# IDENTIFIABILITY AND EXPLORATION IN PERFORMATIVE REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

In performative reinforcement learning, deploying a policy changes the environment it acts in. Repeated-retraining algorithms fit the environment the current policy induces and redeploy, converging to a performatively stable policy. They have no explicit persistent-exploration mechanism, so their data eventually concentrate on the state–action pairs their own policies keep visiting, and we ask what those data can identify. For performative MDPs whose response is an exponential tilt of a base model, the answer is exact in the limit: the data identify the response directions revealed by the pairs that stay visited, and nothing else. The remaining parameter directions form a *blind subspace* that can be computed before any data arrive, from a candidate policy and the known response design. This separates two cases. When no blind direction can change which policy is best, retraining loses coverage but not identifiability, as on the grid-world benchmark used in this literature. When one can, retraining can settle on a policy that its own data cannot refute, and a learner whose cumulative exposure to the revealing pairs stays bounded has linear regret. The test also says what repairs it, and we analyse three remedies with regret guarantees: optimism and posterior sampling, which explore only among plausibly optimal policies yet suffice once rewards are observed, and a structure-directed probe that plays a covering set of revealing pairs and needs only the planner retraining already uses. Empirical results corroborate our results.

## 1 INTRODUCTION

In performative reinforcement learning (Mandal et al., 2023), deploying a policy changes the transitions and rewards of the MDP it acts in. The standard algorithms retrain on the environment the current policy induces and repeat, converging to a performatively stable policy (Mandal et al., 2023; Rank et al., 2024; Mandal and Radanovic, 2025). They have no explicit persistent-exploration mechanism (Brown et al., 2022), and their finite-sample guarantees assume the data cover what they need, through an overlap ratio, a coverage ratio or an occupancy lower bound (Mandal et al., 2023, Asm. 2)(Mandal and Radanovic, 2025, Asm. 4)(Pollatos et al., 2025).

Stability does not say whether the stable policy is good; that depends on what the retraining data can identify. As best response settles, the actions it stops choosing stop producing data. If every aspect of the response that could change the best policy is still revealed by the actions that remain, nothing is lost. If not, retraining can settle on a policy that its own data cannot refute, a possibility noted qualitatively by Hardt and Mendler-Dünner (2025). We give an exact, computable test that tells these cases apart, bound the cost of failing it, and give remedies with guarantees.

The failure concerns learners that stop exploring. Posterior sampling and optimism explore only among policies optimal for some plausible parameter, which in structured bandits can miss informative actions (Russo and Roy, 2018); with rewards observed, the particular separation formalized in Proposition 3 cannot happen here, because a policy's value depends only on the pairs it visits, which its own deployments observe (Appendix N treats the case where it can).

**Contributions.**

- **What converging deployments identify, before any data arrive.** The information accumulated along a converging deployment sequence splits in two: every response direction



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

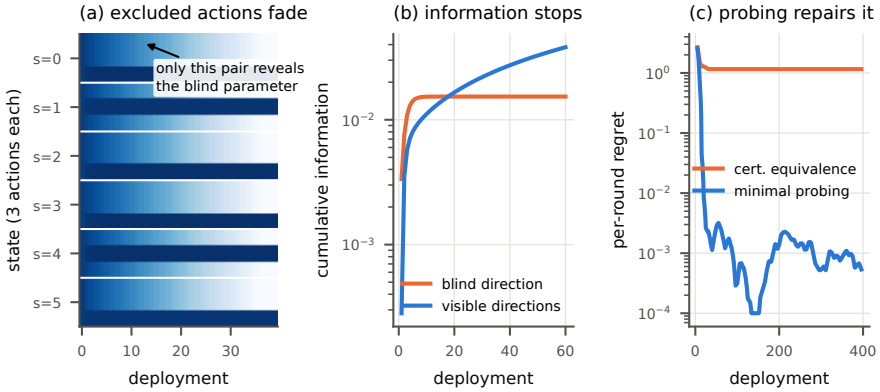


Figure 1: Blind spots in a small performative MDP. (a) How often damped best response visits each state–action pair (rows; three actions per state) over successive deployments, on a log colour scale: actions that best response stops choosing fade geometrically, until one action per state remains. (b) Information gathered about a response parameter that only an abandoned action reveals (orange) stops growing, while the other parameters (blue) keep being learned (Theorem 2). (c) Per-round regret, median over 24 generated environments in which that parameter matters: certainty equivalence stays at a suboptimal policy, while minimal probing (Section 4) recovers.

revealed by the pairs the policies keep visiting is learned at a linear rate, and the rest form a *blind subspace* whose information stays bounded forever and which is computable from a candidate policy and the known response design (Theorems 1 and 2, Corollary 2, Proposition 1). Nullspace characterizations of uninformative policies are classical, in certainty-equivalent control of Markov chains (Borkar and Varaiya, 1979; 1982) and in adaptive LQR (Ziemann and Sandberg, 2020), but there the environment is fixed and only excitation is lost; here the response the learner is trying to identify also decides which pairs are ever seen again. The information geometry extends to any response family smooth in $\epsilon$, with the observation block replaced by a parameter-dependent Jacobian (Appendix L); the exponential tilt additionally makes that block known, so the test can be run in advance.

- **What failing the test costs, and what repairs it.** Stability and identifiability come apart on the path the standard algorithms take: retraining converges to policies whose support contracts to one pair per state, and when a blind direction can change the decision, certainty equivalence settles on a plan that is optimal for a parameter its own data cannot refute, while any learner with bounded exposure to the revealing pairs has linear regret (Proposition 2, Theorem 4). This re-reads the coverage conditions assumed in this literature (Mandal et al., 2023, Asm. 2)(Mandal and Radanovic, 2025, Asm. 4)(Pollatos et al., 2025): they are not regularity conditions, since in the unrestricted class that defines a performative MDP coverage is the entire source of identification (Proposition 7). For repair, observing rewards makes exploring among plausibly optimal plans sufficient (Proposition 3) – the trap that motivates information-directed sampling in structured bandits (Russo and Roy, 2018) does not bind here – so optimism attains $O(\sqrt{T}\log T)$ frequentist regret, posterior sampling attains the same Bayesian-regret rate, and *minimal probing* attains $O(\sqrt{T})$ without the joint policy–parameter optimization oracle required by optimism (Theorems 5–7).
- **Measurements on both sides of the test, including where they cut against us.** On 200 generated environments with five replications, certainty equivalence takes every decision-relevant action in 27% of runs; eight of the nine explicit-exploration baselines reach 99–100% (the count bonus reaches 67%), and its median regret falls from 461 to 45 under minimal probing, a paired difference of $-419$ with 95% interval $[-518, -328]$; on the separation class it recovers in 100% of runs when the unseen effect hurts and 13% when it helps. Optimism nonetheless leads minimal probing by about four regret units, at five times the runtime (Section 5). Beyond the constructed classes we measure how often the failure arises at all: in 500 environments with nothing planted, no blind subspace below the dimension threshold and a decision-relevant one in 66 of 100 above it. Two case stud-

ies show both outcomes: the grid-world used to evaluate performative RL (Mandal et al., 2023) stays decision-relevantly identifiable after retraining abandons 61 of its 256 pairs, as does a battery simulator under the standard plating law, while under a thresholded law the onset turns invisible and probing corrects pessimistic beliefs in 96 hours of excess charging against 244 (Section 5.3).

## 2 SETTING

**Performative MDP.** Let $S, A$ be finite, $\gamma \in (0,1)$, $\rho$ an initial distribution with full support, $P_0$ a full-support base kernel, and $r_0$ a base reward. There are $p$ unknown response parameters $\epsilon_1, \dots, \epsilon_p$. Parameter $j$ comes with a known transition response $B_j \in \mathbb{R}^{S\times A\times S}$ (centred over successors), a reward response $W_j \in \mathbb{R}^{S\times A}$, and a state feature $\phi_j \in \mathbb{R}^S$ that sets how strongly a policy triggers it. Deploying $\pi$ induces

$$\begin{aligned} P^\epsilon_\pi(\cdot \mid s,a) &= \mathrm{softmax}\Big(\log P_0(\cdot \mid s,a) + \textstyle\sum_j \epsilon_j g_j(\pi) B_j(s,a,\cdot)\Big), \\ r^\epsilon_\pi &= r_0 - \kappa \textstyle\sum_j \epsilon_j g_j(\pi) W_j, \end{aligned} \tag{1}$$

with $g_j(\pi) = \langle \phi_j, d^0_\pi \rangle$, where $d^0_\pi = (1-\gamma)\sum_{h\ge 0}\gamma^h \Pr_{P_0,\pi}(s_h = \cdot)$ is the normalized discounted state occupancy of $\pi$ under $P_0$, and $\kappa > 0$. This exponential tilt is a valid kernel for every $\epsilon$. Read it as follows: the policy's aggregate behaviour $g_j(\pi)$ (for example, how much time it spends in some region, weighted by $\phi_j$) scales an effect whose shape on transitions is $B_j$ and on rewards is $W_j$, and the unknown $\epsilon_j$ is the strength of that effect. The tilt is chosen for computability, not generality: Appendix L proves the same factorization and the same identifiability dichotomy for *any* response family smooth in $\epsilon$, with the observation block below replaced by the Jacobian of what a visit reveals, and what the tilt adds is that this block is known and free of $\epsilon$, so the test can be run before any data exist. $J_\epsilon(\pi) = \mathbb{E}[\sum_{h\ge 0}\gamma^h r^\epsilon_\pi(s_h,a_h)]$ is the discounted value of $\pi$ in the environment it induces, and $\pi^\star(\epsilon)$ is a policy maximizing it. We say a policy *visits* a state–action pair (a *pair*) if it takes that action there with positive probability.

**Relation to the standard performative MDP.** The literature assumes only that the deployed policy changes the environment Lipschitz-continuously, in the occupancy (Mandal et al., 2023; Mandal and Radanovic, 2025; Rank et al., 2024; Pollatos et al., 2025) or in the policy (Chen and Huang, 2025; Basu et al., 2025); (1) is a parametric case of that, and meets the policy-level conditions of Chen and Huang (2025) and Basu et al. (2025), whose softmax response has a similar form. It differs from Mandal et al. (2023) and its successors in one respect: their response is driven by the occupancy induced by the response itself, while our trigger uses the occupancy under $P_0$, keeping $g(\pi)$ a known function of the deployed policy. Appendix M treats the induced-occupancy trigger, where $g$ depends on $\epsilon$ too: for a weak enough response the information still factorizes with $D_{g(\pi)}$ replaced by the trigger map's Jacobian, and the blind dimensions carry over. In the unrestricted Lipschitz class, on the other hand, the data bound the response at a pair the learner stops visiting only through the class-wide Lipschitz constant (Appendix L), so no test can separate safe from unsafe cases for that class as a whole: structure is what makes the question answerable, and ours is the favourable case rather than a convenient one.

**Learners.** *Damped best response* iterates $\pi_{t+1} = (1-d)\pi_t + d\,\mathrm{BR}(\pi_t)$, where $\mathrm{BR}(\pi)$ is an optimal deterministic policy of the MDP induced by $\pi$; its fixed points are the *performatively stable* policies $\pi_{\mathrm{PS}}$. *Data-driven retraining* computes BR under the current estimate of $\epsilon$. A *certainty-equivalent* (CE) learner deploys $\pi^\star(\hat\epsilon)$ and re-estimates $\hat\epsilon$ by maximum likelihood from its own data.

**Observation models.** As in the finite-sample performative-RL literature (Mandal et al., 2023; Rank et al., 2024; Mandal and Radanovic, 2025), the learner observes transitions and rewards of the induced MDP, with $\mathcal{N}(0,\sigma_r^2)$ reward noise. We analyse three observation models; the rest of the paper uses (O). (O) *On-policy trajectories*: a deployment of $\pi$ returns one trajectory $(s_0,a_0,r_0,s_1,\dots,s_H)$ of $H$ steps from $s_0 \sim \rho$, and $\bar w_\pi(s,a) = \frac{1}{H}\sum_{h<H}\Pr_\pi(s_h = s)\pi(a \mid s)$ is its time-averaged visitation of pair $(s,a)$. (G) *Generative sampling*: $n$ successors and rewards from every pair. (Q) *Queries*: for every pair, the change in log transition probabilities that the deployment causes, with Gaussian noise of variance $\sigma^2$, and every pair's reward (Appendix D compares the three). Appendix N treats the case where the reward response is not observed.

## 3 INFORMATION AND IDENTIFIABILITY

One visit to a pair $(s, a)$ reveals $\epsilon$ only through how the successor distribution and the reward at $(s, a)$ move with $\epsilon$. Collect this into the *observation block* $O_{sa} = \left[B_{sa};\ (\kappa/\sigma_r)w_{sa}^\top\right] \in \mathbb{R}^{(|S|+1)\times p}$, where $B_{sa}$ has columns $B_j(s, a, \cdot)$ and $w_{sa} = (W_j(s, a))_j$. Let $D_g = \mathrm{diag}(g)$ and $F(P) = \mathrm{diag}(P) - PP^\top$. The true parameter lies in a bounded set, so every induced transition probability is at least some $c_0 > 0$ (Lemma 1). The first result says what one deployment reveals.

**Theorem 1** (Factorization)**.** *For $c \in \{Q, G, O\}$, the Fisher information about $\epsilon$ in one deployment of $\pi$ is $I_c(\pi) = D_{g(\pi)}M_c(\pi)D_{g(\pi)}$ with $M_c(\pi) = \sum_{s,a}\left[\nu_c(s, a)B_{sa}^\top\Lambda_c(s, a)B_{sa} + \nu_c'(s, a)(\kappa/\sigma_r)^2 w_{sa}w_{sa}^\top\right]$, where $\nu_Q = \sigma^{-2}$, $\nu_Q' = 1$, $\Lambda_Q = I$; $\nu_G = \nu_G' = n$, $\Lambda_G = F(P_\pi^\epsilon(\cdot \mid s, a))$; $\nu_O = \nu_O' = H\bar{w}_\pi(s, a)$, $\Lambda_O = F(P_\pi^\epsilon(\cdot \mid s, a))$.*

The policy enters in two ways. It rescales each parameter by $g_j(\pi)$, which cannot change what is identifiable as long as no $g_j(\pi)$ vanishes (an almost-everywhere statement under the measure in Lemma 6), and under on-policy trajectories it also decides which pairs are observed, which can.

**Corollary 1** (What an observation model can reveal)**.** *(i) With $\Gamma_{sa} = B_{sa}^\top F(P_{sa})B_{sa} + (\kappa/\sigma_r)^2 w_{sa}w_{sa}^\top$, $c_0\|O_{sa}u\|^2 \le u^\top\Gamma_{sa}u \le \|O_{sa}u\|^2$. (ii) $\ker M_c(\pi) = \bigcap_{(s,a):\nu_c(s,a)>0}\ker O_{sa}$, so under (O) it depends on $\pi$ only through the pairs $\pi$ visits. (iii) A parameter whose responses vanish on every observed pair gets no information.*

**Converging deployments.** As deployments converge, the parameters split into directions that are learned at a linear rate and a *blind subspace* that is never learned. Let $I_T = \sum_{t\le T} I_O(\pi_t)$ for deployments $\pi_1, \pi_2, \ldots$ fixed in advance rather than adapted to the data (Theorem 3 gives the data-adaptive version).

**Theorem 2** (Identifiability under on-policy deployment)**.** *(a) $\lambda_{\min}(I_T) \to \infty$ if and only if, for every $v \neq 0$, $\sum_t \sum_{s,a} \bar{w}_{\pi_t}(s, a)\,\|O_{sa}D_{g(\pi_t)}v\|^2 = \infty$. (b) Suppose $\pi_t \to \pi_\infty$ and $\sum_t \|g(\pi_t) - g(\pi_\infty)\|_\infty^2 < \infty$. Let $\mathcal{R} = \{(s, a) : \sum_t \bar{w}_{\pi_t}(s, a) = \infty\}$ be the pairs with infinite cumulative expected visitation, and define the blind subspace*

$$\mathcal{N} = \{v :\ O_{sa}D_{g(\pi_\infty)}v = 0 \text{ for all } (s, a) \in \mathcal{R}\}.$$

*Then (i) $\sup_T v^\top I_T v < \infty$ for every $v \in \mathcal{N}$; (ii) $v^\top I_T v \to \infty$ for every $v \notin \mathcal{N}$, linearly if a witnessing pair has $\liminf_t \bar{w}_{\pi_t} > 0$; (iii) $\lambda_{\min}(I_T) \to \infty$ iff $\mathcal{N} = \{0\}$.*

For any policy $\pi$ we write $\mathcal{N}(\pi)$ for the subspace defined this way from the pairs $\pi$ visits. The summability condition in (b) says the policies converge fast enough; damped best response converges geometrically, so it holds. The test is computable before any data arrive: stack the observation blocks of the pairs a policy keeps visiting and take the null space. No estimator is consistent along $\mathcal{N}$ (Corollary 4, Appendix B). For deterministic limits the size of $\mathcal{N}$ is explicit.

