is $T = 200$, and results are averaged over 20 seeds.

Panel (b) of Table 1 shows when the regularised rule helps. Under strongly biased warm-start ($N = 20, 50$), greedy exploits the wrong posterior preference and TS pays for matching the biased posterior; $\mathrm{IDS}_{0.5}$ improves over both and over UCB. When the warm start is already accurate ($N = 100$), UCB is best and $\mathrm{IDS}_{0.5}$ over-regularises the ratio. The experiment thus supports the intended mechanism rather than a uniform dominance claim: regularised IDS is most useful when the offline posterior is informative but biased.

7.3 D4RL Benchmark

The D4RL experiment isolates the effect of the online action-selection rule in a deep offline-to-online pipeline. We keep the offline backbone fixed and replace only the online selector by ROID, the practical IDS_η rule introduced in Section 6. We use the Path B implementation: each run trains a 5-member TD3+BC ensemble offline, repacks the twin critics as $K = 10$ Q-functions, and then

Table 2: D4RL normalised return (offline→online). Bold marks the highest online value per row. ROID numbers are mean over 3 seeds. Baselines are ODT [Zheng et al., 2023], PEX [Zhang et al., 2023], Cal-QL [Nakamoto et al., 2023], RLPD [Ball et al., 2023] and BOORL [Hu et al., 2024]. RLPD is a from-scratch online method, so only its online value is reported. The ROID column is our result.

Task	Type	RLPD	ODT	PEX	Cal-QL	BOORL	**ROID (ours)**
Hopper	medium	107.3	66.9→97.5	63.8→78.6	75.8→100.6	61.9→109.8	51.21→**113.22**
	medium-replay	58.9	86.6→88.8	89.8→103.3	95.4→106.1	75.5→**111.1**	56.47→106.33
Walker2d	medium	108.6	72.1→76.7	79.8→94.8	80.8→89.6	83.6→107.7	82.63→**110.60**
	medium-replay	115.0	68.9→76.8	73.6→89.3	83.8→94.5	69.1→114.4	80.50→**131.58**
HalfCheetah	medium	90.5	42.7→42.1	47.3→67.8	48.0→72.3	47.9→**98.7**	47.48→95.51
	medium-replay	87.6	39.9→40.4	44.1→55.2	46.5→59.5	44.5→91.5	43.90→**91.65**
Sum (online)		567.9	422.3	489.0	522.6	633.2	**648.9**

uses the regularised IDS_η selector. The selector changes only the online action choice; the offline representation and critic ensemble play the role of a posterior surrogate. We evaluate on six D4RL [Fu et al., 2020] `-v2` locomotion tasks: medium and medium-replay variants of HOPPER, WALKER2D, and HALFCHEETAH. Per-environment hyperparameters are given in Appendix E. Evaluation uses the Polyak target actor on 10 deterministic rollouts every $5{,}000$ environment steps.

Table 2 provides evidence that ROID improves online fine-tuning in this D4RL protocol. It obtains the best score on $4/6$ tasks and the highest summed online score, exceeding BOORL by 15.7 points in total. The strongest improvement occurs on WALKER2D-MEDIUM-REPLAY, where ROID reaches 131.58, compared with 115.0 for RLPD and 114.4 for BOORL. The gains concentrate on medium-replay tasks, where the offline data is heterogeneous and residual uncertainty is more consequential; this is exactly the regime targeted by the regret-information criterion. On easier warm starts such as HALFCHEETAH-MEDIUM, ROID remains competitive but does not dominate.

8 Conclusion

Offline-to-online learning changes what remains uncertain after the offline dataset is observed. We formalised this residual uncertainty by $I(\chi; \tau_{1:T} \mid \mathcal{D}_N)$ and showed that IDS targets it through a regret–information tradeoff, yielding an off-to-on Bayesian regret bound with a two-branch known-dynamics linear-reward guarantee that interpolates between standard online learning and warm-start coverage along IDS-induced visitation, and a dominated-probe regime in which IDS strictly separates from Thompson sampling. Bandit and D4RL experiments support the same mechanism: IDS is most useful when offline data narrows the posterior but leaves biased or low-probability residual modes. The residual-information view is community-agnostic and offers a common language for decision-making from offline datasets across offline RL, offline black-box optimization, Bayesian optimization, and contextual bandits.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning (ICML)*, 2023.

Keru Chen, Honghao Wei, Zhigang Deng, and Sen Lin. Towards fast safe online reinforcement learning via policy finetuning. *Transactions on Machine Learning Research*, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. In *arXiv preprint arXiv:2112.13487*, 2021.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning (ICML)*, 2019.

Siyuan Guo, Lixin Zou, Hechang Chen, Bohao Qu, Haotian Chi, Philip S. Yu, and Yi Chang. Sample efficient offline-to-online reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*, 2023. doi: 10.1109/TKDE.2023.3302804.

Botao Hao and Tor Lattimore. Regret bounds for information-directed reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

Hao Hu, Yiqin Yang, Jianing Ye, Chengjie Wu, Ziqing Mai, Yujing Hu, Tangjie Lv, Changjie Fan, Qianchuan Zhao, and Chongjie Zhang. Bayesian design principles for offline-to-online reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19491–19515, 2024.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory (COLT)*, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. *Conference on Learning Theory (COLT)*, 2018.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble. In *Conference on Robot Learning (CoRL)*, 2022.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Jianxiong Li, Xiao Hu, Haoran Xu, Jingjing Liu, Xianyuan Zhan, and Ya-Qin Zhang. PROTO: Iterative policy regularized offline-to-online reinforcement learning. In *arXiv preprint arXiv:2305.15669*, 2023.

Weiting Liu, Han Wu, Yufei Kuang, Xiongwei Han, Tao Zhong, Jianfeng Feng, and Wenlian Lu. Automated optimization modeling via a localizable error-driven perspective. *arXiv preprint arXiv:2602.11164*, 2026.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. Reinforcement learning, bit by bit. *Foundations and Trends in Machine Learning*, 16(6): 733–865, 2023.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. AWAC: Accelerating online reinforcement learning with offline datasets. In *arXiv preprint arXiv:2006.09359*, 2020.

Mitsuhiko Nakamoto, Yuexiang Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

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

Yuda Song, Yifei Zhou, Ayush Sekhari, J. Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. In *International Conference on Learning Representations (ICLR)*, 2023.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. In *Biometrika*, 1933.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations (ICLR)*, 2022.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2023.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Huaiyuan Yao, Longchao Da, Vishnu Nandam, Justin Turnau, Zhiwei Liu, Linsey Pang, and Hua Wei. Comal: Collaborative multi-agent large language models for mixed-autonomy traffic. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*, pages 409–418. SIAM, 2025a.

Huaiyuan Yao, Pengfei Li, Bu Jin, Yupeng Zheng, An Liu, Lisen Mu, Qing Su, Qian Zhang, Yilun Chen, and Peng Li. Lilodriver: A lifelong learning framework for closed-loop motion planning in long-tail autonomous driving scenarios. *arXiv preprint arXiv:2505.17209*, 2025b.

Huaiyuan Yao, Longchao Da, Xiaoou Liu, Charles Fleming, Tianlong Chen, and Hua Wei. Langmarl: Natural language multi-agent reinforcement learning. *arXiv preprint arXiv:2604.00722*, 2026.

Haichao Zhang, We Xu, and Haonan Yu. Policy expansion for bridging offline-to-online reinforcement learning. *arXiv preprint arXiv:2302.00935*, 2023.

Lina Zhao, Jiaxing Bai, Zihao Bian, Qingyue Chen, Yafang Li, Guangbo Li, Min He, Huaiyuan Yao, and Zongjiu Zhang. Autonomous multi-modal llm agents for treatment planning in focused ultrasound ablation surgery. *arXiv preprint arXiv:2505.21418*, 2025.

Han Zheng, Xufang Luo, Pengfei Wei, Xuan Song, Dongsheng Li, and Jing Jiang. Adaptive policy learning for offline-to-online reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2023.

A Proofs from Section 4

We give complete, self-contained proofs of the results in Section 4. Throughout the appendix, all expectations are conditional on $\mathcal{D}_N$ unless otherwise noted; we write $\mathbb{E}^{\text{off}}$ and $\mathbb{P}^{\text{off}}$ for $\mathbb{E}[\cdot \mid \mathcal{D}_N]$ and $\mathbb{P}(\cdot \mid \mathcal{D}_N)$.

A.1 Probability matching

Lemma 9 (Probability matching)**.** *Assume the transition kernel P does not depend on w (the known-dynamics linear-reward instantiation of Section 3). For any deterministic function $F_h(s, a; \beta_t)$ of (s, a) given β_t,*

$$\mathbb{E}_{t,\pi_t^{\text{ref-TS}}}[F_h(s_{t,h}, a_{t,h}; \beta_t)] = \mathbb{E}_{w\sim\beta_t}\mathbb{E}_{\pi_w^*}[F_h(s_h^{\pi_w^*}, a_h^{\pi_w^*}; \beta_t)].$$

Proof. Conditional on β_t, the reference sample $\widetilde{w}_t$ is drawn from the same posterior as the true parameter w. Because the transition kernel does not depend on w, the trajectory law under the policy $\pi^*(\widetilde{w}_t)$ played in the true MDP depends on $\widetilde{w}_t$ only through the policy it selects and not through any separate dynamics parameter. Therefore

$$\mathbb{E}_{t,\pi_t^{\text{ref-TS}}}[F_h(s_{t,h}, a_{t,h}; \beta_t)] = \int \mathbb{E}_{\pi_{\widetilde{w}}^*}\big[F_h(s_h^{\pi_{\widetilde{w}}^*}, a_h^{\pi_{\widetilde{w}}^*}; \beta_t)\big]\, d\beta_t(\widetilde{w}),$$

where the inner expectation is taken under the trajectory law of $\pi_{\widetilde{w}}^*$ in the (parameter-free) dynamics. Renaming the dummy variable $\widetilde{w}$ as w yields the claim. □

A.2 Stage-local decomposition

Assumption 10 (Stage-local conditional independence)**.** For every t, h, $o_{t,h} \perp w_{-h} \mid (w_h, \mathcal{H}_{t,h}, a_{t,h})$, where $w_{-h} := (w_{h'})_{h'\neq h}$.

This assumption holds in the stage-local linear-Gaussian instantiation of Section 3: the regression-target component $y_{t,h} = \phi_{t,h}^\top w_h + \epsilon_{t,h}$ depends on w only through w_h, and the transition component $s_{t,h+1}$ does not carry additional information about w since the dynamics are independent of w.

Lemma 11 (Episode MI decomposition)**.** *Under Assumption 10, for any $\mathcal{H}_t$-measurable policy π, $I(w; \tau_t \mid \mathcal{H}_t, \pi_t = \pi) = \sum_{h=1}^H I(w_h; o_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}, \pi_t = \pi)$.*

Proof. The trajectory factors as $\tau_t = (s_{t,1}, a_{t,1}, o_{t,1}, \ldots, a_{t,H}, o_{t,H})$. Since $s_{t,1} \sim \rho_1$ is independent of w, $I(w; s_{t,1} \mid \mathcal{H}_t, \pi) = 0$. Applying the chain rule of conditional mutual information [Cover and Thomas, 2006, Thm. 2.5.2] repeatedly,

$$I(w; \tau_t \mid \mathcal{H}_t, \pi) = \sum_{h=1}^H \Big[I(w; a_{t,h} \mid \mathcal{H}_{t,h}, \pi) + I(w; o_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}, \pi)\Big].$$

Because π is $\mathcal{H}_t$-measurable and $a_{t,h} = \pi_h(s_{t,h})$ is generated using fresh independent randomness, $a_{t,h} \perp w \mid (\mathcal{H}_{t,h}, \pi)$, so $I(w; a_{t,h} \mid \mathcal{H}_{t,h}, \pi) = 0$ for every h. Applying Assumption 10 to each remaining term, under the conditioning $(\mathcal{H}_{t,h}, a_{t,h}, \pi)$ the observation $o_{t,h}$ is independent of w_{-h} given w_h, so $I(w; o_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}, \pi) = I(w_h; o_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}, \pi)$. Summing over h gives the claim. □

A.3 Closed-form per-stage MI

Proof of Lemma 4. Conditional on $\mathcal{H}_{t,h}$, the prior on w_h from Section 3 ($\mathcal{N}(0, \lambda^{-1}I)$) updated by all offline samples and online observations through episode $t-1$ is again Gaussian: $w_h \mid \mathcal{H}_{t,h} \sim \mathcal{N}(\mu_{h,N+t-1}, \Lambda_{h,N+t-1}^{-1})$.