**Corollary 2** (Deterministic limits)**.** *If $\pi_t \to \pi_\infty$ deterministic, with summable mass on non-limit actions and $\sum_t \|\pi_t - \pi_\infty\|^2 < \infty$, then $\mathcal{R} = \{(s, \pi_\infty(s))\}_{s\in S}$ and $\dim\mathcal{N} = p - \mathrm{rank}[O_{s,\pi_\infty(s)}D_{g(\pi_\infty)}]_s \ge p - |S|^2$, with equality ($\dim\mathcal{N} = \max(0, p - |S|^2)$) for Lebesgue-almost every $(B, W, \phi)$ and every deterministic limit. This covers damped best response once its best responses stop changing, and a CE learner whose plan becomes fixed.*

**Finite horizons.** The dichotomy is asymptotic. At a finite horizon $T$, what matters is how many directions have less information than an information threshold $\tau$ (corresponding standard deviation $\tau^{-1/2}$): $d_{\mathrm{eff}}(T, \tau) = \#\{k : \lambda_k(I_T) < \tau\}$.

**Proposition 1** (Effective blind dimension)**.** *Suppose the deployed policy equals $\pi_\infty$ after round $t_0$. Let $\mu_1 \le \cdots \le \mu_p$ be the eigenvalues of $I_O(\pi_\infty)$ and $A = I_{t_0}$. Then for $T > t_0$, $\#\{k : (T - t_0)\mu_k + \|A\| < \tau\} \le d_{\mathrm{eff}}(T, \tau) \le \#\{k : (T - t_0)\mu_k < \tau\}$, so $d_{\mathrm{eff}}$ reaches $\dim\mathcal{N}$ only once $T \ge t_0 + \tau/\mu_{\dim\mathcal{N}+1}$ when $\dim\mathcal{N} < p$; if $\dim\mathcal{N} = p$, every $\mu_k$ is zero.*

In practice the smallest nonzero eigenvalue is tiny, so many more directions are effectively blind than structurally blind (Remark 1, Appendix B).

**Consequence for decisions.** A certainty-equivalent learner can stop at a plan that its own data cannot refute.

**Proposition 2** (Self-confirming plans)**.** *Let the estimate be the ridge-penalized MLE ($\lambda > 0$) on all data, and suppose the deployed policy equals a deterministic plan $\hat{\pi}$ from round $t_0$ on. Then the post-$t_0$ likelihood (but not necessarily the earlier likelihood or ridge penalty) is flat along $\mathcal{N}(\hat{\pi})$, and almost surely $\hat{\epsilon}_T \to \epsilon'$ with $\epsilon' - \epsilon \in \mathcal{N}(\hat{\pi})$. If $\pi^\star$ is single-valued and locally constant at $\epsilon'$, the plan is self-confirming, $\pi^\star(\epsilon') = \hat{\pi}$, and if $\hat{\pi}$ is suboptimal its regret grows linearly.*

These rest points are self-confirming equilibria (Fudenberg and Levine, 1993); the same object appears in adaptive control of Markov chains (Borkar and Varaiya, 1982) and misspecified MDPs (Esponda and Pouzo, 2021).

**Theorem 3** (Adaptive deployments)**.** *If an algorithm chooses $\pi_t$, possibly at random, from the data of rounds $1, \dots, t-1$, then (i) the Fisher information of all data through round $T$ is $\mathbb{E}_\epsilon[I_T]$; (ii) Theorem 2(b) holds pathwise on every realization satisfying its hypotheses; and (iii) Corollary 4(a) holds with $C$ any bound on $\mathbb{E}_\epsilon[v^\top I_T v]$.*

Part (ii) is pathwise. It does not by itself bound $\mathbb{E}_\epsilon[v^\top I_T v]$, since the remaining realizations may carry unbounded information, which is why (iii) takes such a bound as a hypothesis.

## 4 What failing the test costs, and what repairs it

**Exploration is necessary.** The *exposure* of a direction $v$, $\mathcal{V}_T(v) = \mathbb{E}\big[\sum_{t\le T}\sum_{h<H} \|O_{s_h a_h} D_{g(\pi_t)} v\|^2\big]$, measures how much the deployed trajectories reveal along $v$; it is at most a constant times the expected probability mass the learner puts on the pairs that reveal $v$.

**Theorem 4** (A learner that does not explore has linear regret)**.** *Let $\epsilon^0$ and $\epsilon^1 = \epsilon^0 + hv$, $\|v\| = 1$, have no common optimal policy, let $\mathrm{reg}_i(\pi) = \max_{\pi'} J_{\epsilon^i}(\pi') - J_{\epsilon^i}(\pi)$ and $\Delta = \min_\pi[\mathrm{reg}_0(\pi) + \mathrm{reg}_1(\pi)] > 0$. Every algorithm that chooses policies from data satisfies*

$$\max_{i\in\{0,1\}} \mathbb{E}_{\epsilon^i}\Big[\sum_{t\le T} \mathrm{reg}_i(\pi_t)\Big] \;\ge\; \frac{\Delta T}{8}\exp\Big(-\frac{h^2}{2}\mathcal{V}_T(v)\Big),$$

*with $\mathcal{V}_T$ computed under $\epsilon^0$. So a learner whose exposure of $v$ stays bounded, such as a retraining or CE learner whose cumulative exposure to the revealing pairs is bounded, has linear regret on one of the two instances.*

**Which exploration escapes.** Posterior sampling and optimism deploy only plans that are optimal for some plausible parameter; in structured bandits that can miss informative actions (Russo and Roy, 2018). Let $\Pi_\Theta$ be the deterministic plans optimal for some $\epsilon$ in a set $\Theta$ of plausible parameters, and $E_v(\pi) = \{(s,a) : O_{sa} D_{g(\pi)} v \neq 0\}$ the pairs at which deploying $\pi$ reveals $v$.

**Proposition 3** (Plausibly optimal plans see what matters)**.** *If $\epsilon^0$ and $\epsilon^1 = \epsilon^0 + hv$ have no common optimal deterministic plan, then at least one endpoint $i \in \{0,1\}$ has the property that every optimal plan $\pi$ at $\epsilon^i$ puts mass on $E_v(\pi)$ and therefore exposes $v$. In particular, if both endpoints lie in $\Theta$, some plan in $\Pi_\Theta$ exposes $v$.*

The reason: a plan's value depends on $\epsilon$ only through the environment at its own pairs, which its deployments observe; if both optimal plans avoided their revealing pairs, their values would not change along $v$, and one would be optimal at both parameters. Thus, with observed rewards, a learner remains blind only if it fails to expose the decision-relevant direction; visiting unrelated pairs does not suffice. This conclusion fails when the plan causing a performative effect on the objective does not observe it (Appendix N).

**Remedies.** We analyse three remedies: optimism, posterior sampling and minimal probing. They use the following assumptions. (B1) The planner maximizes over deterministic policies, and $\pi^\star(\epsilon)$ is constant on a ball of radius $r$ around $\epsilon$. (B2) $\epsilon$ lies in a known ball $\Theta = \{\|\epsilon\| \le S_\Theta\}$, the estimator is the ridge-penalized MLE constrained to $\Theta$, and $c_\Theta > 0$ is the smallest transition probability over $\Theta$. (B3) $\rho$ is uniform and $|g_j(\pi)| \ge g_{\mathrm{lo}} > 0$ for all $\pi$ and $j$. (B4) Every deterministic policy has a cover

(defined after Theorem 5). All three remedies rest on a confidence bound. Let $V_t = \sum_\ell X_\ell^\top X_\ell$ over the steps $\ell$ observed before round $t$, with $X_\ell = O_{s_\ell a_\ell} D_{g(\pi)}$ for the plan $\pi$ deployed at step $\ell$, and $\bar{R} = \bar{\phi}^2 \max_{sa} \|O_{sa}\|^2$ with $\bar{\phi} = \max_j \|\phi_j\|_\infty$.

**Theorem 5** (Confidence bound for on-policy trajectories)**.** *Under (B2), for $\delta \in (0,1)$ and $\lambda > 0$, with probability at least $1-\delta$, for all $t$, $\|\hat{\epsilon}_t - \epsilon\|_{V_t+\lambda I} \le \sqrt{\beta_t(\delta)} := c_\Theta^{-1}\big(\sqrt{2\log(1/\delta) + \log\det(I + V_t/\lambda)} + \sqrt{\lambda} S_\Theta\big)$, and $\log\det(I + V_t/\lambda) \le p \log(1 + n_t \bar{R}/\lambda)$ after $n_t$ observed steps.*

The constant $1/c_\Theta$ makes the bound very conservative: on our benchmark, the confidence radius is about $10^{10}$ times the empirical estimation error in the corresponding norm. Faury et al. (2020) show how to remove most of this dependence for logistic models.

*Minimal probing* (PBS, for probing the blind subspace; Algorithm 1) deploys the CE plan $\hat{\pi}_t$, but at the states of a *cover* it moves probability $\alpha_t$ (the *probe mass*, distinct from the initial distribution $\rho$) onto the cover's actions (the name refers to the near-cheapest cover it probes). A cover of $\hat{\pi}$ is a set $\mathcal{Q}$ of pairs $\hat{\pi}$ does not visit such that $\operatorname{rank}[O_{sa} D_g]_{\mathcal{R}(\hat{\pi}) \cup \mathcal{Q}} = p$: adding them makes every direction visible. Its cost is the sum of the pairs' advantage gaps under the current estimate.

**Proposition 4** (Covers)**.** *(a) Choosing pairs greedily by rank gained per unit cost gives a cover costing at most $\mathsf{H}(|S|) = \sum_{i \le |S|} 1/i$ times the cheapest one (Wolsey, 1982), where costs are clipped below at $c_{\min} > 0$ so that a pair with no estimated gap is the cheapest rather than free, ties go to the first pair in a fixed order, and candidates are the pairs $\hat{\pi}$ does not visit. This approximation factor concerns the clipped estimated cover cost, not its cost under the true parameter. (b) Under a cover, information grows at least in proportion to the probe mass $\sum_t \alpha_t$, and (c) probing costs at most $L_J \sum_t \alpha_t$ in value, where $L_J$ is the Lipschitz constant of $J_\epsilon$ in the policy (Appendix B).*

**Theorem 6** (Minimal probing has $\sqrt{T}$ regret)**.** *Under (B1)–(B4), let PBS re-plan every round, probe any cover of the current blind subspace, and use $\alpha_t = \min(\frac{1}{2}, t^{-1/2})$. With probability at least $1 - 2\delta$, its regret is at most $\Delta_{\max} t^* + 2 L_J \sqrt{T}$, where $\Delta_{\max}$ bounds the per-round regret and $t^* = O(\log^2 T)$ is a burn-in given explicitly in Appendix B. With $\delta = 1/T$ the expected regret is $O(\sqrt{T})$.*

**Corollary 3** (Certified stopping)**.** *Under the assumptions of Theorem 6, with $T$ and $r$ known, stop probing once $\lambda_{\min}(V_t + \lambda I) \ge \bar{\beta}_T(\delta)/r^2$, where $\bar{\beta}_T(\delta)$ is the radius of Theorem 5 with its log-determinant replaced by the deterministic bound at $TH$ steps, so that the threshold is known in advance. With probability at least $1 - 2\delta$, probing stops by $t^*$, the plan is then optimal, and the regret is at most $\Delta_{\max} t^* = O(\log^2 T)$.*

*Optimism* deploys $\pi_t$ from a pair $(\pi_t, \tilde{\epsilon}_t)$ maximizing $J_{\epsilon'}(\pi)$ over deterministic plans and parameters $\epsilon'$ in the confidence set $\mathcal{C}_t = \{\epsilon' \in \Theta : \|\epsilon' - \hat{\epsilon}_t\|_{V_t+\lambda I} \le \sqrt{\beta_t(\delta)}\}$; the theorem assumes an exact joint policy–parameter optimization oracle. *Posterior sampling* deploys $\pi^\star(\tilde{\epsilon}_t)$ for $\tilde{\epsilon}_t$ drawn from the exact posterior under a prior on $\Theta$. Regret is measured against the best deterministic plan, with a fixed tie-breaking rule.

**Theorem 7** (Optimism and posterior sampling)**.** *Assume (B2), $\sigma_r > 0$ and $\lambda = pH\bar{R}$. For fixed $p$ and $\gamma$ and $H = \lceil 2\log T/(1-\gamma) \rceil$, with probability at least $1 - 2/T$ the regret of optimism is $O(\sqrt{T}\log T)$, and the Bayesian regret of posterior sampling is $O(\sqrt{T}\log T)$; neither needs (B1), (B3) or (B4). Appendix B gives the bound with all constants.*

The proof bounds each round's regret by the confidence radius times the information the deployed plan gathers at its own pairs, which is the structure of Proposition 3. Optimism needs a planner over plans and parameters jointly; posterior sampling and minimal probing do not require this joint optimization oracle, and minimal probing also admits certified stopping.

## 5 Experiments

All generated environments have 6 states and 3 actions, $\gamma = 0.9$, and deployments of 540-step trajectories; the learner observes transitions and rewards (reward noise $\sigma_r = 1$) and estimates $\epsilon$ by maximum likelihood. The experiment below uses 100 environments per class and 5 replications. Every comparison's protocol and test were fixed in its script header before the run and are reported

whatever the outcome. The earlier registrations were sequential, later ones designed after earlier outcomes, so as a family they are exploratory; Appendix N gives that history, including the ones that failed, the arm voided by an implementation flaw and the results we withdrew. The experiment reported here is separate: its three primary tests, its environments and its non-inferiority margin were registered before any of its environments were generated, on seeds disjoint from every earlier run, and it is the paper's confirmatory evidence. The unit of analysis is the environment, and tests are paired over environments. Regret is measured against the best policy found by a search that includes $\pi_{\mathrm{PS}}$. Enumerating all $3^6 = 729$ deterministic plans shows that reference is never below the exact optimum over deterministic plans, the class our regret bounds use; a search $20\times$ larger improves it by at most $1.4$ of cumulative regret (Appendix D).

### 5.1 Does the test predict what the data identify?

The structural dimension prediction and finite-horizon bracket matched the corresponding numerical diagnostics in all 27 configurations we checked: at the limit of damped best response the computed blind subspace has exactly $\max(0, p - |S|^2)$ dimensions at $p = 8$, 32 and 64 and in 24 larger cases ($|S| \in \{6, 10, 15\}$), and Proposition 1's finite-horizon bracket contains the observed count in all 24; near the threshold, per-round information on the least-informed direction falls 100-fold at $p = 32$ (Figure 4). On the grid-world benchmark of Mandal et al. (2023), retraining stops visiting 61 of 256 pairs yet the response stays identifiable, because the other agents respond to states and every state stays visited, and the test predicts exactly this (Appendix O).

**How often does this happen when nobody builds it in?** In 500 unplanted draws (100 at each $p \in \{4, 8, 16, 32, 40\}$, $|S| = 6$; pre-registered, Appendix D), the blind subspace was trivial in all 400 below the threshold $|S|^2 = 36$. At $p = 40$ it was nontrivial in 76 of 100, and a blind direction changed the best plan in 66, that is in 87% of those that had one. Under this generator, the failure occurred frequently above the threshold; below it, failure requires additional structure, which the classes below plant and the gated battery supplies physically.

### 5.2 Does exploration repair it, and at what price?

**Environments.** Two generated classes, 100 environments each, 400 deployments, five replications. *Hidden opportunities*: one action at some state would be the best action, but the parameter that shows this is revealed only by taking that action, and the learner starts out believing it is the worst. *Separation*: built so that, if rewards were not observed, a parameter that changes the best decision would show up only at a pair that no plausibly optimal plan visits (with rewards observed, Proposition 3 rules this out); in half the environments the unseen effect helps the plan that ignores it, in half it hurts.