Given $a_{t,h}$, the component of $o_{t,h}$ that carries information about w_h is the Gaussian regression target $y_{t,h} = \phi_{t,h}^\top w_h + \epsilon_{t,h}$ with $\epsilon_{t,h} \sim \mathcal{N}(0, 1)$; the remaining component (the next state $s_{t,h+1}$) is independent of w_h given $(s_{t,h}, a_{t,h})$ by the scope assumption of Section 3. Hence $I(w_h; o_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}) = I(w_h; y_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h})$. By the Gaussian-channel mutual-information identity [Cover

and Thomas, 2006, Thm. 9.1.1], $I(X; X+Z) = \frac{1}{2}\log(1+\sigma_X^2/\sigma_Z^2)$. Apply it with $X = \phi_{t,h}^\top w_h$ (variance $\phi_{t,h}^\top \Lambda_{h,N+t-1}^{-1}\phi_{t,h}$) and $Z = \epsilon_{t,h}$:

$$I(\phi_{t,h}^\top w_h;\, y_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}) = \tfrac{1}{2}\log\big(1+\phi_{t,h}^\top \Lambda_{h,N+t-1}^{-1}\phi_{t,h}\big).$$

Finally, $y_{t,h}$ depends on w_h only through the scalar projection $\phi_{t,h}^\top w_h$, so this projection is a sufficient statistic of w_h for $y_{t,h}$ given $(\mathcal{H}_{t,h}, a_{t,h})$. The Markov chain $w_h \to \phi_{t,h}^\top w_h \to y_{t,h}$ together with the converse data-processing identity for sufficient statistics [Cover and Thomas, 2006, Sec. 2.8] therefore gives

$$I(w_h; y_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}) = I(\phi_{t,h}^\top w_h;\, y_{t,h} \mid \mathcal{H}_{t,h}, a_{t,h}).$$

□

A.4 Reference ratio constant

Proof of Lemma 5. We will show that

$$\Delta_t(\pi_t^{\text{ref-TS}}) \le \Gamma \sum_{h=1}^{H} \sqrt{\mathbb{E}_{t,\pi_t^{\text{ref-TS}}}[J_{t,h}^w]}, \qquad J_{t,h}^w := \tfrac{1}{2}\log\big(1+\phi_{t,h}^\top \Lambda_{h,N+t-1}^{-1}\phi_{t,h}\big),$$

with $\Gamma = \sqrt{4d/\log 2}$, and then deduce the ratio bound by Cauchy–Schwarz over stages. The proof uses no high-probability concentration event: it works at the level of posterior second moments.

Step 1 (regret to posterior linear differences). Let $\widetilde{w}_t = (\widetilde{w}_{t,h})_{h=1}^H$ be the posterior sample used by the reference TS policy, drawn independently of w from β_t. Probability matching (Lemma 9) gives

$$\Delta_t(\pi_t^{\text{ref-TS}}) = \mathbb{E}_t\big[V_1^{\pi^*_{\widetilde{w}_t}}(s_{t,1};\widetilde{w}_t) - V_1^{\pi^*_{\widetilde{w}_t}}(s_{t,1};w)\big],$$

where the two value functions are evaluated under the same policy $\pi^*_{\widetilde{w}_t}$ at different model parameters. Applying the linear-reward value-difference inequality (1) with $u = \widetilde{w}_t$, $v = w$, and $\pi = \pi^*_{\widetilde{w}_t} = \pi_t^{\text{ref-TS}}$,

$$\Delta_t(\pi_t^{\text{ref-TS}}) \le \sum_{h=1}^{H} \mathbb{E}_{t,\pi_t^{\text{ref-TS}}}\big[\big|\phi_{t,h}^\top(\widetilde{w}_{t,h} - w_h)\big|\big]. \tag{3}$$

Step 2 (Cauchy–Schwarz in the Λ-norm). For each stage h, Cauchy–Schwarz in the $\Lambda_{h,N+t-1}$-norm gives the pointwise bound

$$\big|\phi_{t,h}^\top(\widetilde{w}_{t,h} - w_h)\big| \le \sqrt{\phi_{t,h}^\top \Lambda_{h,N+t-1}^{-1}\phi_{t,h}} \;\cdot\; \|\widetilde{w}_{t,h} - w_h\|_{\Lambda_{h,N+t-1}}.$$

A second application of Cauchy–Schwarz to the joint expectation over the trajectory and the parameter pair yields

$$\mathbb{E}_{t,\pi_t^{\text{ref-TS}}}\big[\big|\phi_{t,h}^\top(\widetilde{w}_{t,h} - w_h)\big|\big] \le \sqrt{\mathbb{E}_{t,\pi_t^{\text{ref-TS}}}\big[\phi_{t,h}^\top \Lambda_{h,N+t-1}^{-1}\phi_{t,h}\big]}\,\sqrt{\mathbb{E}_t\big[\|\widetilde{w}_{t,h} - w_h\|^2_{\Lambda_{h,N+t-1}}\big]}. \tag{4}$$

Step 3 (Bayesian second-moment identity). Conditional on $\mathcal{H}_t$, w_h and $\widetilde{w}_{t,h}$ are independent draws from the posterior $\mathcal{N}(\mu_{h,N+t-1}, \Lambda_{h,N+t-1}^{-1})$, the former by definition of β_t and the latter by independent posterior sampling. Hence $\widetilde{w}_{t,h} - w_h$ is conditionally Gaussian with mean zero and covariance $2\Lambda_{h,N+t-1}^{-1}$, and (since $\Lambda_{h,N+t-1}$ is $\mathcal{H}_t$-measurable)

$$\mathbb{E}_t\big[\|\widetilde{w}_{t,h} - w_h\|^2_{\Lambda_{h,N+t-1}}\big] = \operatorname{Tr}\big(\Lambda_{h,N+t-1}\cdot 2\Lambda_{h,N+t-1}^{-1}\big) = 2d. \tag{5}$$

This step is exact, no concentration needed.