**Methods.** Four groups, specified in Appendix G: *no exploration* (certainty equivalence, CE; data-driven retraining); *undirected* ($\epsilon$-greedy, Boltzmann dithering, a count bonus); *value-directed* (Thompson sampling, sample-based optimism, information-directed sampling, IDS); and *structure-directed* (D-optimal probing; minimal probing with and without its heuristic stopping rule). The pre-registered primary tests were that minimal probing has lower regret than CE on hidden opportunities, and that Thompson sampling and optimism do better on the separation class when rewards are observed than when they are not (Proposition 3).

**Results.** All three primary tests were met (Holm-corrected; Table 1). Effect sizes below are paired Hodges–Lehmann differences over environments with 95% bootstrap intervals, which are not differences of the medians in the table.

- *Without exploration, learners get stuck* (P1). On hidden opportunities, CE's final plan takes every opportunity the true-parameter plan takes in only 27% of runs; eight of the nine explicit-exploration baselines reach 99–100%, while the count bonus reaches 67%; minimal probing has median regret 45 against CE's 461, a difference of $-419$, CI $[-518, -328]$, lower in 88% of environments.
- *The blind spot is orientation-specific.* On the separation class, CE recovers the decision action in 100% of runs when the unseen effect hurts, because its plan then plays the harmful action and observes the harm, but in 13% when the effect helps. Probing with its stopping

Table 1: Median cumulative regret over 400 deployments, and the run-level share whose final plan takes every decision-relevant action that the plan at the true parameter takes (the opportunity actions in one class, the decision actions in the other), here called *recovery*; 100 environments per class and 5 replications, split for the separation class by whether the unseen effect hurts or helps. Lower regret is better. $\epsilon$-greedy reaches those actions in every run but keeps exploring, hence its high regret. Seconds: computation per run, averaged over the two classes.

| | hidden opportunities | | separation | | |
|---|---|---|---|---|---|
| method | regret | recovery | regret | recovery (hurts / helps) | seconds |
| *no exploration* | | | | | |
| certainty equivalence (CE) | 461 | 27% | 94 | 100% / 13% | 28 |
| data-driven retraining | 75 | 90% | 41 | 95% / 94% | 15 |
| *undirected* | | | | | |
| $\epsilon$-greedy | 489 | 100% | 460 | 99% / 100% | 29 |
| Boltzmann dithering | 82 | 99% | 52 | 98% / 100% | 28 |
| count bonus | 130 | 67% | 62 | 97% / 80% | 29 |
| *value-directed* | | | | | |
| Thompson sampling | 48 | 100% | 32 | 98% / 100% | 28 |
| optimism | 36 | 100% | 27 | 97% / 100% | 140 |
| information-directed sampling (IDS) | 40 | 100% | 25 | 99% / 100% | 94 |
| *structure-directed* | | | | | |
| D-optimal probing | 77 | 100% | 71 | 98% / 100% | 40 |
| minimal probing, no stopping | 45 | 100% | 73 | 99% / 100% | 28 |
| minimal probing, heuristic stopping | 38 | 99% | 30 | 98% / 100% | 32 |

rule beats CE there by $-216$, CI $[-313, -139]$ (P2). That rewards are what lets value-directed methods escape at all, as Proposition 3 predicts, is shown by the transitions-only registrations, where Thompson sampling and optimism stay blind (Appendix N).

- *The classical remedy helps but does not close the gap.* Reward-biased maximum likelihood (Kumar and Becker, 1982), which biases estimation toward high-value parameters instead of exploring, cuts CE's regret by $-362$, CI $[-511, -221]$, yet reaches the decision-relevant action in only 75% and 35% of runs on the two classes, against 99–100% for methods that explore (Appendix I, on the 24-environment classes).
- *Minimal probing is close to the best value-directed methods and much cheaper.* With its heuristic stopping rule it has lower regret than CE, data-driven retraining, all three undirected methods and D-optimal probing on both classes, every interval excluding zero. PBS-STOP incurs $4.3$ more regret than optimism, CI $[1.7, 7.3]$, on hidden opportunities and $4.1$ more, CI $[2.4, 6.1]$, on separation. It incurs $4.3$ more than IDS, CI $[2.7, 6.5]$, on separation, while the two are indistinguishable on hidden opportunities ($1.8$, CI $[-0.8, 4.9]$). Those gaps are real but small: the pre-registered non-inferiority test at a margin of 10 regret units, fixed before the run, was met (P3), and probing spends a fifth of optimism's runtime and a third of IDS's. Without the stopping rule it keeps paying for probes it no longer needs, which costs it $42$ on the separation class.

### 5.3 A BATTERY INSTANCE: FAST CHARGING WITH GATED LITHIUM PLATING

A battery charger that keeps its rate below the believed onset of lithium plating learns where the onset is only from charges that plate. This is a mechanistic stress test, not a validation instance of our theory: the simulator is not an instance of the model of §2, so Theorem 2 does not apply. We simulated a 5 Ah cell with an electrochemical model (Sulzer et al., 2021) and the degradation parameters of O'Kane et al. (2022) (Appendix C). We instead ask the theorem's diagnostic question numerically: at the operating points a policy keeps using, does any observable move with the unknown parameter? Read the cases below as that diagnostic. *With the standard plating law, nothing is hidden*: the plating parameters change voltage and capacity even under safe charging, so a cautious charger can learn them. *If lithium deposits only below a nucleation overpotential* $U_{\mathrm{nuc}}$, a thresholded model motivated by Arora et al. (1999) rather than their exact law, *a blind spot appears*: below the onset no observable we measured moves with $U_{\mathrm{nuc}}$, and above it, at 0–15 °C, plating rises sharply.

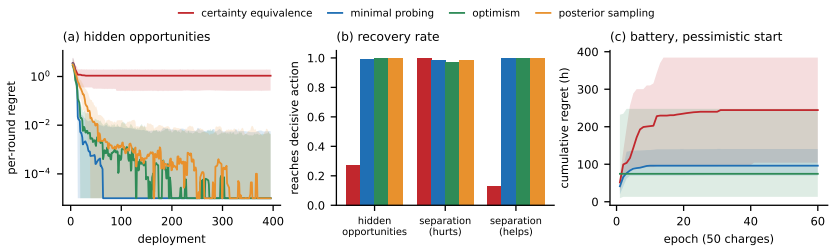


Figure 2: (a) Per-round regret on the hidden-opportunity class: median over the 100 environments, band the interquartile range, smoothed over 9 deployments; curves at the floor are zero. Certainty equivalence does not improve, because the pair that would refute its plan is one it no longer plays. (b) How often each method's final plan takes every decision-relevant action of the plan at the true parameter. Certainty equivalence recovers when the unseen effect hurts, since its own plan then suffers it, and rarely when it helps. (c) The battery loop from pessimistic beliefs: median cumulative regret in hours, interquartile band, over 60 epochs of 50 charges.

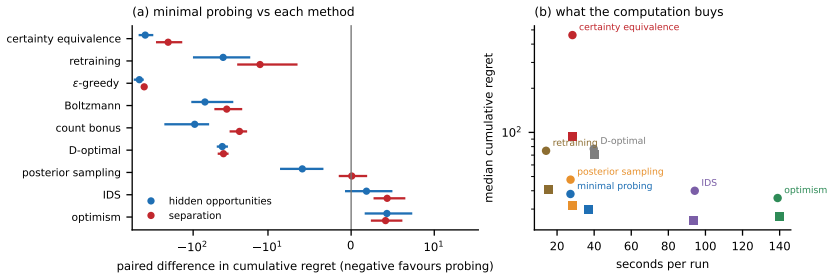


Figure 3: (a) In 500 environments drawn with nothing planted, the share whose blind subspace is nontrivial and the share where a blind direction can change the best plan; below the threshold $|S|^2 = 36$ neither occurs in 400 draws. (b) Certainty equivalence's final-round regret after 200 deployments, each point an environment, bars at the medians: unplanted environments the test clears, unplanted environments where it fires, and the constructed class of §5. The first two are indistinguishable, so a decision-relevant blind direction is not by itself enough to trap retraining; it must also be a direction along which the learner's own estimate settles on the wrong side of the decision boundary (Proposition 2), which is what the constructed class adds.

Invisibility below the onset follows from the gate, so it is a consequence of that assumption, not a finding. In closed loop over 3000 charges, with loss equal to charging time plus a plating penalty, a first registered experiment measured only capacity and was too noisy for any learner to identify $U_{\mathrm{nuc}}$ ($p = 0.11$); a second, registered after that failure, added a detector for the voltage plateau that plated lithium leaves during rest (Petzl and Danzer, 2014; Yang et al., 2018). From pessimistic beliefs CE then corrects slowly (median regret 244 h) while minimal probing corrects within about five epochs (96 h; paired difference $-183$ h, CI $[-304, -98]$; Figure 5b), as do optimism (75 h) and posterior sampling (88 h). This holds at a detection threshold of 0.2 mAh but not 5 mAh.

### 5.4 Discussion and Future Work

We characterized what the data of a converging performative learner can identify: the response directions revealed by the pairs it keeps visiting, and nothing else, with the remainder a blind subspace computable in advance, a dimension law at deterministic limits, and the same geometry for any smooth response family. When a blind direction can change the decision, retraining settles on a plan its own data cannot refute and bounded exposure forces linear regret. A possible line of future inquiry includes: certifying the second condition as cheaply as the first, constants that are not vacuous at practical scale, and a response design estimated rather than assumed.

### AI USE STATEMENT

We used generative AI tools (Claude) to edit manuscript, write and review the code, to draft and check several appendix derivations, and to assist with running and monitoring experiments. The research questions, the method, and the experimental design are the authors'. All derivations were verified by the authors, all numbers are generated from logged results by scripts the authors reviewed, and the authors take full responsibility for the content.

## REFERENCES

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *NeurIPS*.

Abeille, M. and Lazaric, A. (2017). Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2).

Arora, P., Doyle, M., and White, R. E. (1999). Mathematical modeling of the lithium deposition overcharge reaction in lithium-ion batteries using carbon-based negative electrodes. *Journal of The Electrochemical Society*, 146(10):3543–3553.

Basu, D., Das, U., Driss, B., and Mukherjee, U. (2025). Performative policy gradient: Optimality in performative reinforcement learning. arXiv:2512.20576.

Becker, A., Kumar, P. R., and Wei, C.-Z. (1985). Adaptive control with the stochastic approximation algorithm: Geometry and convergence. *IEEE Transactions on Automatic Control*, 30(4):330–338.

Berge, C. (1963). *Topological Spaces*. Oliver and Boyd.

Berkenkamp, F., Turchetta, M., Schoellig, A. P., and Krause, A. (2017). Safe model-based reinforcement learning with stability guarantees. In *NeurIPS*.

Bittanti, S., Bolzern, P., and Campi, M. (1990). Recursive least-squares identification algorithms with incomplete excitation: convergence analysis and application to adaptive control. *IEEE Transactions on Automatic Control*, 35(12):1371–1373.

Borkar, V. and Varaiya, P. (1979). Adaptive control of Markov chains, I: Finite parameter set. *IEEE Transactions on Automatic Control*, 24(6):953–957.

Borkar, V. and Varaiya, P. (1982). Identification and adaptive control of markov chains. *SIAM Journal on Control and Optimization*, 20(4):470–489.

Broder, J. and Rusmevichientong, P. (2012). Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980.

Brown, G., Hod, S., and Kalemaj, I. (2022). Performative prediction in a stateful world. In *AISTATS*, volume 151, pages 6045–6061.

Chen, Z. and Huang, H. (2025). Achieve performatively optimal policy for performative reinforcement learning. arXiv:2510.04430.

Cheng, G., Hardt, M., and Mendler-Dünner, C. (2024). Causal inference out of control: Estimating performativity without treatment randomization. In *ICML*.

Combes, R., Magureanu, S., and Proutiere, A. (2017). Minimal exploration in structured stochastic bandits. In *NeurIPS*.

den Boer, A. V. and Zwart, B. (2014). Simultaneously learning and optimizing using controlled variance pricing. *Management Science*, 60(3):770–783.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, volume 81, pages 160–171.

Esponda, I. and Pouzo, D. (2021). Equilibrium in misspecified markov decision processes. *Theoretical Economics*, 16(2):717–757.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In *ICML*.

Fudenberg, D. and Levine, D. K. (1993). Self-confirming equilibrium. *Econometrica*, 61(3):523–545.

Gevers, M., Bazanella, A. S., Bombois, X., and Mišković, L. (2009). Identification and the information matrix: how to get just sufficiently rich? *IEEE Transactions on Automatic Control*, 54(12):2828–2840, 2009. *IEEE Transactions on Automatic Control*, 54(12):2828–2840.

Gill, R. D. and Levit, B. Y. (1995). Applications of the van trees inequality: a bayesian cramér–rao bound. *Bernoulli*, 1(1–2):59–79.

Gustavsson, I., Ljung, L., and Söderström, T. (1977). Identification of processes in closed loop—identifiability and accuracy aspects. *Automatica*, 13(1):59–75.

Hamidi, N. and Bayati, M. (2020). On frequentist regret of linear Thompson sampling. arxiv:2006.06790, 2020. arXiv:2006.06790.

Hardt, M. (2026). Retraining seeks stable signals. arXiv:2607.15623.

Hardt, M. and Mendler-Dünner, C. (2025). Performative prediction: Past and future. *Statistical Science*, 40(3).

Izzo, Z., Ying, L., and Zou, J. (2021). How to learn when data reacts to your model: Performative gradient descent. In *ICML*.

Jagadeesan, M., Zrnic, T., and Mendler-Dünner, C. (2022). Regret minimization with performative feedback. In *ICML*.

Jiang, R., Chiappa, S., Lattimore, T., György, A., and Kohli, P. (2019). Degenerate feedback loops in recommender systems. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390.

Keskin, N. B. and Zeevi, A. (2018). On incomplete learning and certainty-equivalence control. *Operations Research*, 66(4):1136–1167.

Kilbertus, N., Gomez-Rodriguez, M., Schölkopf, B., Muandet, K., and Valera, I. (2020). Fair decisions despite imperfect predictions. In *AISTATS*, volume 108, pages 277–287.

Kirschner, J. and Krause, A. (2018). Information directed sampling and bandits with heteroscedastic noise. In *COLT*, volume 75, pages 358–384.

Kumar, P. R. (1985). A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, 23(3):329–380.

Kumar, P. R. and Becker, A. (1982). A new family of optimal adaptive controllers for Markov chains. *IEEE Transactions on Automatic Control*, 27(1):137–146.

Kumar, P. R. and Varaiya, P. (1986). *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice-Hall. Reissued as Classics in Applied Mathematics 75, SIAM, 2016.

Lai, T. L. and Robbins, H. (1981). Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 56:329–360.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *KDD*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lin, W., Kumar, P. R., and Seidman, T. I. (1985). Will the self-tuning approach work for general cost criteria? *Systems & Control Letters*, 6(2):77–85, 1985. *Systems & Control Letters*, 6(2):77–85.

Liu, C.-Y. and Li, L. (2016). On the prior sensitivity of Thompson sampling. In *Algorithmic Learning Theory (ALT)*, pages 321–336.

Mandal, D. and Radanovic, G. (2025). Performative reinforcement learning with linear markov decision process. In *AISTATS*, volume 258, pages 3232–3240.

Mandal, D., Triantafyllou, S., and Radanovic, G. (2023). Performative reinforcement learning. In *ICML*.

Miller, J. P., Perdomo, J. C., and Zrnic, T. (2021). Outside the echo chamber: Optimizing the performative risk. In *ICML*.

Nguyen-Tang, T. and Arora, R. (2024). Learning in Markov games with adaptive adversaries: Policy regret, fundamental barriers, and efficient algorithms. In *NeurIPS*.

O'Kane, S. E. J., Ai, W., Madabattula, G., Alonso-Alvarez, D., Timms, R., Sulzer, V., Edge, J. S., Wu, B., Offer, G. J., and Marinescu, M. (2022). Lithium-ion battery degradation: how to model it. *Physical Chemistry Chemical Physics*, 24(13):7909–7922.

Pei, A., Zheng, G., Shi, F., Li, Y., and Cui, Y. (2017). Nanoscale nucleation and growth of electrodeposited lithium metal. *Nano Letters*, 17(2):1132–1139.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *ICML*.

Petzl, M. and Danzer, M. A. (2014). Nondestructive detection, characterization, and quantification of lithium plating in commercial lithium-ion batteries. *Journal of Power Sources*, 254:80–87.