Step 4 (variance to information gain). Since $\|\phi\| \le 1$ and $\Lambda_{h,N+t-1} \succeq \lambda I$ with $\lambda \ge 1$, $x := \phi_{t,h}^\top \Lambda_{h,N+t-1}^{-1}\phi_{t,h} \in [0,1]$. On $[0,1]$ the inequality $\log(1+x) \ge x\log 2$ holds (this is $2^x \le 1+x$, which follows from concavity of $1+x-2^x$ on $[0,1]$ and its vanishing at the endpoints), so

$$x \le \tfrac{2}{\log 2}\cdot\tfrac{1}{2}\log(1+x) = \tfrac{2}{\log 2}\, J_{t,h}^w. \tag{6}$$

Combining (4), (5), and (6),

$$\mathbb{E}_{t,\pi_t^{\text{ref-TS}}}\big[\big|\phi_{t,h}^\top(\widetilde{w}_{t,h}-w_h)\big|\big] \le \sqrt{\tfrac{2}{\log 2}\,\mathbb{E}_{t,\pi_t^{\text{ref-TS}}}[J_{t,h}^w]}\cdot\sqrt{2d} = \Gamma\sqrt{\mathbb{E}_{t,\pi_t^{\text{ref-TS}}}[J_{t,h}^w]}, \qquad \Gamma := \sqrt{\tfrac{4d}{\log 2}}.$$

Substituting into (3),

$$\Delta_t(\pi_t^{\text{ref-TS}}) \le \Gamma \sum_{h=1}^{H} \sqrt{\mathbb{E}_{t,\pi_t^{\text{ref-TS}}}[J_{t,h}^w]}.$$

Step 5 (Cauchy–Schwarz over stages). By $(\sum_h a_h)^2 \le H\sum_h a_h^2$,

$$\Delta_t(\pi_t^{\text{ref-TS}})^2 \le H\Gamma^2 \sum_{h=1}^{H} \mathbb{E}_{t,\pi_t^{\text{ref-TS}}}[J_{t,h}^w] = H\Gamma^2\, I_t^w(\pi_t^{\text{ref-TS}}),$$

where the equality uses Lemmas 11–4. This is Assumption 1 with $G=\Omega$ and $C_w = H\Gamma^2 = \frac{4Hd}{\log 2} = O(Hd)$. □

A.5 Master inequality and interpolation theorem

Proof of Proposition 3. By Lemma 2, on G and for every $t \le T$, $\Delta_t^2 \le C_\chi(I_t^\chi + \eta)$. Taking square roots and using $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, then summing over t and applying Cauchy–Schwarz,

$$\sum_{t=1}^{T} \Delta_t\, \mathbf{1}\{G\} \le \sqrt{TC_\chi}\sqrt{\sum_{t=1}^{T}(I_t^\chi+\eta)}.$$

Apply $\mathbb{E}^{\text{off}}[\cdot]$ with Jensen on the right. Since $\pi_t^{\text{IDS}_\eta,\chi}$ is $\mathcal{H}_t$-measurable, $\mathbb{E}^{\text{off}}[I_t^\chi(\pi_t^{\text{IDS}_\eta,\chi}) \mid \mathcal{H}_t] = I(\chi;\tau_t \mid \mathcal{H}_t)$, and the conditional-MI chain rule [Cover and Thomas, 2006, Thm. 2.5.2] gives $\sum_t I(\chi;\tau_t \mid \mathcal{H}_t) = I(\chi;\tau_{1:T} \mid \mathcal{D}_N)$. Since rewards are in $[0,1]$ over horizon H, $\Delta_t \le H$ on G^c, contributing at most $HT\mathbb{P}^{\text{off}}(G^c)$. Combining gives the proposition. □

Proof of equation (2). By Lemma 11 and Lemma 4, $I(w;\tau_{1:T} \mid \mathcal{D}_N) = \sum_{h,t}\mathbb{E}^{\text{off}}[\frac{1}{2}\log(1+\phi_{t,h}^\top\Lambda_{h,N+t-1}^{-1}\phi_{t,h})]$. The rank-1 log-determinant identity [Abbasi-Yadkori et al., 2011, Lem. 11] gives $\log\det(A+vv^\top) = \log\det A + \log(1+v^\top A^{-1}v)$, so the sum over t telescopes to $\log\det\Lambda_{h,N+T} - \log\det\Lambda_{h,N}$, giving the claim. □

Proof of Theorem 7. Specialising Proposition 3 to vanilla IDS_0 ($\eta = 0$) and combining with Lemma 5 (which holds with $G=\Omega$, so the failure term $HT\mathbb{P}^{\text{off}}(G^c)$ vanishes) gives

$$\text{BayesRegret}_T^{\text{off}}(\pi^{\text{IDS}_0,w}) \le \sqrt{T\,C_w\,I(w;\tau_{1:T} \mid \mathcal{D}_N)}, \qquad C_w = \tfrac{4Hd}{\log 2} = O(Hd).$$

Branch 1 (elliptical potential). The elliptical-potential bound gives $\frac{1}{2}\sum_h \mathbb{E}^{\text{off}}[\log(\det\Lambda_{h,N+T}/\det\Lambda_{h,N})] \le \frac{Hd}{2}L_N(T)$. Substituting and using $C_w = O(Hd)$ yields $\sqrt{T\cdot Hd\cdot Hd\,L_N(T)} = \widetilde{O}\big(Hd\sqrt{TL_N(T)}\big)$.

Branch 2 (IDS_0-induced coverage). The trace bound combined with the conditional-trace identity below gives $\frac{1}{2}\sum_h \mathbb{E}^{\text{off}}[\log(\det\Lambda_{h,N+T}/\det\Lambda_{h,N})] \le \frac{TdH}{2}\cdot C_{\beta,\text{IDS}_0}^\dagger/N$, yielding $\sqrt{T\cdot Hd\cdot THd\,C_{\beta,\text{IDS}_0}^\dagger/N} = \widetilde{O}\big(HdT\sqrt{C_{\beta,\text{IDS}_0}^\dagger(N,T)/N}\big)$.

Conditional-trace identity. By linearity, $\mathbb{E}^{\text{off}}[\sum_t \phi_{t,h}^\top\Lambda_{h,N}^{-1}\phi_{t,h}] = T\,\text{Tr}(\Lambda_{h,N}^{-1}\overline{\Sigma}_{h,T}^{\text{IDS}_0})$. Set $A := \Lambda_{h,N}/n_h$ and $B := \overline{\Sigma}_{h,T}^{\text{IDS}_0}$. Then $\text{Tr}(\Lambda_{h,N}^{-1}B) = n_h^{-1}\,\text{Tr}(A^{-1/2}BA^{-1/2}) \le \frac{d}{n_h}\lambda_{\max}(A^{-1/2}BA^{-1/2}) = \frac{d}{n_h}C_{\beta,h}^{\text{IDS}_0}$. Taking the minimum of the two branches gives the bound. □

A.6 Regret bound for regularised IDS_η

Corollary 12 (Regularised IDS_η)**.** *Under the conditions of Theorem 7, for every $\eta > 0$, regularised IDS_η satisfies*

$$\mathrm{BayesRegret}_T^{\mathrm{off}}(\pi^{\mathrm{IDS}_\eta,w}) \le \widetilde{O}\Big(Hd\,\min\big\{\sqrt{TL_N(T)},\ T\sqrt{C^\dagger_{\beta,\mathrm{IDS}_\eta}(N,T)/N}\big\}\Big) + T\sqrt{C_w\eta},$$

where $C_w = 4Hd/\log 2$ from Lemma 5. The price of regularisation is the additive slack $T\sqrt{C_w\eta}$, which vanishes as $\eta \downarrow 0$ and recovers Theorem 7.

Proof. Apply Proposition 3 at $\eta > 0$. Since $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$, the master inequality splits as

$$\mathrm{BayesRegret}_T^{\mathrm{off}}(\pi^{\mathrm{IDS}_\eta,w}) \le \sqrt{TC_w\, I(w;\tau_{1:T} \mid \mathcal{D}_N)} + T\sqrt{C_w\eta}.$$

The first term is treated exactly as in the proof of Theorem 7, with the only change that the trajectory law (and hence the visitation-induced coverage coefficient) is now under IDS_η rather than IDS_0. The conditional-trace identity therefore yields the trace branch with $C^\dagger_{\beta,\mathrm{IDS}_\eta}(N,T)$ in place of $C^\dagger_{\beta,\mathrm{IDS}_0}(N,T)$; the elliptical-potential branch is unchanged. Combining gives the stated bound. □

B Proofs from Section 5

B.1 Formal assumptions

Assumption 13 (Cell-optimality, B1)**. (B1.1)** For every w with $\theta = 0$, π_0 is the unique globally optimal policy. **(B1.2)** For every w with $\theta = 1$, π_1 is the unique globally optimal policy. **(B1.3)** π_P is strictly suboptimal for every w with $\theta \in \{0,1\}$.

Assumption 14 (Conditional gaps, B2)**.** A history h is unresolved if $0 < p_h := \mathbb{P}(\theta = 1 \mid h, \mathcal{D}_N) < 1$. Constants $\underline{\Delta}_+, \underline{\Delta}_-, c_0, c_1 > 0$ satisfy, for every unresolved h, $\mathbb{E}[V_1^*(w) - V_1^{\pi_0}(w) \mid h, \theta = 1] \ge \underline{\Delta}_+$, $\mathbb{E}[V_1^*(w) - V_1^{\pi_1}(w) \mid h, \theta = 0] \ge \underline{\Delta}_-$, $\mathbb{E}[V_1^*(w) - V_1^{\pi_P}(w) \mid h, \theta = 0] \le c_0$, and $\mathbb{E}[V_1^*(w) - V_1^{\pi_P}(w) \mid h, \theta = 1] \le c_1$.

Assumption 15 (Information structure, B3)**. (B3.1)** Under π_0 the trajectory law is identical on $\{\theta = 0\}$ and $\{\theta = 1\}$, so $I(\theta;\tau \mid \pi_0, h) = 0$ and $\mathbb{P}(\theta = 1 \mid h, \tau, \pi_0) = p_h$ a.s. **(B3.2)** Each of π_1, π_P is perfectly informative: $I(\theta;\tau \mid \pi_1, h) = I(\theta;\tau \mid \pi_P, h) = H_2(p_h)$. Here $H_2(p) := -p\log p - (1-p)\log(1-p)$ is the binary entropy in nats.

Assumption 16 (Vanilla IDS convention, B4)**.** $g = 0 < \Delta \Rightarrow \Psi = +\infty$; $g = \Delta = 0 \Rightarrow \Psi = 0$.

B.2 Proof of Theorem 8

Full proof. Part (i): TS never plays π_P and the discovery time is geometric. TS plays $\pi^*(\widetilde{w})$ with $\widetilde{\theta} \in \{0,1\}$. By (B1.1)–(B1.2) the unique optimiser is $\pi^*(\widetilde{w}) = \pi_0$ if $\widetilde{\theta} = 0$ and π_1 if $\widetilde{\theta} = 1$, and (B1.3) rules out π_P. For the geometric law: on any TS trajectory in which only π_0 has been played up to (but not including) episode t, (B3.1) keeps the conditional distribution of θ given the history at p. Hence $\mathbb{P}(\text{TS plays } \pi_1 \text{ at episode } t \mid \text{not yet}) = p$ i.i.d. across t, so $T^* \sim \mathrm{Geom}(p)$.

Part (ii): IDS picks π_P at episode 1 and pays bounded regret. At episode 1, $p_1 = p \in (0,1)$. Compute each Ψ:

- π_0: $\Delta(\pi_0) \ge p\underline{\Delta}_+ > 0$, $g(\pi_0) = 0$, so $\Psi(\pi_0) = +\infty$ by (B4).
- π_1: $\Delta(\pi_1) \ge (1-p)\underline{\Delta}_-$, $g(\pi_1) = H_2(p)$, so $\Psi(\pi_1) \ge ((1-p)\underline{\Delta}_-)^2/H_2(p)$.
- π_P: $\Delta(\pi_P) \le (1-p)c_0 + pc_1$, $g(\pi_P) = H_2(p)$, so $\Psi(\pi_P) \le ((1-p)c_0 + pc_1)^2/H_2(p)$.

The strict-probe condition $(1-p)c_0 + pc_1 < (1-p)\underline{\Delta}_-$ gives $\Psi(\pi_P) < \Psi(\pi_1)$. Combined with $\Psi(\pi_0) = +\infty$, IDS strictly prefers π_P.

By (B3.2), one episode of π_P is perfectly informative, so the posterior on θ collapses to a Dirac on the true value. At every subsequent episode the history is resolved; the corresponding $\pi_{\theta_{\mathrm{true}}} \in \Pi^\dagger$ has $\Delta = 0$ and $g = 0$, so $\Psi = 0$ by (B4). Therefore $\mathrm{BayesRegret}_T^{\mathrm{IDS}_0}(\mathcal{D}_N) \le \Delta(\pi_P) \le (1-p)c_0 + pc_1$.