Pollatos, V., Mandal, D., and Radanovic, G. (2025). On corruption-robustness in performative reinforcement learning. In *AAAI*, volume 39, pages 19939–19947.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. SIAM Classics in Applied Mathematics.

Rank, B., Triantafyllou, S., Mandal, D., and Radanovic, G. (2024). Performative reinforcement learning in gradually shifting environments. In *UAI*, volume 244, pages 3041–3075.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Rodemann, J., Fischer-Abaigar, U., Bailie, J., and Muandet, K. (2026). Performative learning theory. arXiv:2602.04402.

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

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

Simchowitz, M. and Foster, D. (2020). Naive exploration is optimal for online lqr. In *ICML*.

Simchowitz, M., Tosh, C., Krishnamurthy, A., Hsu, D., Lykouris, T., Dudík, M., and Schapire, R. E. (2021). Bayesian decision-making under misspecified priors with applications to meta-learning. arXiv:2107.01509.

Sulzer, V., Marquis, S. G., Timms, R., Robinson, M., and Chapman, S. J. (2021). Python battery mathematical modelling (pybamm). *Journal of Open Research Software*.

Wolsey, L. A. (1982). An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4):385–393.

Yang, X.-G., Ge, S., Liu, T., Leng, Y., and Wang, C.-Y. (2018). A look into the voltage plateau signal for detection and quantification of lithium plating in lithium-ion cells. *Journal of Power Sources*, 395:251–261.

Ziemann, I. and Sandberg, H. (2020). On uninformative optimal policies in adaptive lqr with unknown $b$-matrix. arXiv:2011.09288.

*Roadmap.*Appendix K holds the statements moved from the main text and all proofs. Appendix B holds the statements moved from the main text and all proofs. Appendix C documents the battery environments and their closed-loop registrations (§5.3). Appendix D collects the theory checks, the scale studies and the diagnostic and withdrawn experiments (§5.1). Appendix E records the analysis unit, aggregation, intervals, tests, and multiplicity correction, and Appendix F summarizes where the theoretical assumptions bind. Appendix G specifies the exploration methods compared in Table 1 and places minimal probing among them. Appendices H–K discuss coverage assumptions, reward-biased estimation, scale, and the closest classical results. Appendices L and M extend the information geometry beyond the exponential tilt and to induced-occupancy triggers. Appendix N treats the model in which the reward response is not observed, where information and decisions can separate, and reports the nine earlier registrations run in it. Appendix O reports the grid-world benchmark (§5.1).

# A Related work

**Closed-loop identification and incomplete learning.** Feedback destroying identifiability is classical: closed-loop identification (Gustavsson et al., 1977; Gevers et al., 2009); adaptive control of Markov chains, where certainty-equivalent estimates converge to a parameter the limiting closed loop cannot distinguish from the truth (Borkar and Varaiya, 1979; 1982), which makes self-tuning fail for general costs (Lin et al., 1985), shapes the limit set of stochastic-approximation estimates (Becker et al., 1985), and is classically repaired by biasing estimation toward high-value parameters (Kumar and Becker, 1982; Kumar, 1985; Kumar and Varaiya, 1986); least squares with incomplete excitation (Bittanti et al., 1990; Lai and Robbins, 1981); uninformative optimal policies in adaptive LQR, with a $\sqrt{T}$ lower bound (Ziemann and Sandberg, 2020); incomplete learning and minimal exploration in pricing, LQR and structured bandits (Broder and Rusmevichientong, 2012; den Boer and Zwart, 2014; Keskin and Zeevi, 2018; Simchowitz and Foster, 2020; Combes et al., 2017); self-confirming (Fudenberg and Levine, 1993) and misspecified-MDP equilibria (Esponda and Pouzo, 2021); and feedback loops in recommendation, policing and lending (Jiang et al., 2019; Ensign et al., 2018; Lakkaraju et al., 2017; Kilbertus et al., 2020). Appendix K compares the closest of these in detail.

**Performative prediction and RL.** We build on Perdomo et al. (2020); Miller et al. (2021); Izzo et al. (2021); procedures converging to stable points do not attain sublinear performative regret (Jagadeesan et al., 2022), whose confidence-bound algorithm is for prediction rather than MDPs. Stability-seeking performative RL (Mandal et al., 2023; Rank et al., 2024; Mandal and Radanovic, 2025; Pollatos et al., 2025) assumes the coverage it needs rather than characterizing it, and methods targeting the performative optimum (Chen and Huang, 2025; Basu et al., 2025) assume occupancy lower bounds or a known response gradient; optimism attains $\sqrt{T}$ policy regret when the response is state-local and observed (Nguyen-Tang and Arora, 2024). Also related: estimating performativity, what retraining converges to, and its learning-theoretic trade-offs (Cheng et al., 2024; Hardt, 2026; Rodemann et al., 2026).

# B Proofs

Throughout, $\mathcal{N}^\perp$ is the orthogonal complement of a subspace $\mathcal{N}$, $\Pi$ the orthogonal projector onto $\mathcal{N}^\perp$, $\|\cdot\|$ the Euclidean norm (operator norm for matrices), and for policies $\delta(\pi,\pi') = \max_s \frac{1}{2}\|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|_1$. The tilt at $(s,a)$ is $t_{sa} = \sum_j \epsilon_j g_j(\pi) B_j(s,a,\cdot) = B_{sa} D_{g(\pi)}\epsilon$, and the induced reward is $r^\epsilon_\pi(s,a) = r_0(s,a) - \kappa\, w_{sa}^\top D_{g(\pi)}\epsilon$, observed with $\mathcal{N}(0,\sigma_r^2)$ noise independent of everything else given the pair.

## B.1 Statements moved from the main text

**Corollary 4** (The cost of a blind spot)**.** *(a) Suppose the deployment sequence does not depend on $\epsilon$ for $\epsilon$ in a segment $\{\epsilon_0 + \theta v : |\theta| < a\}$, $\|v\| = 1$, and $v^\top I_T(\epsilon) v \le C$ for all $T$, uniformly on the*

*segment (for $v \in \mathcal{N}$, Theorem 2(b)(i) gives this after bounding $\bar{w}_{\pi_t}(s,a) \le \pi_t(a \mid s)$, which does not depend on $\epsilon$). Then every estimator satisfies* $\sup_{|\theta|<a} \mathbb{E}_{\epsilon_0+\theta v}\big[(v^\top(\hat{\epsilon}_T - \epsilon))^2\big] \ge (C + \pi^2/a^2)^{-1}$ *for all $T$. (b) If $\mathcal{N} = \{0\}$ and every pair in $\mathcal{R}$ has $\liminf_t \bar{w}_{\pi_t} > 0$, then $I_T \succeq cT\, I$ for some $c > 0$ and all large $T$.*

**Remark 1** (Why near-blindness is common)**.** $\mu_{\dim \mathcal{N}+1}$ is small for three multiplicative reasons. (i) Block geometry: the retained blocks reduce to an $|S|^2 \times p$ matrix of generic entries whose smallest nonzero singular values sit, as a heuristic (its rows are not identically distributed), near a Marchenko–Pastur hard edge of order $(1 - \sqrt{\psi})^2$, where $\psi = p/|S|^2$ is the aspect ratio, small whenever $p$ is near the threshold. (ii) Feature scale: a coordinate is seen through $g_j^2$, so a spread in $|g_j|$ compresses the spectrum by up to $(\min|g_j| / \max|g_j|)^2$. (iii) Visitation and outcome probabilities weight each block by $\bar{w}\, F$. On our benchmark, (i) alone matches the hard edge within a small factor, while (ii) and (iii) remove another one and a half to four orders of magnitude (§5.1, Appendix D).

**Precise statement of Theorem 6.** PBS re-plans $\hat{\pi}_t = \pi^\star(\hat{\epsilon}_t)$ every round and deploys $\pi_t = \mathrm{mix}(\hat{\pi}_t, \mathcal{Q}_t, \alpha_t)$, which moves probability $\alpha_t = \min(\frac{1}{2}, t^{-1/2})$ onto the actions of $\mathcal{Q}_t$, any cover of $\mathcal{N}(\hat{\pi}_t)$. Let $\kappa_{\min} > 0$ be the smallest value of $\lambda_{\min}(\sum_{(s,a)\in\mathcal{R}(\hat{\pi})\cup\mathcal{Q}} O_{sa}^\top O_{sa})$ over the finitely many deterministic plans $\hat{\pi}$ and their covers $\mathcal{Q}$; $c = (1 - e^{-1})(g_{\mathrm{lo}})^2 \kappa_{\min}/(|S||A|)$; $L_{T,\delta} = \bar{R}\,(p\log(6TH\bar{R}) + \log(1/\delta)) + 1$; $\bar{\beta}_T(\delta)$ the radius $\beta_T(\delta)$ of Theorem 5 with the log-determinant replaced by its deterministic bound at $n_T = TH$; and $t^*$ the first $t$ with $c(2\sqrt{t} - 4) \ge \bar{\beta}_T(\delta)/r^2 + L_{T,\delta}$. Then with probability at least $1 - 2\delta$, $\sum_{t\le T}[J_\epsilon(\pi^\star(\epsilon)) - J_\epsilon(\pi_t)] \le \Delta_{\max} t^* + 2L_J\sqrt{T}$. Since $\bar{\beta}_T$ and $L_{T,\delta}$ are $O(\log T)$ for fixed $\delta$, $t^* = O(\log^2 T)$.

**Precise statement of Theorem 7.** Let $R_{\max}$ bound $|r_\pi^\epsilon|$ over $\Theta$, $J_{\max} = R_{\max}/(1-\gamma)$, $L_V = (\sigma_r + \gamma J_{\max})/(1-\gamma)$, $d_T = p\log_2(1 + TH\bar{R}/\lambda)$ and $M = pH\bar{R}/\lambda$. Assume (B2), $\sigma_r > 0$ and $\lambda \ge \bar{R}$. With probability at least $1 - 2\delta$, optimism's regret satisfies

$$\mathrm{Reg}_T \le 2J_{\max}\, d_T + 2L_V\sqrt{\bar{\beta}_T(\delta)\, T\,\big[2(1-\gamma)(2+M)\big(d_T + \log(1/\delta)\big) + pT\gamma^H\big]},$$

and posterior sampling satisfies the same bound in expectation over $\epsilon$ drawn from the prior, plus $8\delta T J_{\max}$. With $H = \lceil 2\log T/(1-\gamma)\rceil$, $\lambda = pH\bar{R}$ (so $M \le 1$) and $\delta = 1/T$, both are $O(\sqrt{T}\log T)$ for fixed $p$ and $\gamma$.

### B.2 Supporting lemmas

**Lemma 1** (Well-posedness)**.** *For every $\epsilon$ and $\pi$: (i) $P_\pi^\epsilon(\cdot \mid s,a)$ is a probability vector with full support; (ii) $C\big[\log P_\pi^\epsilon(\cdot \mid s,a) - \log P_0(\cdot \mid s,a)\big] = t_{sa}$, where $C = I - \frac{1}{|S|}\mathbf{1}\mathbf{1}^\top$; (iii) if $\|t_{sa}\|_\infty \le b$ then $P_\pi^\epsilon(s' \mid s,a) \ge P_0(s' \mid s,a)e^{-2b} \ge c_0$.*

*Proof.* (i) A softmax of a finite vector has positive entries summing to one. (ii) $\log\mathrm{softmax}(z) = z - (\log\sum_i e^{z_i})\mathbf{1}$ and $C\mathbf{1} = 0$, so the left side is $Ct_{sa} = t_{sa}$, because every $B_j(s,a,\cdot)$ is centred. (iii) $P(s') = P_0(s')e^{t(s')}/\sum_{s''} P_0(s'')e^{t(s'')} \ge P_0(s')e^{-b}/e^b$. □

**Lemma 2** (Monotone sums)**.** *If $A_t \succeq 0$ and $S_T = \sum_{t\le T} A_t$, then $\lambda_{\min}(S_T) \to \infty$ if and only if $v^\top S_T v \to \infty$ for every unit $v$.*

*Proof.* If $\lambda_{\min}(S_T) \to \infty$ then $v^\top S_T v \ge \lambda_{\min}(S_T) \to \infty$. Conversely, $\lambda_{\min}(S_T)$ is nondecreasing; if it stays below $L$, pick unit $v_T$ with $v_T^\top S_T v_T \le L$ and a subsequence with $v_T \to v^*$. For fixed $T_0$ and $T \ge T_0$ in the subsequence, $v_T^\top S_{T_0} v_T \le v_T^\top S_T v_T \le L$; letting $T \to \infty$ gives $v^{*\top} S_{T_0} v^* \le L$ for every $T_0$, so $v^{*\top} S_T v^* \not\to \infty$. □

**Lemma 3** (Features are Lipschitz)**.** $\|d_\pi^0 - d_{\pi'}^0\|_1 \le \frac{2\gamma}{1-\gamma}\delta(\pi,\pi')$ *and* $|g_j(\pi) - g_j(\pi')| \le \|\phi_j\|_\infty \frac{2\gamma}{1-\gamma}\delta(\pi,\pi')$.

*Proof.* With $P_\pi(s,s') = \sum_a \pi(a \mid s)P_0(s' \mid s,a)$, the occupancy is the row vector $d_\pi = (1-\gamma)\rho^\top(I - \gamma P_\pi)^{-1}$, so $d_\pi(I - \gamma P_\pi) = (1-\gamma)\rho^\top$. Subtracting the same identity for $\pi'$ gives

$(d_\pi - d_{\pi'})(I - \gamma P_{\pi'}) = \gamma d_\pi (P_\pi - P_{\pi'})$. Since $(I - \gamma P_{\pi'})^{-1} = \sum_k \gamma^k P_{\pi'}^k$ and a stochastic matrix does not increase the $\ell_1$ norm of a row vector, $\|x(I - \gamma P_{\pi'})^{-1}\|_1 \le \|x\|_1/(1-\gamma)$. Also $\|d_\pi(P_\pi - P_{\pi'})\|_1 \le \sum_s d_\pi(s)\|P_\pi(s,\cdot) - P_{\pi'}(s,\cdot)\|_1 \le \max_s \|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|_1 = 2\delta(\pi,\pi')$. Finally $|\langle \phi_j, d_\pi - d_{\pi'}\rangle| \le \|\phi_j\|_\infty \|d_\pi - d_{\pi'}\|_1$. □

**Lemma 4** (Performative value is Lipschitz). *For fixed $\epsilon$, $J_\epsilon(\pi)$ is Lipschitz in $\pi$ with respect to $\delta$: there is $L_J$ with $|J_\epsilon(\pi) - J_\epsilon(\pi')| \le L_J\, \delta(\pi,\pi')$.*

*Proof.* $J_\epsilon(\pi) = \rho^\top (I - \gamma \tilde{P}_\pi)^{-1} \tilde{r}_\pi$ with $\tilde{P}_\pi(s,s') = \sum_a \pi(a \mid s) P_\pi^\epsilon(s' \mid s,a)$ and $\tilde{r}_\pi(s) = \sum_a \pi(a \mid s) r_\pi^\epsilon(s,a)$. The map $\pi \mapsto g(\pi)$ is rational in the entries of $\pi$, with denominator $\det(I - \gamma \tilde{P}_\pi) \neq 0$ on an open neighbourhood of the compact policy set (the spectral radius of $\gamma P_\pi$ is $\gamma < 1$ on the set and is continuous). $P_\pi^\epsilon$ and $r_\pi^\epsilon$ are smooth functions of $g(\pi)$, and $(I - \gamma \tilde{P}_\pi)^{-1}$ is smooth wherever it exists. So $J_\epsilon$ is continuously differentiable on an open neighbourhood of the compact convex policy set, and the mean value theorem gives a Lipschitz constant on the set. All norms on the finite-dimensional policy space are equivalent, so the constant can be taken with respect to $\delta$. □

**Lemma 5** (Damped best response). *Let $\pi_{t+1} = (1-d)\pi_t + d\,\mathrm{BR}(\pi_t)$ with deterministic* BR. *(i) If $a \neq \mathrm{BR}(\pi_\tau)(s)$ for all $\tau \in [t_0, t)$, then $\pi_t(a \mid s) = (1-d)^{t-t_0}\pi_{t_0}(a \mid s)$. (ii) If $\mathrm{BR}(\pi_\tau) = \pi_\infty$ for all $\tau \ge t_0$, then $\pi_t - \pi_\infty = (1-d)^{t-t_0}(\pi_{t_0} - \pi_\infty)$ for $t \ge t_0$.*