Part (iii): TS lower bound. Decompose by T^*. For $1 \le t < T^*$, TS plays π_0 with per-episode regret $\ge p\underline{\Delta}_+$. At $t = T^*$ (if $T^* \le T$), TS plays π_1, contributing $\ge (1-p)\underline{\Delta}_-$. So

$$\mathrm{BayesRegret}_T^{\mathrm{TS}}(\mathcal{D}_N) \ge p\underline{\Delta}_+ \mathbb{E}[\min(T^*-1, T)] + (1-p)\underline{\Delta}_- \mathbb{P}(T^* \le T).$$

For $T^* \sim \mathrm{Geom}(p)$, $\mathbb{P}(T^* \le T) = 1-(1-p)^T$ and $\mathbb{E}[\min(T^*-1,T)] = (1-p)(1-(1-p)^T)/p$. Substituting, $\mathrm{BayesRegret}_T^{\mathrm{TS}} \ge (1-p)(1-(1-p)^T)(\underline{\Delta}_+ + \underline{\Delta}_-)$. For $T \ge \lceil 1/p \rceil$, $(1-p)^T \le e^{-1}$, so $\mathrm{BayesRegret}_T^{\mathrm{TS}} \ge (1-p)(1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)$.

Part (iv). Combine the upper bound of (ii) with the lower bound of (iii) under the separation condition. □

B.3 Warm-start corollary and policy-class extension

Corollary 17 (Warm-start threshold)**.** *Let $\mathcal{D}_N$ vary with $N \in \mathbb{N}$ inducing $p_N \downarrow 0$, with the four assumptions uniform. If $c_0 < \underline{\Delta}_-$ and $c_0 < (1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)$, then there is a threshold $N_\star$ such that for every $N \ge N_\star$, with $T_N := \lceil 1/p_N \rceil$, $\liminf_{N\to\infty} \mathrm{BayesRegret}_{T_N}^{\mathrm{TS}}/\mathrm{BayesRegret}_{T_N}^{\mathrm{IDS}_0} \ge (1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)/c_0 > 1$.*

Proof. The strict-probe condition is continuous in p and reduces as $p \to 0^+$ to $c_0 < \underline{\Delta}_-$, which holds by assumption; similarly the separation condition reduces to $c_0 < (1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)$. Apply Theorem 8; the lim-inf follows from $((1-p_N)(1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-))/((1-p_N)c_0 + p_N c_1) \to (1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-)/c_0$. □

Corollary 18 (Domination under information-structure closure)**.** *Let $\Pi \supseteq \Pi^\dagger = \{\pi_0, \pi_1, \pi_P\}$ and assume **(IS-Closure)**: for every $\pi \in \Pi$ and every unresolved h, $g_1(\pi; \theta \mid h) \in \{0, H_2(p_h)\}$. Under **(B1)–(B4)** and the separation condition,* $\mathrm{BayesRegret}_T^{\mathrm{TS},\Pi} = \mathrm{BayesRegret}_T^{\mathrm{TS},\Pi^\dagger}$, $\mathrm{BayesRegret}_T^{\mathrm{IDS}_0,\Pi} \le (1-p)c_0 + pc_1$, *and* $\mathrm{BayesRegret}_T^{\mathrm{IDS}_0,\Pi} < \mathrm{BayesRegret}_T^{\mathrm{TS},\Pi}$ *for* $T \ge \lceil 1/p \rceil$.

Proof. TS plays $\pi^*(\widetilde{w}) \in \Pi^\dagger$ by (B1), so $\mathrm{BayesRegret}_T^{\mathrm{TS},\Pi} = \mathrm{BayesRegret}_T^{\mathrm{TS},\Pi^\dagger}$. For IDS, let $\pi^\star \in \Pi$ be its episode-1 choice. Since $\pi_P \in \Pi$, $\Psi(\pi^\star) \le \Psi(\pi_P) < \infty$, and (IS-Closure) plus (B4) force $g_1(\pi^\star; \theta) = H_2(p)$. Then $\Delta(\pi^\star) \le (1-p)c_0 + pc_1$, the observation perfectly resolves θ, and $\Psi = 0$ thereafter. The TS lower bound from Theorem 8 (iii) closes the argument. □

Remark 19. The closure condition is needed only when the online candidate class is enlarged beyond $\Pi^\dagger$. Without it, a partially informative policy π_Q with $0 < g_1(\pi_Q; \theta \mid h) < H_2(p_h)$ and slightly smaller immediate regret than π_P can satisfy $\Psi(\pi_Q) < \Psi(\pi_P)$, be selected by IDS, and leave residual uncertainty. The concrete construction in Appendix C satisfies the separation theorem on the restricted candidate class $\Pi^\dagger = \{\pi_0, \pi_1, \pi_P\}$. If the behaviour action O from that construction is also included in the online candidate class, then **(IS-Closure)** need not hold, since O can be partially informative about θ.

C A Concrete Linear-Q Instance

We exhibit an explicit $H = 2$ Bayesian linear-Q instance that satisfies **(B1)–(B4)** for every large enough N. The state-action geometry is $\mathcal{S}_1 = \{s_1\}$, $\mathcal{S}_2 = \{\perp, g_0, g_1, y_+, y_-\}$, $\mathcal{A}(s_1) = \{S, R, P, O\}$, with feature map sending each action to a distinct one-hot vector $\phi(s_1, \cdot) \in \{e_1, e_2, e_3, e_4\} \subset \mathbb{R}^4$ (so $d = 4$). The stage-1 parameter takes the form $\theta_1^{(\theta)} = (1/2, \theta, 1/2 - c, 0)$ with $c \in (0, 1/2)$ and $\theta \in \{0, 1\}$. Transitions are deterministic: $S, R \to \perp$, $P \to g_\theta$ (which itself reveals θ), and under O the next state is y_- with probability 1 if $\theta = 0$ and y_+ with probability $q \in (0, 1)$ (else y_-) if $\theta = 1$. Stage-2 rewards distinguish the two cells.

This finite construction is used only for the structural separation result; the information that R and P each carry about θ is delivered through the reward (for R) and the deterministic next state (for P), and is therefore separate from the linear-Gaussian observation model used in the log-determinant regret bound of Section 4.

Conditional on having observed N no-y_+ trajectories under the behaviour policy that plays only O, the posterior over θ satisfies

$$p_N = \frac{(1-q)^N}{1+(1-q)^N} \to 0.$$

At episode 1 of online IDS,

$$\Delta(S) = p_N/2, \qquad \Delta(R) = (1-p_N)/2, \qquad \Delta(P) = p_N/2 + c, \qquad \Delta(O) = 1/2 + p_N/2,$$

while $g(S) = 0$ and $g(R) = g(P) = H_2(p_N)$, since the reward of R equals θ (revealing θ) and the next state of P is g_θ (also revealing θ). IDS selects P if and only if $p_N/2 + c < (1-p_N)/2$, equivalently $p_N < 1/2 - c$, so $N \geq N_\star(q,c) := \min\{N : p_N < 1/2 - c\}$ is sufficient.

Identifying

$$\pi_0 \equiv S, \qquad \pi_1 \equiv R, \qquad \pi_P \equiv P,$$

the constants of Assumption 14 are read off as

$$\underline{\Delta}_+ = \underline{\Delta}_- = \tfrac{1}{2}, \qquad c_0 = c, \qquad c_1 = \tfrac{1}{2} + c,$$

since $\mathbb{E}[V_1^* - V_1^{\pi_0} \mid \theta = 1] = 1 - 1/2 = 1/2$, $\mathbb{E}[V_1^* - V_1^{\pi_1} \mid \theta = 0] = 1/2 - 0 = 1/2$, $\mathbb{E}[V_1^* - V_1^{\pi_P} \mid \theta = 0] = 1/2 - (1/2 - c) = c$, and $\mathbb{E}[V_1^* - V_1^{\pi_P} \mid \theta = 1] = 1 - (1/2 - c) = 1/2 + c$. Consistently, $\Delta(\pi_0) = p_N \underline{\Delta}_+ = p_N/2$, $\Delta(\pi_1) = (1-p_N)\underline{\Delta}_- = (1-p_N)/2$, and $\Delta(\pi_P) = (1-p_N)c_0 + p_N c_1 = c + p_N/2$. The warm-start hypotheses of Corollary 17 reduce to $c < \underline{\Delta}_- = 1/2$ and $c < (1-e^{-1})(\underline{\Delta}_+ + \underline{\Delta}_-) = 1 - e^{-1} \approx 0.632$, both implied by $c \in (0, 1/2)$. Therefore, for any $c \in (0, 1/2)$ and all $N \geq N_\star(q,c)$, the conclusions of Theorem 8 apply.

D Additional Bandit Results

Full η sweep for the hidden-mode bandit. Table 3 reports the complete η sweep behind Table 1. The main text focuses on IDS_0 because the structural separation theorem is stated for vanilla IDS and relies on the boundary convention $g = 0 < \Delta \Rightarrow \Psi = +\infty$. The sweep shows why this convention matters: when the residual mode probability is extremely small ($N = 1000$), regularised IDS_η with $\eta > 0$ can make the uninformative default action appear cheap, while IDS_0 still probes.

Table 3: Full η sweep on the hidden-mode bandit. Cumulative regret over $T = 500$ online steps, averaged over 10 seeds. A value near 0.15 means the agent probed once and then acted optimally.

N	p_N	greedy	UCB	IDS_0	$\mathrm{IDS}_{0.01}$	$\mathrm{IDS}_{0.05}$	$\mathrm{IDS}_{0.1}$	TS
100	0.377	0.15	0.50	0.15	0.15	0.15	0.15	1.06
200	0.268	0.15	0.15	0.15	0.15	0.15	0.15	1.28
300	0.182	0.15	0.15	0.15	0.15	0.15	0.15	1.29
1000	0.0066	3.31	3.31	0.15	3.31	3.31	3.31	1.79

Full η sweep for the biased contextual bandit. Table 4 reports the complete η sweep behind Table 1. The main text reports $\eta = 0.5$ as a representative regularised IDS setting because it is the strongest choice in the most biased warm-start regimes ($N = 20, 50$). The sweep also shows that the best regularisation level is regime-dependent: when the warm start is more accurate ($N = 100$), a smaller regulariser performs better, while large regularisation over-emphasises immediate regret.

Table 4: Full η sweep for the biased linear contextual bandit. Final cumulative regret over $T = 200$, mean over 20 seeds. Bold marks the best IDS variant per row.

N	greedy	UCB	TS	IDS_0	$\mathrm{IDS}_{0.01}$	$\mathrm{IDS}_{0.05}$	$\mathrm{IDS}_{0.1}$	$\mathrm{IDS}_{0.5}$
20	$42.85_{\pm 34.83}$	$4.80_{\pm 1.37}$	$10.64_{\pm 1.52}$	$7.29_{\pm 1.79}$	$6.17_{\pm 1.42}$	$5.19_{\pm 1.23}$	$4.69_{\pm 1.43}$	$\mathbf{3.57}_{\pm \mathbf{1.67}}$
50	$18.90_{\pm 26.91}$	$1.94_{\pm 1.02}$	$4.73_{\pm 0.76}$	$2.51_{\pm 0.65}$	$2.16_{\pm 0.88}$	$1.90_{\pm 0.91}$	$2.18_{\pm 1.50}$	$\mathbf{1.67}_{\pm \mathbf{1.89}}$
100	$0.88_{\pm 2.08}$	$0.113_{\pm 0.33}$	$0.445_{\pm 0.52}$	$0.152_{\pm 0.27}$	$\mathbf{0.126}_{\pm \mathbf{0.32}}$	$0.194_{\pm 0.59}$	$0.632_{\pm 1.94}$	$0.641_{\pm 1.95}$

Algorithm 1 Off-to-on IDS, deep-ensemble path (used for D4RL). The linear-Gaussian path is identical except that Δ and g use the BLR closed forms.

Inputs. Offline dataset $\mathcal{D}_N$; ensemble size K; noise variance σ^2; gain temperature α; regulariser η; online steps T; candidates per step M; perturbation σ_a; action bounds $[-a_{\max}, a_{\max}]$; clip $q_{\max}$; UTD ratio U; mix ratio ρ.

Stage 1 (offline pretraining).

1. For $i = 1, \ldots, K'$, train $(\pi_\psi^{(i)}, Q^{(i)})$ on $\mathcal{D}_N$ with TD3+BC [Fujimoto and Gu, 2021], with each transition admitted to member i's minibatch with probability $p_{\text{boot}} = 0.9$.
2. Repack the twin heads of each critic as $K = 2K'$ independent ensemble members $\{Q_k\}$. Choose $\pi_\psi := \pi_\psi^{(1)}$ as the anchor actor.
3. Calibrate $\sigma^2 \leftarrow \mathrm{Var}(r + \gamma \bar{Q}(s', \bar{\pi}(s')) - \bar{Q}(s, a))$ and α on an offline holdout, where $\bar{Q}$ is the ensemble mean and $\bar{\pi}$ is the target actor.