*Proof.* (i) $\mathrm{BR}(\pi_\tau)(a \mid s) = 0$, so $\pi_{\tau+1}(a \mid s) = (1-d)\pi_\tau(a \mid s)$; induct. (ii) $\pi_{t+1} - \pi_\infty = (1-d)\pi_t + d\pi_\infty - \pi_\infty = (1-d)(\pi_t - \pi_\infty)$. □

**Lemma 6** (Genericity). *Let $U \in \mathbb{R}^{|S|\times(|S|-1)}$ have orthonormal columns spanning $\mathbf{1}^\perp$, and parameterize the centred blocks as $B_{sa} = UC_{sa}$ with $C_{sa} \in \mathbb{R}^{(|S|-1)\times p}$. For Lebesgue-almost every $(C, W, \phi)$, every deterministic policy $\pi$ has $g_j(\pi) \neq 0$ for all $j$ and $\mathrm{rank}[O_{s,\pi(s)} D_{g(\pi)}]_{s\in S} = \min\{p, |S|^2\}$.*

*Proof.* Fix a deterministic $\pi$. $g_j(\pi) = \langle \phi_j, d_\pi^0 \rangle$, and $d_\pi^0 \neq 0$ does not depend on $(C, \phi)$, so $\{g_j(\pi) = 0\}$ is a hyperplane in $\phi_j$ and is null. When $D_{g(\pi)}$ is invertible, the stack equals $(I_{|S|} \otimes \mathrm{diag}(U, \kappa/\sigma_r))\,[C_{s,\pi(s)}; w_{s,\pi(s)}^\top]_s\, D_{g(\pi)}$. The left factor has orthogonal columns of nonzero norm, so the rank is that of $[C_{s,\pi(s)}; w_{s,\pi(s)}^\top]_s$, an $|S|^2 \times p$ matrix whose entries are distinct free coordinates, because the pairs $(s, \pi(s))$ are distinct. Its rank falls below the maximum only where every maximal minor vanishes, which is the zero set of a nonzero polynomial and so is null. There are finitely many deterministic policies, and a finite union of null sets is null. □

**Lemma 7** (Divergence of tilted transitions). *For $z \in \mathbb{R}^{|S|}$ and centred $\Delta$, $\mathrm{KL}\big(\mathrm{softmax}(z) \,\|\, \mathrm{softmax}(z+\Delta)\big) \le \frac{1}{2}\|\Delta\|^2$. For rewards, $\mathrm{KL}(\mathcal{N}(m,\sigma_r^2)\|\mathcal{N}(m',\sigma_r^2)) = (m-m')^2/(2\sigma_r^2)$.*

*Proof.* With $A(z) = \log \sum_i e^{z_i}$, $\mathrm{KL}(\mathrm{softmax}(z)\|\,\mathrm{softmax}(z')) = A(z') - A(z) - \nabla A(z)^\top (z' - z) = \int_0^1 (1-u)\Delta^\top \nabla^2 A(z + u\Delta)\Delta\, du$. Since $\nabla^2 A(y) = F(\mathrm{softmax}(y))$ and $\Delta^\top F(P)\Delta = \mathrm{Var}_P(\Delta) \le \max_i \Delta_i^2 \le \|\Delta\|^2$, the integral is at most $\frac{1}{2}\|\Delta\|^2$. □

### B.3 Proof of Theorem 1

Fix $\pi$ and write $P_{sa} = P_\pi^\epsilon(\cdot \mid s,a)$ and $g = g(\pi)$. Because $g$ is computed under $P_0$, it does not depend on $\epsilon$, and $\partial t_{sa}/\partial \epsilon_j = g_j B_j(s,a,\cdot)$. The log-softmax gradient gives the score of one transition from $(s,a)$ to $s'$ as $\nabla_\epsilon \log P_{sa}(s') = D_g B_{sa}^\top (e_{s'} - P_{sa})$. It has mean zero, and since $\mathbb{E}[(e_{s'} - P_{sa})(e_{s'} - P_{sa})^\top] = F(P_{sa})$, its covariance is $D_g B_{sa}^\top F(P_{sa}) B_{sa} D_g$. A reward observed at $(s,a)$ has score $-\kappa D_g w_{sa}(r - r_\pi^\epsilon(s,a))/\sigma_r^2$, with mean zero and covariance $(\kappa/\sigma_r)^2 D_g w_{sa} w_{sa}^\top D_g$, and given the pair it is independent of the successor, so the two scores are uncorrelated. One visit therefore carries $I_{sa} = D_g\big[B_{sa}^\top F(P_{sa}) B_{sa} + (\kappa/\sigma_r)^2 w_{sa} w_{sa}^\top\big] D_g$. The outcome distributions have smooth densities, so the Fisher information equals the covariance of the score.

*(G).* The $n$ draws at each pair are independent, so the information is $n \sum_{s,a} I_{sa}$.

*(O).* A trajectory $x = (s_0, a_0, y_0, \dots, s_H)$, with $y_h$ the observed reward, has density $\rho(s_0)\prod_h \pi(a_h \mid s_h)\varphi_{\sigma_r}(y_h - r^{\epsilon}_{\pi}(s_h, a_h))P_{s_h a_h}(s_{h+1})$, and only the reward and transition factors depend on $\epsilon$. The score is $\sum_h \zeta_h$, with $\zeta_h$ the sum of the reward and transition scores at step $h$. With $\mathcal{F}_h = \sigma(s_0, a_0, y_0, \dots, s_h, a_h)$, $\mathbb{E}[\zeta_h \mid \mathcal{F}_h] = 0$ and $\zeta_h$ is $\mathcal{F}_{h+1}$-measurable. So for $h < k$, $\mathbb{E}[\zeta_h \zeta_k^\top] = \mathbb{E}[\zeta_h \mathbb{E}[\zeta_k^\top \mid \mathcal{F}_k]] = 0$, and the information is $\sum_h \mathbb{E}[\zeta_h \zeta_h^\top] = \sum_h \mathbb{E}[I_{s_h a_h}] = \sum_{s,a}\sum_h \Pr(s_h = s, a_h = a) I_{sa} = H \sum_{s,a} \bar{w}_\pi(s,a) I_{sa}$.

*(Q).* By Lemma 1(ii), the log-ratio observation is $y = X\epsilon + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2 I)$ and $X = [g_1 \mathrm{vec} B_1, \dots, g_p \mathrm{vec} B_p]$, a Gaussian linear model with information $\sigma^{-2} X^\top X$, whose $(j,k)$ entry is

$$g_j g_k \langle \mathrm{vec} B_j, \mathrm{vec} B_k \rangle = g_j g_k \textstyle\sum_{s,a} (B_{sa}^\top B_{sa})_{jk};$$

the independent reward observations add $(\kappa/\sigma_r)^2 D_g \sum_{s,a} w_{sa} w_{sa}^\top D_g$.

In every case $I_c(\pi) = D_g M_c D_g$, and $(D_g M D_g)_{jk} = g_j g_k M_{jk}$, which is $(gg^\top) \odot M$. □

### B.4 Proof of Corollary 1

(i) $x = B_{sa} u$ satisfies $\mathbf{1}^\top x = 0$. Then $x^\top F(P) x = \sum_i P_i x_i^2 - (\sum_i P_i x_i)^2 = \mathrm{Var}_P(x) = \min_c \sum_i P_i (x_i - c)^2$. The minimum is at least $c_0 \min_c \sum_i (x_i - c)^2 = c_0 \|x - \bar{x}\mathbf{1}\|^2 = c_0 \|x\|^2$, and at most $\sum_i P_i x_i^2 \le \|x\|^2$. Adding the reward term $(\kappa/\sigma_r)^2 (w_{sa}^\top u)^2$ to both sides and using $c_0 \le 1$ gives the bounds in terms of $\|O_{sa} u\|^2 = \|B_{sa} u\|^2 + (\kappa/\sigma_r)^2 (w_{sa}^\top u)^2$. (ii) Each summand of $M_c$ is positive semidefinite, so $v \in \ker M_c$ iff every summand with $\nu_c(s,a) > 0$ annihilates $v$ in the quadratic form, iff $O_{sa} v = 0$ for those pairs, by (i) (or trivially when $\Lambda = I$). Under Q and G, $\nu_c > 0$ at every pair; under O, $\nu_O(s,a) = H \bar{w}_\pi(s,a) > 0$ exactly at the pairs visited with positive probability. (iii) Under O, every trajectory with positive probability uses only visited pairs, where neither the tilt nor the reward involves $\epsilon_j$. So the law of the data does not depend on $\epsilon_j$ at all, and in particular $e_j^\top I_c e_j = 0$. Under Q and G every pair is observed, so the hypothesis forces $B_j = 0$ and $W_j = 0$. □

### B.5 Proof of Theorem 2

Write $g_t = g(\pi_t)$, $\bar{w}_t = \bar{w}_{\pi_t}$, $g_\infty = g(\pi_\infty)$, $\beta_O = \max_{s,a} \|O_{sa}\|^2$.

(a) By Theorem 1 and Corollary 1(i), for every $v$,

$$c_0 H \sum_{s,a} \bar{w}_t(s,a) \|O_{sa} D_{g_t} v\|^2 \le v^\top I_O(\pi_t) v \le H \sum_{s,a} \bar{w}_t(s,a) \|O_{sa} D_{g_t} v\|^2.$$

Summing over $t$, $v^\top I_T v \to \infty$ iff the stated series diverges. Apply Lemma 2.

(b)(i) Let $v \in \mathcal{N}$. For $(s,a) \in \mathcal{R}$, $\|O_{sa} D_{g_t} v\| = \|O_{sa}(D_{g_t} - D_{g_\infty})v\| \le \sqrt{\beta_O}\, \|g_t - g_\infty\|_\infty \|v\|$. For $(s,a) \notin \mathcal{R}$, $\|O_{sa} D_{g_t} v\|^2 \le \beta_O \bar{\phi}^2 \|v\|^2$, where $\bar{\phi} = \max_j \|\phi_j\|_\infty$ bounds every $|g_j|$ because $d_\pi^0$ is a probability vector. Using the upper bound in (a) and $\sum_{s,a} \bar{w}_t(s,a) = 1$,

$$\sup_T v^\top I_T v \le H \beta_O \|v\|^2 \Big( \sum_t \|g_t - g_\infty\|_\infty^2 + \bar{\phi}^2 \sum_t \sum_{(s,a) \notin \mathcal{R}} \bar{w}_t(s,a) \Big).$$

The first sum is finite by hypothesis. The second is a finite sum, over finitely many pairs, of series that are finite by definition of $\mathcal{R}$.

(ii) Let $v \notin \mathcal{N}$ and pick $(s,a) \in \mathcal{R}$ with $\kappa_v := \|O_{sa} D_{g_\infty} v\|^2 > 0$. Since $g_t \to g_\infty$, there is $t_1$ with $\|O_{sa} D_{g_t} v\|^2 \ge \kappa_v/2$ for $t \ge t_1$. By the lower bound in (a), $v^\top I_T v \ge \frac{c_0 H \kappa_v}{2} \sum_{t_1 \le t \le T} \bar{w}_t(s,a)$, which diverges because $(s,a) \in \mathcal{R}$. If $\liminf_t \bar{w}_t(s,a) = w > 0$, pick $t_2 \ge t_1$ with $\bar{w}_t(s,a) \ge w/2$ for $t \ge t_2$; the sum is then at least $w(T - t_2)/2$.

(iii) $\mathcal{N}$ is a linear subspace. If $\mathcal{N} = \{0\}$, (ii) and Lemma 2 give $\lambda_{\min}(I_T) \to \infty$. Otherwise, for a unit $v \in \mathcal{N}$, $\lambda_{\min}(I_T) \le v^\top I_T v$, which is bounded by (i). □

### B.6 Proof of Corollary 2

For $a \ne \pi_\infty(s)$, $\bar{w}_t(s,a) \le \pi_t(a \mid s)$, which is summable by hypothesis, so $(s,a) \notin \mathcal{R}$. For $a = \pi_\infty(s)$, the $h = 0$ term gives $\bar{w}_t(s,a) \ge \rho(s)\pi_t(a \mid s)/H \to \rho(s)/H > 0$, so $(s,a) \in \mathcal{R}$ with

positive $\liminf$. By Lemma 3, $\sum_t \|g_t - g_\infty\|_\infty^2 \le c\sum_t \delta(\pi_t,\pi_\infty)^2 < \infty$, since all norms on the policy space are equivalent. Theorem 2(b) therefore applies with $\mathcal{R} = \{(s,\pi_\infty(s))\}_s$. Each $B_{sa}$ has $\mathbf{1}^\top B_{sa} = 0$, so its rank is at most $|S|-1$; the reward row adds at most one, so each $O_{sa}$ has rank at most $|S|$, and the stack of $|S|$ blocks has rank at most $|S|^2$. Equality for almost every $(B,W,\phi)$, simultaneously for every deterministic limit, is Lemma 6; this matters because $\pi_\infty$ itself depends on $(B,W,\phi)$.

For damped best response whose best responses equal $\pi_\infty$ from $t_0$ on, Lemma 5(ii) gives $\delta(\pi_t,\pi_\infty) \le (1-d)^{t-t_0}$ and $\pi_t(a \mid s) \le (1-d)^{t-t_0}$ for $a \ne \pi_\infty(s)$, so both summability hypotheses hold. For a CE learner whose plan equals $\pi_\infty$ from some round on, they hold trivially. □

### B.7 Proof of Corollary 4

(a) Parameterize the segment by $\theta \in (-a,a)$, $\epsilon_\theta = \epsilon_0 + \theta v$. The deployments do not depend on $\theta$, so the data have a law $p_\theta$ that is smooth in $\theta$, with Fisher information $I_T(\theta) = v^\top I_T(\epsilon_\theta) v \le C$. Take the prior $\lambda(\theta) = a^{-1}\cos^2(\pi\theta/2a)$, which vanishes at $\pm a$. Its Fisher information is $J(\lambda) = \int \lambda'^2/\lambda = \frac{\pi^2}{a^3}\int_{-a}^{a} \sin^2(\pi\theta/2a)\,d\theta = \pi^2/a^2$. The van Trees inequality (Gill and Levit, 1995) states that for any estimator $\tilde\theta$ of $\theta$, $\int \mathbb{E}_\theta(\tilde\theta-\theta)^2\lambda(\theta)d\theta \ge \big(\int I_T(\theta)\lambda(\theta)d\theta + J(\lambda)\big)^{-1} \ge (C+\pi^2/a^2)^{-1}$. Apply it with $\tilde\theta = v^\top(\hat\epsilon_T - \epsilon_0)$, for which $\tilde\theta - \theta = v^\top(\hat\epsilon_T - \epsilon_\theta)$. The supremum over $\theta$ is at least the prior average.

(b) $\kappa := \min_{\|v\|=1}\sum_{(s,a)\in\mathcal{R}} \|O_{sa}D_{g_\infty}v\|^2 > 0$, because it is the minimum of a continuous function on the unit sphere that vanishes only on $\mathcal{N} = \{0\}$. The matrices $K_t = \sum_{\mathcal{R}} D_{g_t}O_{sa}^\top O_{sa} D_{g_t}$ converge in norm to $K_\infty$, so $\lambda_{\min}(K_t) \ge \kappa/2$ for $t \ge t_1$. With $w = \min_{\mathcal{R}} \liminf_t \bar w_t > 0$, we have $\bar w_t \ge w/2$ on $\mathcal{R}$ for $t \ge t_2$. The lower bound of Theorem 2(a) then gives $I_O(\pi_t) \succeq \frac{c_0 H w \kappa}{4} I$ for $t \ge t_3 = \max(t_1,t_2)$, hence $I_T \succeq \frac{c_0 H w \kappa}{8} T\, I$ for $T \ge 2t_3$. □

### B.8 Proof of Theorem 3

Let $\mathcal{D}_T = (U,\pi_1,X_1,\dots,\pi_T,X_T)$, where $U$ is the algorithm's internal randomness and $X_t$ the trajectory of round $t$. Its density factors as $p(U)\prod_t \kappa_t(\pi_t \mid U,\mathcal{D}_{t-1})\,p_\epsilon(X_t \mid \pi_t)$, and the policy kernels $\kappa_t$ do not depend on $\epsilon$. The score is therefore $\sum_t s_t$ with $s_t = \nabla_\epsilon \log p_\epsilon(X_t \mid \pi_t)$. Conditional on $(\mathcal{D}_{t-1},\pi_t)$, $s_t$ has mean zero and covariance $I_O(\pi_t)$ (Theorem 1). For $t < u$, $s_t$ is measurable with respect to $(\mathcal{D}_{u-1},\pi_u)$, so $\mathbb{E}[s_t s_u^\top] = \mathbb{E}[s_t\mathbb{E}[s_u^\top \mid \mathcal{D}_{u-1},\pi_u]] = 0$. Hence the Fisher information is $\sum_t \mathbb{E}[s_t s_t^\top] = \mathbb{E}[I_T]$, which is (i). Part (ii) holds because the proof of Theorem 2(b) uses only the realized sequence. For (iii), the van Trees argument of Corollary 4(a) applies to the full data, whose Fisher information along $v$ is $\mathbb{E}[v^\top I_T v]$ by (i); the algorithm is the same map from data to policies for every $\epsilon$. □

### B.9 Proof of Theorem 4

$\Delta > 0$: $\mathrm{reg}_0 + \mathrm{reg}_1$ is continuous on the compact policy set (Lemma 4), and it is positive everywhere, because a zero would be a common optimal policy. Let $P^i$ be the law of the data under $\epsilon^i$ and the same algorithm. In the chain rule for the divergence (Lattimore and Szepesvári, 2020, Chapter 15), the policy kernels, initial states and actions contribute nothing, since they do not depend on $\epsilon$, which leaves $\mathrm{KL}(P^0\|P^1) = \mathbb{E}_0\sum_t\sum_h \big[\mathrm{KL}(P^{\epsilon^0}_{s_h a_h}\|P^{\epsilon^1}_{s_h a_h}) + \mathrm{KL}_{\mathrm{rew}}\big]$. The two tilts at $(s_h,a_h)$ differ by the centred vector $hB_{s_h a_h}D_{g(\pi_t)}v$ and the two reward means by $h\kappa w_{s_h a_h}^\top D_{g(\pi_t)}v$, so Lemma 7 bounds each step's divergence by $\frac{h^2}{2}\|O_{s_h a_h}D_{g(\pi_t)}v\|^2$, and $\mathrm{KL}(P^0\|P^1) \le \frac{h^2}{2}\mathcal{V}_T(v)$.

Fix $t$ and let $A_t = \{\mathrm{reg}_0(\pi_t) \ge \Delta/2\}$, an event determined by the data before round $t$'s trajectory. Since $\mathrm{reg}_0 + \mathrm{reg}_1 \ge \Delta$, $\mathrm{reg}_1(\pi_t) > \Delta/2$ on $A_t^c$. By the Bretagnolle–Huber inequality (Lattimore and Szepesvári, 2020, Theorem 14.2) and data processing,

$$\mathbb{E}_0\mathrm{reg}_0(\pi_t) + \mathbb{E}_1\mathrm{reg}_1(\pi_t) \ge \tfrac{\Delta}{2}\big(P^0(A_t) + P^1(A_t^c)\big) \ge \tfrac{\Delta}{4}e^{-\mathrm{KL}(P^0\|P^1)}.$$

Summing over $t$ and using $\max(a,b) \ge (a+b)/2$ gives the bound. For the exposure bound, let $v$ be any unit vector: $\sum_h \mathbb{E}\|O_{s_h a_h}D_g v\|^2 = H\sum_{s,a}\bar w_\pi(s,a)\|O_{sa}D_g v\|^2$, which vanishes off $E$, while $\|D_g v\| \le \bar\phi$, $\|O_{sa}\|^2 \le \beta_O$ and $\bar w_\pi(s,a) \le \pi(a \mid s)$. □

### B.10 Proof of Theorem 5

Index the steps observed before round $t$ as $k = 1, \dots, n_t$, in order. At step $k$ the pair is $(s_k, a_k)$, the policy in force has features $g_k$, and two things are observed: the successor, coded as $y_k = e_{s'_k}$, and the normalized reward $\tilde{y}_k = (r_k - r_0(s_k, a_k))/\sigma_r$. Write $Z_k = B_{s_k a_k} D_{g_k}$, $\tilde{z}_k = -(\kappa/\sigma_r) D_{g_k} w_{s_k a_k}$, so that the theorem's $X_k = O_{s_k a_k} D_{g_k}$ stacks $Z_k$ and $-\tilde{z}_k^\top$, and $p_k(\theta) = \mathrm{softmax}(\log P_0(\cdot \mid s_k, a_k) + Z_k \theta)$. Let $\mathcal{F}_k$ be generated by everything up to and including the choice of $(s_k, a_k)$ and of the policy in force. Then $\mathbb{E}[y_k \mid \mathcal{F}_k] = p_k(\epsilon)$ and $\tilde{y}_k = \tilde{z}_k^\top \epsilon + \xi_k$ with $\xi_k \sim \mathcal{N}(0,1)$ independent of $y_k$ given $\mathcal{F}_k$.

*Sub-Gaussian increments.* For $w \in \mathbb{R}^{|S|}$, $w^\top y_k$ lies in an interval of length at most $2\|w\|_\infty \le 2\|w\|$, so Hoeffding's lemma gives $\mathbb{E}[\exp(w^\top(y_k - p_k(\epsilon))) \mid \mathcal{F}_k] \le \exp(\|w\|^2/2)$; and $\mathbb{E}[\exp(c\xi_k) \mid \mathcal{F}_k] = \exp(c^2/2)$. By conditional independence, with $w = Z_k u$, $c = \tilde{z}_k^\top u$, $S_n = \sum_{k \le n} [Z_k^\top (y_k - p_k(\epsilon)) + \tilde{z}_k \xi_k]$ and $V_n = \sum_{k \le n} (Z_k^\top Z_k + \tilde{z}_k \tilde{z}_k^\top)$, each $M_n(u) = \exp(u^\top S_n - \frac{1}{2} u^\top V_n u)$ is a nonnegative supermartingale. Note $V_n = \sum_k D_{g_k} O_{s_k a_k}^\top O_{s_k a_k} D_{g_k}$, the $V_t$ of the theorem.

*Method of mixtures.* Integrating $u \sim \mathcal{N}(0, \lambda^{-1} I)$ gives $\bar{M}_n = \det(I + V_n/\lambda)^{-1/2} \exp(\frac{1}{2}\|S_n\|^2_{(V_n + \lambda I)^{-1}})$, again a nonnegative supermartingale with $\bar{M}_0 = 1$ (Tonelli). By Ville's inequality, with probability at least $1 - \delta$, for all $n$, $\|S_n\|^2_{(V_n+\lambda I)^{-1}} \le 2\log(1/\delta) + \log\det(I + V_n/\lambda)$ (Abbasi-Yadkori et al., 2011, Theorem 1).

*From score to parameter.* The log-likelihood, up to terms free of $\theta$, is $\ell(\theta) = \sum_k [\log p_k(\theta)^\top y_k - \frac{1}{2}(\tilde{y}_k - \tilde{z}_k^\top \theta)^2]$. Let $G = \ell - \frac{\lambda}{2}\|\cdot\|^2$, concave, with $\nabla \ell(\theta) = \sum_k [Z_k^\top (y_k - p_k(\theta)) + \tilde{z}_k (\tilde{y}_k - \tilde{z}_k^\top \theta)]$. The constrained maximizer $\hat{\epsilon}$ over the convex set $\Theta$ satisfies $\nabla G(\hat{\epsilon})^\top (\epsilon - \hat{\epsilon}) \le 0$. With $\Delta = \hat{\epsilon} - \epsilon$, the softmax Jacobian gives $p_k(\hat{\epsilon}) - p_k(\epsilon) = \bar{F}_k Z_k \Delta$, where $\bar{F}_k = \int_0^1 F(p_k(\epsilon + u\Delta)) du$, and $\tilde{y}_k - \tilde{z}_k^\top \hat{\epsilon} = \xi_k - \tilde{z}_k^\top \Delta$, so $\nabla G(\hat{\epsilon}) = S_n - \bar{H}\Delta - \lambda\hat{\epsilon}$ with $\bar{H} = \sum_k (Z_k^\top \bar{F}_k Z_k + \tilde{z}_k \tilde{z}_k^\top)$. The optimality condition rearranges to $\Delta^\top (\bar{H} + \lambda I)\Delta \le (S_n - \lambda\epsilon)^\top \Delta$. The segment lies in $\Theta$, so every probability along it is at least $c_\Theta$, and by Corollary 1(i) and $c_\Theta \le 1$, $\Delta^\top \bar{H} \Delta \ge c_\Theta \sum_k \|Z_k \Delta\|^2 + \sum_k (\tilde{z}_k^\top \Delta)^2 \ge c_\Theta \|\Delta\|^2_{V_n}$. Write $B = V_n + \lambda I$. The left side is at least $c_\Theta \|\Delta\|_B^2$. The right side is at most $(\|S_n\|_{B^{-1}} + \lambda\|\epsilon\|_{B^{-1}})\|\Delta\|_B \le (\|S_n\|_{B^{-1}} + \sqrt{\lambda} S_\Theta)\|\Delta\|_B$. Divide by $c_\Theta \|\Delta\|_B$.

*Log-determinant.* By the AM–GM inequality on eigenvalues, $\det(I + V_n/\lambda) \le (1 + \mathrm{tr}\, V_n/(p\lambda))^p$, and $\mathrm{tr}\, V_n = \sum_k \|O_{s_k a_k} D_{g_k}\|_F^2 \le p \max_k \|O_{s_k a_k} D_{g_k}\|^2 n \le pn\bar{R}$, using $\|M\|_F^2 \le p\|M\|^2$ and $\|D_{g_k}\| \le \bar{\phi}$. □

**Lemma 8** (Gram growth under probing)**.** *Under (B3)–(B4) and the deployments of Theorem 6, with probability at least* $1 - \delta$*, for all* $t \le T$*,* $\lambda_{\min}(V_t) \ge c \sum_{\tau < t} \alpha_\tau - L_{T,\delta}$.

### B.11 Proof of Lemma 8

Fix a unit $v$. For round $\tau$ let $x_\tau(v) = \|O_{s_0 a_0} D_{g(\pi_\tau)} v\|^2 \in [0, \bar{R}]$, using that round's first step. Its conditional mean given the past is $m_\tau(v) = \sum_{s,a} \rho(s) \pi_\tau(a \mid s) \|O_{sa} D_{g(\pi_\tau)} v\|^2$. On the cover $\mathcal{C}_\tau = \mathcal{R}(\hat{\pi}_\tau) \cup \mathcal{Q}_\tau$, $\pi_\tau(a \mid s) \ge \alpha_\tau/|A|$, as in Proposition 4(b). Using the round's own features for every pair of its cover, $m_\tau(v) \ge \frac{\alpha_\tau}{|S||A|} (D_{g(\pi_\tau)} v)^\top K_{\mathcal{C}_\tau} (D_{g(\pi_\tau)} v) \ge \frac{\alpha_\tau}{|S||A|} \kappa_{\min}(g_{\mathrm{lo}})^2$, where $K_{\mathcal{C}} = \sum_{(s,a) \in \mathcal{C}} O_{sa}^\top O_{sa}$. For $X \in [0, \bar{R}]$ with conditional mean $m$, convexity gives $e^{-X/\bar{R}} \le 1 - (1 - e^{-1})X/\bar{R}$, hence $\mathbb{E}[e^{-X/\bar{R}}] \le \exp(-(1 - e^{-1})m/\bar{R})$. So $\exp\big(\sum_{\tau < t} [(1 - e^{-1}) m_\tau(v) - x_\tau(v)]/\bar{R}\big)$ is a nonnegative supermartingale started at 1, and by Ville's inequality, with probability at least $1 - \delta'$, for all $t$, $v^\top V_t v \ge \sum_{\tau < t} x_\tau(v) \ge c \sum_{\tau < t} \alpha_\tau - \bar{R} \log(1/\delta')$.

Let $\mathcal{U}$ be an $\eta$-net of the unit sphere with $\eta = 1/(2TH\bar{R})$ and $|\mathcal{U}| \le (3/\eta)^p$. Apply the bound to every $u \in \mathcal{U}$ with $\delta' = \delta/|\mathcal{U}|$. For any unit $v$ and $u \in \mathcal{U}$ with $\|v - u\| \le \eta$, $|v^\top V_t v - u^\top V_t u| \le 2\eta \|V_t\| \le 2\eta TH\bar{R} = 1$. Hence $\lambda_{\min}(V_t) \ge c \sum_{\tau < t} \alpha_\tau - \bar{R}(p \log(6TH\bar{R}) + \log(1/\delta)) - 1$. □

### B.12 Proof of Theorem 6

Work on the intersection of the events of Theorem 5 and Lemma 8, which has probability at least $1 - 2\delta$. For $t \ge 5$, $\sum_{\tau < t} \alpha_\tau \ge \int_4^t x^{-1/2} dx = 2\sqrt{t} - 4$, so for $t \ge t^*$, $\lambda_{\min}(V_t) \ge \bar{\beta}_T / r^2$. Then

$\|\hat{\epsilon}_t - \epsilon\|^2 \le \beta_t/\lambda_{\min}(V_t + \lambda I) \le r^2$, since $\beta_t \le \bar{\beta}_T$ by the deterministic log-determinant bound, and $\hat{\pi}_t = \pi^\star(\epsilon)$ by (B1). For $t \ge t^*$, $\delta(\pi_t, \pi^\star(\epsilon)) \le \alpha_t$, and Lemma 4 bounds the round's regret by $L_J \alpha_t$. Earlier rounds cost at most $\Delta_{\max}$ each. Summing, with $\sum_{t \le T} t^{-1/2} \le 2\sqrt{T}$, gives the bound. Solving for $t^*$ gives $t^* \le \big((\bar{\beta}_T/r^2 + L_{T,\delta})/(2c) + 2\big)^2 + 1 = O(\log^2 T)$. The expectation bound follows by adding $2\delta T \Delta_{\max}$ with $\delta = 1/T$. □

### B.13 Proof of Corollary 3

Either probing has already stopped before $t^*$, or the deployments through $t^*$ are those of Theorem 6 and Lemma 8 applies, giving $\lambda_{\min}(V_{t^*}) \ge \bar{\beta}_T/r^2$. Either way probing stops at some $t_s \le t^*$. $V_t$ is nondecreasing, so on the event of Theorem 5, for every $t \ge t_s$, $\|\hat{\epsilon}_t - \epsilon\|^2 \le \bar{\beta}_T/\lambda_{\min}(V_t + \lambda I) \le \bar{\beta}_T/\lambda_{\min}(V_{t_s} + \lambda I) \le r^2$. So the plan is $\pi^\star(\epsilon)$ and no probing occurs, and every round after $t_s$ has zero regret. Rounds before $t_s$ cost at most $\Delta_{\max}$ each. □

### B.14 Proof of Proposition 1

After $t_0$ every deployment is $\pi_\infty$, so $I_T = A + (T - t_0) I_O(\pi_\infty)$ with $A \succeq 0$. Weyl's inequalities give $(T - t_0)\mu_k \le \lambda_k(I_T) \le (T - t_0)\mu_k + \|A\|$, and counting the indices below $\tau$ gives the sandwich. If no $g_j(\pi_\infty)$ vanishes (generic by Lemma 6, and implied by (B3)), $\ker I_O(\pi_\infty) = \mathcal{N}$ by Corollary 1(ii) and Corollary 2, so exactly $\dim \mathcal{N}$ of the $\mu_k$ vanish. The upper count equals $\dim \mathcal{N}$ once $(T - t_0)\mu_{\dim \mathcal{N}+1} \ge \tau$. □

### B.15 Proof of Proposition 2

Let $\mathcal{N} = \mathcal{N}(\hat{\pi})$, $\hat{g} = g(\hat{\pi})$, and let $\Pi$ project onto $\mathcal{N}^\perp$. The data are on-policy trajectories. We take the Gaussian reward terms of the log-likelihood without their $\epsilon$-free normalizing constants, so the log-likelihood of each trajectory is at most zero. Let $\ell_0$ be the log-likelihood of the data before $t_0$, which is fixed, and $\ell_T = \sum_{t_0 \le t < T} \ell^{(t)}$ that of the later data. Both are concave, being sums of log-softmax terms of affine functions of $\epsilon$ and negative squares of affine functions of $\epsilon$. The estimate maximizes $G_T(\epsilon) = L_0(\epsilon) + \ell_T(\epsilon)$, where $L_0 = \ell_0 - \frac{\lambda}{2}\|\cdot\|^2$.