Stage 2 (online IDS). For $t = 1, \ldots, T$:

1. Observe s_t. Set anchor $\bar{a} \leftarrow \pi_\psi(s_t)$ and candidates $A_t = \{\bar{a}\} \cup \{\mathrm{clip}(\bar{a} + \sigma_a \xi_m, -a_{\max}, a_{\max})\}_{m=2}^{M}$. Draw a wider set $A_{V^\star}$ of size $M_{V^\star}$ from the same proposal.
2. For each $a \in A_t \cup A_{V^\star}$, query the ensemble at the inference path so that $\widehat{Q}_k(s_t, a) = \mathrm{clip}(Q_k(s_t, a), -q_{\max}, q_{\max})$ (the clip is *not* applied during training forward passes). Compute $\Delta(a) \leftarrow \frac{1}{K} \sum_k (\max_{a' \in A_{V^\star}} \widehat{Q}_k(s_t, a') - \widehat{Q}_k(s_t, a))$, $g(a) \leftarrow \frac{1}{2} \log(1 + \alpha \mathrm{Var}_k[\widehat{Q}_k(s_t, a)]/\sigma^2)$.
3. Pick $a_t \leftarrow \arg\min_{a \in A_t} \Delta(a)^2/(g(a) + \eta)$, add execution noise $\mathcal{N}(0, \sigma_{\text{ex}}^2)$, step the environment.
4. Push $(s_t, a_t, r_t, s_{t+1}, \mathrm{done}_t)$ to the online replay buffer.
5. Slow fine-tune (every step, U gradient steps): sample a batch with mix ratio ρ from the online replay and $1 - \rho$ from $\mathcal{D}_N$; update each Q_k with bootstrap masking against the TD target $r + \gamma \bar{Q}(s', \pi_\psi^{\text{tgt}}(s'))$; update π_ψ with $-\bar{Q}(s, \pi_\psi(s))/|\bar{Q}|_{\text{mean}}$ (no BC term in the online phase); Polyak-update π_ψ^{tgt} with rate τ.
6. Every E env steps, evaluate π_ψ^{tgt} deterministically on n_{eval} rollouts and log the D4RL normalised score.

E Algorithm, Implementation, and Hyperparameters

Path A details (linear contextual bandit). We use a fixed feature map $\phi(s, a) \in \mathbb{R}^d$ constructed as follows: state and action are concatenated with quadratic and interaction terms, projected by a random Gaussian matrix, passed through a $\tanh$ nonlinearity, and ℓ_2-normalised. The resulting feature dimension is $d = 128$.

Rewards follow a linear-Gaussian model

$$r(s, a) = \phi(s, a)^\top w^\star + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

The offline dataset is collected by a mismatched behaviour policy. We initialise a Bayesian linear regression posterior with prior precision $\lambda = 1$ and noise variance $\sigma^2 = 1$, then update it using standard closed-form sufficient statistics:

$$\Lambda = \lambda I + \sum \phi \phi^\top, \quad \mu = \Lambda^{-1} \sum \phi r.$$

During the online phase, the posterior is updated after each observation via rank-1 updates (Cholesky form). All methods share this posterior.

For IDS, we draw $S = 64$ posterior samples to estimate $\Delta(a)$ by Monte Carlo. The information gain $g(a)$ is computed in closed form from the posterior covariance. The candidate set size is $M = 64$.

Path B details (deep ensemble, D4RL). We use a TD3+BC-based ensemble of actor–critic models as a posterior surrogate. Offline training runs $K' = 5$ independent actor–critic pairs with bootstrap

masking probability 0.9 for 5×10^5 gradient steps. Each critic has twin Q-heads, which are repacked into a $K = 10$ ensemble $\{Q_k\}$.

Online, we discard the BLR head and use the ensemble for action evaluation. After each environment step, both the actor and critics are updated using a replay buffer with a $50/50$ mix of offline and online data and an update-to-data ratio of 5.

At each state, a candidate set A_t (size $M = 64$) is sampled around the actor output, and $V_k^\star(s)$ is computed using an enlarged set ($M_{V^\star} = 256$). To stabilise disagreement estimates, we apply inference-time clipping

$$\widehat{Q}_k(s,a) = \mathrm{clip}(Q_k(s,a), \pm q_{\max}), \quad q_{\max} = 10^4.$$

The IDS surrogate uses ensemble variance with scale parameters (σ^2, α) calibrated on a held-out subset of $\mathcal{D}_N$. Target networks use Polyak averaging with $\tau = 0.005$.

Environment hyperparameters (D4RL). Three knobs are tuned per environment; all others are shared. The proposal sigma σ_a controls the candidate spread, the IDS regulariser η trades exploration against $\arg\min \Delta$, and the ensemble Q clip $q_{\max}$ caps OOD extrapolation. The choice $\eta = 0.05$ was robust on every environment except HALFCHEETAH-MEDIUM-REPLAY, where multimodal data raises ensemble disagreement and the ratio $\Delta^2/(g+\eta)$ becomes unstable at small η; raising η to 0.5 recovered stable behaviour.

Table 5: Per-environment hyperparameters for Path B. All other settings are shared.

Environment (-v2)	σ_a	η	$q_{\max}$
walker2d-medium / -replay	0.1	0.05	10^4
hopper-medium / -replay	0.1	0.05	10^4
halfcheetah-medium	0.1	0.05	10^3
halfcheetah-medium-replay	0.1	0.5	10^3

Total env steps $T = 10^6$; warmup $25{,}000$ steps; evaluation every $5{,}000$ env steps over 10 deterministic rollouts using π_ψ^{tgt}; action exploration noise $\mathcal{N}(0, 0.1^2)$; slow fine-tune every step with $U = 5$ gradient updates; learning rate 3×10^{-4} for both ensemble members and actor; online batch size 256; offline-mix ratio $\rho = 0.5$ on HALFCHEETAH and $\rho = 0.0$ on WALKER2D/HOPPER; online BC weight 0; Polyak rate $\tau = 0.005$; policy frequency 2; gain temperature α for g calibrated on a $5{,}000$-transition offline holdout against the empirical TD-residual variance; Q-clamp applied at inference only. The D4RL data is loaded through the original release [Fu et al., 2020] at -v2. Each (env, seed) D4RL run uses one NVIDIA Ada6000 GPU; bandit experiments run on a single CPU under one hour per seed.