*Flatness.* After $t_0$, every observed step is at a pair $(s, \hat{\pi}(s))$, and its log-likelihood depends on $\epsilon$ only through $O_{s\hat{\pi}(s)} D_{\hat{g}} \epsilon$. For $u \in \mathcal{N}$ this is unchanged by $\epsilon \mapsto \epsilon + u$. Hence $\ell^{(t)}(\epsilon + u) = \ell^{(t)}(\epsilon)$ for every realization, and $\ell_T(\epsilon) = \ell_T(\Pi\epsilon)$.

*Existence and decomposition.* $G_T$ is strictly concave and tends to $-\infty$ as $\|\epsilon\| \to \infty$, since $\ell_0, \ell_T \le 0$, so it has a unique maximizer $\hat{\epsilon}_T$. For $x \in \mathcal{N}^\perp$ and $u \in \mathcal{N}$, $G_T(x + u) = L_0(x + u) + \ell_T(x)$. Let $h(x) = \max_{u \in \mathcal{N}} L_0(x + u)$ and $u^*(x)$ its maximizer. The maximizer is unique by strict concavity and continuous in $x$ by Berge's maximum theorem (Berge, 1963), using coercivity for local compactness; $h$ is concave and finite. Hence $\hat{\epsilon}_T = \hat{x}_T + u^*(\hat{x}_T)$ with $\hat{x}_T = \arg\max_{x \in \mathcal{N}^\perp} \{h(x) + \ell_T(x)\}$.

*Consistency off $\mathcal{N}$.* Realize the data by drawing, for each of the finitely many deterministic plans $\pi$, an i.i.d. sequence of $\pi$-trajectories in the induced MDP, and letting the $k$-th deployment of $\pi$ use the $k$-th trajectory of its sequence. On the event that $\hat{\pi}$ is deployed from $t_0$ on, the later data are consecutive terms of the $\hat{\pi}$-sequence, which is i.i.d. whether or not $\hat{\pi}$ was chosen from data; apply the following argument to that sequence. Let $n = T - t_0$. For each $x$, $\ell^{(t)}(x)$ is integrable (its reward terms are quadratic in Gaussian noise), so by the strong law $f_n(x) := (h(x) + \ell_T(x))/n \to \mathcal{L}(x) := \mathbb{E}\,\ell^{(t)}(x)$ almost surely. Almost-sure convergence holds simultaneously on a countable dense subset of $\mathcal{N}^\perp$, and because the $f_n$ are concave and $\mathcal{L}$ is finite and concave, it is then uniform on compacts (Rockafellar, 1970, Theorem 10.8).

$\mathcal{L}$ has the unique maximizer $x^* = \Pi\epsilon$ on $\mathcal{N}^\perp$. Indeed, $\mathcal{L}(x^*) - \mathcal{L}(x) = \sum_h \mathbb{E}\big[\mathrm{KL}\big(P_{s_h a_h} \,\|\, P^x_{s_h a_h}\big) + \mathrm{KL}_{\mathrm{rew}}\big] \ge 0$, where $P^x$ is the kernel with tilt $B_{sa} D_{\hat{g}} x$ and the reward term compares means $r_0 - \kappa w_{sa}^\top D_{\hat{g}} \epsilon$ and $r_0 - \kappa w_{sa}^\top D_{\hat{g}} x$. This uses that at visited pairs the true tilt and reward mean depend on $\epsilon$ only through $O_{sa} D_{\hat{g}} \epsilon = O_{sa} D_{\hat{g}} \Pi\epsilon$. Equality requires the two softmaxes and the two reward means to agree at every pair $(s, \hat{\pi}(s))$, all of which are visited because $\rho$ has full support. That means $B_{s\hat{\pi}(s)} D_{\hat{g}}(x - x^*) \in \mathrm{span}(\mathbf{1})$, hence $= 0$ by centring, and $w_{s\hat{\pi}(s)}^\top D_{\hat{g}}(x - x^*) = 0$; so $x - x^* \in \mathcal{N} \cap \mathcal{N}^\perp = \{0\}$.

Now fix $\delta > 0$, let $\eta = \mathcal{L}(x^*) - \max_{\|y-x^*\|=\delta} \mathcal{L}(y) > 0$, and take $n$ large enough that $|f_n - \mathcal{L}| < \eta/3$ on the closed ball of radius $\delta$. Then $f_n(x^*) > f_n(y)$ on the sphere. Any $x$ outside the ball has $y = x^* + \tau(x - x^*)$ on the sphere for some $\tau \in (0, 1)$, and concavity gives $f_n(y) \geq (1-\tau) f_n(x^*) + \tau f_n(x)$, hence $f_n(x) < f_n(x^*)$. So $\|\hat{x}_T - x^*\| < \delta$ eventually, almost surely.

*Limit.* By continuity of $u^*$, $\hat{\epsilon}_T \to \epsilon' := x^* + u^*(x^*)$ almost surely. Then $\epsilon' - \epsilon = u^*(x^*) - (I - \Pi)\epsilon \in \mathcal{N}$, and $\epsilon'$ depends on the data only through $L_0$.

*Rest points.* If $\pi^\star$ is single-valued and constant on a neighbourhood $V$ of $\epsilon'$, then almost surely $\hat{\epsilon}_T \in V$ for all large $T$, so the CE plan is $\pi^\star(\epsilon')$ from some round on. By hypothesis the deployed policy at those rounds is $\hat{\pi}$, so $\pi^\star(\epsilon') = \hat{\pi}$: the plan is optimal for a parameter its own data cannot refute. It is then deployed forever, and each round costs $J_\epsilon(\pi^\star(\epsilon)) - J_\epsilon(\hat{\pi})$. □

**Proposition 4(b),(c), precisely.** (b) Let $\rho$ be uniform, let the base policies converge to a deterministic $\pi$ with support $\mathcal{R}$ and features $g$, let $\mathcal{Q}$ be a cover for $\pi$, and let each deployment move probability $\alpha_t \leq \frac{1}{2}$, $\alpha_t \to 0$, onto the probe actions of $\mathcal{Q}$. Then $I_T = \sum_{t \leq T} I_O(\pi_t)$ satisfies, for some $t_0$ and all $T > t_0$, $\lambda_{\min}(I_T) \geq \frac{c_0 \kappa_\mathcal{Q}}{2|S||A|} \sum_{t_0 < t \leq T} \alpha_t$, where $\kappa_\mathcal{Q} = \lambda_{\min}(\sum_{\mathcal{R} \cup \mathcal{Q}} D_g O_{sa}^\top O_{sa} D_g) > 0$. (c) By Lemma 4, probing costs at most $L_J \sum_{t \leq T} \alpha_t$ in value.

### B.16 Proof of Proposition 4

The greedy rule is the one we run: cost $c(s, a) = \max(\hat{\Delta}(s, a), c_{\min})$ with $c_{\min} = 10^{-3}$ and $\hat{\Delta}$ the estimated advantage gap, candidates $(s, a) \notin \mathcal{R}(\hat{\pi})$, rank gains computed at numerical tolerance $10^{-9}$, and ties broken by a fixed enumeration order. Wolsey's guarantee applies to these clipped costs, so it bounds the estimated cost of the cover, not its cost under the true parameter; a misestimated gap costs efficiency, not information growth, since (b) depends only on the cover's rank. (a) Let $V_x$ be the row space of $O_x D_g$ and $W_\mathcal{Q} = V_\mathcal{R} + \sum_{q \in \mathcal{Q}} V_q$, where $V_\mathcal{R}$ is the sum over the base pairs. Then $f(\mathcal{Q}) = \dim W_\mathcal{Q}$ is nondecreasing and integer-valued. For $\mathcal{A} \subseteq \mathcal{B}$ and $q \notin \mathcal{B}$, $f(\mathcal{A} \cup q) - f(\mathcal{A}) = \dim V_q - \dim(V_q \cap W_\mathcal{A}) \geq \dim V_q - \dim(V_q \cap W_\mathcal{B}) = f(\mathcal{B} \cup q) - f(\mathcal{B})$, because $W_\mathcal{A} \subseteq W_\mathcal{B}$. So $f$ is submodular. A marginal is at most $\dim V_q \leq \operatorname{rank} O_q \leq |S|$, since $\mathbf{1}^\top B_q = 0$ and the reward adds one row. Wolsey's theorem for the greedy algorithm on submodular set cover with integer-valued nondecreasing $f$ (Wolsey, 1982) bounds the greedy cost by $\mathsf{H}\big(\max_q [f(\{q\}) - f(\emptyset)]\big) \leq \mathsf{H}(|S|) \leq 1 + \ln |S|$ times the optimum, where $\mathsf{H}(k) = \sum_{i \leq k} 1/i$.

(b) With $\rho$ uniform, the $h = 0$ term gives $\bar{w}_{\pi_t}(s, a) \geq \pi_t(a \mid s)/(H|S|)$. A probe pair has $\pi_t(a \mid s) \geq \alpha_t/|A|$. A base pair has $\pi_t(a \mid s) \geq (1 - \alpha_t)\pi_t^{\text{base}}(a \mid s) \geq \frac{1}{4}$ once the base puts at least $\frac{1}{2}$ on its limit action. Since $\alpha_t/|A| \leq \frac{1}{4}$, every pair in $\mathcal{R} \cup \mathcal{Q}$ has $\bar{w}_{\pi_t} \geq \alpha_t/(H|S||A|)$ for $t$ large. Because $\alpha_t \to 0$ and the base converges, $\pi_t \to \pi$, so $g_t \to g$ by Lemma 3, and $K_t = \sum_{\mathcal{R} \cup \mathcal{Q}} D_{g_t} O_{sa}^\top O_{sa} D_{g_t} \to K$ with $\lambda_{\min}(K) = \kappa_\mathcal{Q} > 0$ ($\mathcal{Q}$ is a cover). Hence $\lambda_{\min}(K_t) \geq \kappa_\mathcal{Q}/2$ for $t > t_0$. The lower bound of Theorem 2(a) gives, for unit $v$, $v^\top I_O(\pi_t) v \geq c_0 H \frac{\alpha_t}{H|S||A|} v^\top K_t v \geq \frac{c_0 \kappa_\mathcal{Q}}{2|S||A|} \alpha_t$. Sum over $t > t_0$. This is a pathwise statement about the accumulated information of the realized deployments, and it holds whether or not the base policies were chosen from data.

(c) Mixing moves at most $\alpha_t$ probability at each state, so $\delta(\pi_t, \pi_t^{\text{base}}) \leq \alpha_t$, and Lemma 4 bounds the value lost in round $t$ by $L_J \alpha_t$. □

### B.17 Proof of Proposition 3

Fix a deterministic plan $\pi$ and write $g = g(\pi)$, which does not depend on $\epsilon$. Its value is $J_\epsilon(\pi) = \rho^\top (I - \gamma \tilde{P}_\pi)^{-1} \tilde{r}_\pi$ with $\tilde{P}_\pi(s, \cdot) = P_\pi^\epsilon(\cdot \mid s, \pi(s))$ and $\tilde{r}_\pi(s) = r_\pi^\epsilon(s, \pi(s))$, so it depends on $\epsilon$ only through the tilts $B_{s\pi(s)} D_g \epsilon$ and the reward means $r_0(s, \pi(s)) - \kappa w_{s\pi(s)}^\top D_g \epsilon$, $s \in S$. If $\pi$ puts no mass on $E_v(\pi)$, then $O_{s\pi(s)} D_g v = 0$ for every $s$, both quantities are unchanged by $\epsilon \mapsto \epsilon + \theta v$, and $J_{\epsilon + \theta v}(\pi) = J_\epsilon(\pi)$ for all $\theta$. Suppose optimal plans $\pi^0$ at $\epsilon^0$ and $\pi^1$ at $\epsilon^1 = \epsilon^0 + hv$ both put no mass on their own $E_v$. Then $J_{\epsilon^1}(\pi^0) = J_{\epsilon^0}(\pi^0) \geq J_{\epsilon^0}(\pi^1) = J_{\epsilon^1}(\pi^1) = \max_{\pi'} J_{\epsilon^1}(\pi')$, so $\pi^0$ is optimal at $\epsilon^1$, contradicting the hypothesis. Hence all optimal plans at $\epsilon^0$, or all optimal plans at $\epsilon^1$, put mass on their own $E_v$. Finally, a plan $\pi$ with $(s, \pi(s)) \in E_v(\pi)$ has $\bar{w}_\pi(s, \pi(s)) \geq \rho(s)/H > 0$ from the first step, because $\rho$ has full support, so each deployment of it contributes at least $\rho(s) \|O_{s\pi(s)} D_g v\|^2 > 0$ to $\mathcal{V}_T(v)$. □

### B.18 Proof of Theorem 7

*Value difference.* Fix a deterministic plan $\pi$ with features $g$, parameters $\epsilon, \epsilon' \in \Theta$, and write $\Delta = \epsilon' - \epsilon$, $u_s = O_{s\pi(s)} D_g \Delta$. With $P, r$ and $P', r'$ the induced kernels and rewards along $\pi$ and $V' = (I - \gamma P')^{-1} r'$, the simulation identity $V' - V = (I - \gamma P)^{-1}[(r' - r) + \gamma(P' - P)V']$ gives $J_{\epsilon'}(\pi) - J_\epsilon(\pi) = \frac{1}{1-\gamma} \sum_s d(s) \big[(r' - r)(s) + \gamma(P' - P)(s, \cdot)V'\big]$, where $d = (1-\gamma)\rho^\top (I - \gamma P)^{-1}$ is a probability vector. The reward term is $-\kappa w_{s\pi(s)}^\top D_g \Delta$, of absolute value at most $\sigma_r \|u_s\|$. The two tilts differ by the centred vector $B_{s\pi(s)} D_g \Delta$, so by Lemma 7 and Pinsker's inequality $\|P'(s, \cdot) - P(s, \cdot)\|_1 \le \|B_{s\pi(s)} D_g \Delta\| \le \|u_s\|$, and $|(P' - P)(s, \cdot)V'| \le \|u_s\| J_{\max}$. Hence, by Jensen's inequality,

$$|J_{\epsilon'}(\pi) - J_\epsilon(\pi)| \le L_V \sum_s d(s) \|u_s\| \le L_V \|\Delta\|_{G_\pi}, \qquad G_\pi = \sum_s d(s) D_g O_{s\pi(s)}^\top O_{s\pi(s)} D_g.$$

*Occupancy.* Write $W_\pi = \sum_s \bar{w}_\pi(s, \pi(s)) D_g O_{s\pi(s)}^\top O_{s\pi(s)} D_g$, so that $H W_{\pi_t}$ is the conditional mean of the round's Gram increment $X_t$. Since $d(s) = (1 - \gamma) \sum_k \gamma^k \Pr(s_k = s) \le (1 - \gamma) H \bar{w}_\pi(s, \pi(s)) + \mathrm{tail}(s)$ with $\sum_s \mathrm{tail}(s) \le \gamma^H$, and every summand is at most $\bar{R} I$, $G_\pi \preceq (1 - \gamma) H W_\pi + \gamma^H \bar{R} I$.

*Per-round regret.* Work on the event of Theorem 5, so $\epsilon \in \mathcal{C}_t$ for all $t$. Optimism gives $J_\epsilon(\pi^\star) \le J_{\tilde{\epsilon}_t}(\pi_t)$, so the regret of round $t$ is at most $J_{\tilde{\epsilon}_t}(\pi_t) - J_\epsilon(\pi_t) \le L_V \|\tilde{\epsilon}_t - \epsilon\|_{G_{\pi_t}}$. With $A_t = V_t + \lambda I$ and $\|\tilde{\epsilon}_t - \epsilon\|_{A_t} \le 2\sqrt{\beta_t}$, $\|\tilde{\epsilon}_t - \epsilon\|_{G_{\pi_t}}^2 \le 4\beta_t \operatorname{tr}(A_t^{-1} G_{\pi_t}) \le 4\bar{\beta}_T \big[(1 - \gamma) b_t + p\gamma^H\big]$, where $b_t = \operatorname{tr}(A_t^{-1} H W_{\pi_t})$ and we used $\bar{R}/\lambda \le 1$.

*Potential.* Split round $t$'s increment into its steps, $X_t = \sum_h x_{t,h}$, with $A_{t,h}$ the matrix before step $h$. Since $\|A_{t,h}^{-1/2} x_{t,h} A_{t,h}^{-1/2}\| \le \bar{R}/\lambda \le 1$ and $\log_2(1 + a) \ge a$ on $[0, 1]$, $\sum_{t,h} \operatorname{tr}(A_{t,h}^{-1} x_{t,h}) \le \log_2 \det(A_{T+1}) / \det(\lambda I) \le d_T$, using $\operatorname{tr} V_{T+1} \le THp\bar{R}$ and the AM–GM inequality. Call a round *bad* if $\det A_{t+1} > 2 \det A_t$; there are at most $d_T$ of them, and each costs at most $2J_{\max}$. On a good round, $A_{t,h} \preceq A_{t+1} \preceq 2A_t$ (every eigenvalue of $A_t^{-1/2} A_{t+1} A_t^{-1/2}$ is at least one and their product is at most two), so $a_t := \operatorname{tr}(A_t^{-1} X_t) \le 2 \sum_h \operatorname{tr}(A_{t,h}^{-1} x_{t,h})$, and $\sum_{\text{good}} a_t \le 2d_T$. Every $a_t$ lies in $[0, M]$, because $\operatorname{tr} X_t \le Hp\bar{R}$, so $\sum_t a_t \le (2 + M) d_T$.

*From realized to expected.* $b_t = \mathbb{E}[a_t \mid \text{past}]$, and for $X \in [0, M]$ convexity gives $\mathbb{E} e^{-X/M} \le \exp(-(1 - e^{-1}) \mathbb{E}X/M)$, so $\exp(\sum_{\tau<t}[(1 - e^{-1}) b_\tau - a_\tau]/M)$ is a nonnegative supermartingale. By Ville's inequality, with probability at least $1 - \delta$, $\sum_t b_t \le (\sum_t a_t + M \log(1/\delta))/(1 - e^{-1}) \le 2(2 + M)(d_T + \log(1/\delta))$.

*Sum.* On both events (probability at least $1 - 2\delta$), summing the per-round bound over good rounds and applying Cauchy–Schwarz, $\mathrm{Reg}_T \le 2J_{\max} d_T + 2L_V \sqrt{\bar{\beta}_T \, T \sum_t [(1 - \gamma) b_t + p\gamma^H]}$, which is the claim ($\beta_t \le \bar{\beta}_T$ by the log-determinant bound of Theorem 5).

*Posterior sampling.* Let $U_t(\pi) = \max_{\epsilon' \in \mathcal{C}_t} J_{\epsilon'}(\pi)$, which depends only on past data. The filtration of Theorem 5 allows the policy in force, here chosen from the sample $\tilde{\epsilon}_t$, to depend on its own randomness. Given the past, $\pi_t$ and $\pi^\star(\epsilon)$ have the same distribution under exact posterior sampling, so $\mathbb{E}[J_\epsilon(\pi^\star) - J_\epsilon(\pi_t)] = \mathbb{E}[U_t(\pi_t) - J_\epsilon(\pi_t)] + \mathbb{E}[J_\epsilon(\pi^\star) - U_t(\pi^\star)]$ (Russo and Roy, 2014). Theorem 5 holds for every $\epsilon \in \Theta$, hence for $\epsilon$ drawn from the prior. On the two events above, the second term is at most zero and the first obeys the argument for optimism, with $\tilde{\epsilon}_t$ the maximizer defining $U_t(\pi_t)$ and bad rounds costing at most $2J_{\max}$. Off the events, which have probability at most $2\delta$, each round costs at most $4J_{\max}$. □

## C Battery instance

This appendix supports §5.3: the electrochemical model, the two plating laws, the closed-loop registrations, and the earlier stylized model. The simulator is not an instance of the model of §2, and none of the results of §3 are applied to it. The checks below ask the question those results answer, numerically: holding the charging policy fixed, we vary $U_{\mathrm{nuc}}$ over its grid and measure whether any observable the charger records changes by more than the solver and measurement noise.

**Electrochemical model.** PyBaMM 26.8 (Sulzer et al., 2021), Doyle–Fuller–Newman model with the O'Kane et al. (2022) LG M50 parameters (5 Ah), all of their degradation options (solvent-diffusion-limited SEI, SEI on cracks, particle cracking, stress-driven loss of active material, partially reversible plating) and a lumped thermal model. Each table entry is one full cycle from the discharged state: constant-current charge at rate $c$ to 4.2 V, constant-voltage hold to C/20, rest, 1C discharge, rest; losses are irreversible lithium-inventory losses, and the numerical mass-balance residual of the lithium budget stays below $10^{-18}$ Ah, against a solver tolerance of $10^{-8}$.

**Standard plating law: the test passes.** In O'Kane et al. (2022)'s plating law, plated lithium has an equilibrium at every anode potential, so a reversible reservoir forms even under safe charging. Across plating rate constants 0.1–10× nominal, the constant-voltage tail current shifts by 84–155 mA and per-cycle discharge capacity by 2–20 mAh at safe rates, well above assumed measurement noise (10 mA, 0.5 mAh): a charger that never plates can still learn the plating parameters.

**Gated plating law: the test fails.** We replaced the plating law by one in which lithium deposits only when the plating overpotential is below a nucleation overpotential $U_{\mathrm{nuc}} \leq 0$ (Arora et al., 1999), with a softplus gate of width 1 mV; stripping acts only on plated lithium. Nucleation overpotentials of tens of mV are reported for lithium on copper (Pei et al., 2017); we take $U_{\mathrm{nuc}} \in [-60, -5]$ mV. Across that range, at 13 safe operating points, terminal voltage differs by at most $3.5 \times 10^{-5}$ mV, constant-voltage current by $1.3 \times 10^{-3}$ mA (one point: 1.0 mA, from event timing) and discharge capacity by 0.023 mAh. Above the onset, dead-lithium loss rises by six to eight orders of magnitude within 0.1–0.2C and decreases strictly as $U_{\mathrm{nuc}}$ becomes more negative. Self-heating keeps plating negligible at 25 °C and above, so the closed loop uses 0–15 °C, where plating reaches 0.06 mAh of dead lithium and 20 mAh of reversibly plated lithium per cycle. Invisibility below the onset follows from the gating assumption; the simulation confirms it and measures its size.

**Closed loop.** Nine hypotheses for $U_{\mathrm{nuc}}$, truths in $\{-10, \ldots, -50\}$ mV, starts pessimistic (−5 mV), correct or optimistic (−60 mV); 60 epochs of 50 charges with ambient temperature uniform on $\{0, 5, 10, 15\}$ °C; per-charge loss = charge time + 1000 h per mAh of dead lithium; regret against the plan that knows $U_{\mathrm{nuc}}$. Arms: CE; probing one rate above the plan at rate $\rho_e = \min(\frac{1}{2}, e^{-1/2})$ (PBS); probing only while plausible hypotheses disagree on the plan (PBS-STOP); optimism over the plausible set; posterior sampling from the exact posterior. Both versions were pre-registered; v2's measurement was chosen after v1's result.

| median regret (h), 50 runs per start | CE | PBS | PBS-STOP | OPT | TS |
|---|---|---|---|---|---|
| v1 (capacity checks), pessimistic | 1524 | 1464 | 1466 | 4409 | 1619 |
| v1, correct | 1334 | 1417 | 1416 | 4409 | 1619 |
| v2 (+ plateau detector), pessimistic | 244 | 172 | 96 | 75 | 88 |
| v2, correct | 49 | 111 | 43 | 75 | 88 |
| v2, optimistic | 90 | 121 | 85 | 75 | 88 |

*v1*: every observation is an epoch capacity check with 0.5 mAh noise. At the true plan, neighbouring hypotheses differ by 0.004–0.1 mAh per epoch, so no learner identifies $U_{\mathrm{nuc}}$ (final estimate correct in 14–42% of runs; CE started at the truth drifts away in 86%). Primary test (PBS-STOP < CE, pessimistic) not met, $p = 0.11$; optimism pays 6082 h of plating penalty, and PBS-STOP has lower regret than optimism ($p = 1.3 \times 10^{-4}$). *v2*: after each charge a plateau detector fires with probability $0.01 + 0.99\,\Phi((\ln q - \ln q_{\mathrm{det}})/0.5)$ for reversibly plated lithium $q$, with $q_{\mathrm{det}} = 1$ mAh assumed, since the plateau appears only above a threshold (Yang et al., 2018) whose value for this cell is not published. Primary met ($p = 3.6 \times 10^{-7}$). Secondary tests: no difference from CE at correct starts ($p = 0.23$); at optimistic starts a small difference (median +1 h, $p = 0.003$), against the prediction of none; PBS-STOP lower than PBS ($p = 2 \times 10^{-7}$); no difference from optimism ($p = 0.77$) or posterior sampling ($p = 0.55$); final estimate correct in 84–94% of runs for all arms. Detection threshold 0.2 mAh: met ($p = 2 \times 10^{-5}$); 5 mAh: not met ($p = 0.17$).

## D ADDITIONAL EXPERIMENTS

Checks of the theory beyond those summarized in §5.1, the scale studies behind Proposition 1, a comparison of observation models, and two experiments we withdrew. Unless marked as standard-model, the checks in this appendix observe transitions only (the model of Appendix N).

**Theory checks in full.** The factorization matches $4 \times 10^4$ simulated trajectories to $0.7\%$ in maximum entrywise relative error and the analytic score matches finite differences to $8 \times 10^{-10}$. For the blind subspace, damped best response ($d = 0.5$, 400 rounds) on the dense benchmark matches the predicted blind dimensions 0, 2 and 34 at $p = 8$, 32 and 64 (Table 2 in Appendix D).

Where best response settles, the predicted dimension matches the directions whose information stops growing, up to one or two that grow slowly; a finite horizon cannot separate bounded from slowly growing. Where it cycles, the theorem's convergence hypothesis fails and the prediction need not hold: with seed 1 the cycle stays inside the predicted support and the two blind directions form anyway, while with seed 2 it visits more actions and none form. Along the damped-best-response path over 108 environments (Figure 4), per-round information is, with transitions only, flat below the threshold $p = |S|(|S| - 1) = 30$, even as the least-visited pair's weight falls from $4.2 \times 10^{-2}$ to $7.6 \times 10^{-7}$, and collapses above it. The median factor over 15 rounds is $5.7 \times 10^{-4}$ at $p = 32$ and $3.6 \times 10^{-5}$ at $p = 64$. Slower damping keeps more information, by a factor 1.2 below the threshold and 1.5–1.8 above it. At larger scale ($|S| \in \{6, 10, 15\}$, $|A| \in \{3, 5\}$, $p$ four below and four above $|S|(|S|-1)$; Appendix D), the predicted blind dimension at the limit equals $\max(0, p-|S|(|S|-1))$ in all 24 cases. But the number of directions whose information stops growing over 300 rounds is larger in 22 of the 24 cases and grows with $|S|$: up to 31 at $|S| = 15$ where the law predicts 4, and 0–19 below the threshold. The effective dimension of Proposition 1 accounts for this exactly. Its bracket contains the observed count in all 24 cases, and at precision $\sigma = 0.1$ the horizon needed to reach the structural count is $10^6$–$10^{11}$ deployments. Removing the feature scaling and visitation weights puts the smallest eigenvalue back near the Marchenko–Pastur edge; with them it is one and a half to four orders of magnitude smaller (Remark 1, Appendix D). Figure 1b shows Theorem 2(b)(i) directly: a coordinate visible only through an excluded pair saturates, at a level that scales roughly as $1/d$.

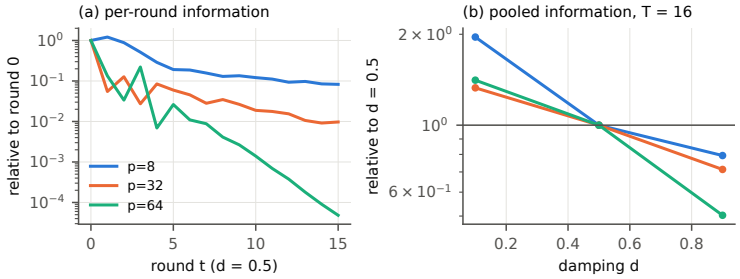


Figure 4: Standard observation model. Along damped best response in 108 environments, per-round information on the least-informed direction (a) falls by a factor of 12 at $p = 8$, 100 at $p = 32$ (just below $|S|^2 = 36$) and $2 \times 10^4$ at $p = 64$; (b) pooled information at $T = 16$ relative to $d = 0.5$. Exact trajectory Fisher information per transition ($H = 540$). Figure 1b shows the saturation of a single blind coordinate.

Table 2: Blind dimension on the dense benchmark ($|S| = 6$, $|A| = 3$; damped best response, $d = 0.5$, 400 rounds).

| | $p = 8$ | $p = 32$ (seed 1) | $p = 32$ (seed 2) | $p = 64$ (seed 1) | $p = 64$ (seed 3) |
|---|---|---|---|---|---|
| best response settles | yes | no (cycles) | no (cycles) | yes | yes |
| predicted $\dim \mathcal{N}$ (Cor. 2) | 0 | 2 | 2 | 34 | 34 |
| directions growing $< 1.5\times$ from $T = 100$ to 400 | 0 | 2 | 0 | 36 | 35 |

**Effective blind dimension.** (`effdim.py`) damped best response for 60 rounds, then its deterministic rounding; $\tau = 100$ ($\sigma = 0.1$), $H = 50$. Entries are lower bound / observed / upper bound from Proposition 1; $T_{\mathrm{res}} = t_0 + \tau/\mu_{\dim\mathcal{N}+1}$ is the horizon at which the upper bound reaches $\dim\mathcal{N}$.

| $\lvert S\rvert$ | $\lvert A\rvert$ | $p$ | seed | $\dim\mathcal{N}$ | $T = 300$ | $T = 3000$ | $T = 30000$ | $T_{\mathrm{res}}$ |
|---|---|---|---|---|---|---|---|---|
| 6 | 3 | 26 | 1 | 0 | 26/26/26 | 18/18/18 | 9/9/9 | $1.1 \times 10^{8}$ |
| 6 | 3 | 26 | 2 | 0 | 23/23/25 | 15/16/16 | 7/9/9 | $1.8 \times 10^{10}$ |
| 6 | 3 | 34 | 1 | 4 | 26/29/31 | 17/21/21 | 11/14/14 | $2.2 \times 10^{7}$ |
| 6 | 3 | 34 | 2 | 4 | 27/31/31 | 18/22/22 | 11/13/13 | $1.3 \times 10^{7}$ |
| 6 | 5 | 26 | 1 | 0 | 24/24/25 | 14/15/15 | 8/9/9 | $2.4 \times 10^{6}$ |
| 6 | 5 | 26 | 2 | 0 | 21/24/25 | 12/16/16 | 6/9/9 | $7.4 \times 10^{6}$ |
| 6 | 5 | 34 | 1 | 4 | 32/32/32 | 23/23/24 | 16/17/17 | $2.7 \times 10^{8}$ |
| 6 | 5 | 34 | 2 | 4 | 31/31/32 | 22/23/23 | 15/17/17 | $1.1 \times 10^{7}$ |
| 10 | 3 | 86 | 1 | 0 | 86/86/86 | 67/68/69 | 40/42/42 | $9.7 \times 10^{10}$ |
| 10 | 3 | 86 | 2 | 0 | 86/86/86 | 69/71/71 | 46/47/47 | $1.7 \times 10^{10}$ |
| 10 | 3 | 94 | 1 | 4 | 93/93/94 | 68/72/72 | 43/46/46 | $1.6 \times 10^{9}$ |
| 10 | 3 | 94 | 2 | 4 | 93/93/94 | 76/78/79 | 49/52/52 | $5.6 \times 10^{9}$ |
| 10 | 5 | 86 | 1 | 0 | 86/86/86 | 70/73/73 | 45/47/47 | $1.1 \times 10^{10}$ |
| 10 | 5 | 86 | 2 | 0 | 86/86/86 | 67/68/68 | 43/44/44 | $2.7 \times 10^{10}$ |
| 10 | 5 | 94 | 1 | 4 | 94/94/94 | 76/77/78 | 48/49/49 | $2.0 \times 10^{8}$ |
| 10 | 5 | 94 | 2 | 4 | 94/94/94 | 80/80/80 | 55/56/56 | $1.0 \times 10^{9}$ |
| 15 | 3 | 206 | 1 | 0 | 206/206/206 | 183/185/186 | 122/124/124 | $2.4 \times 10^{11}$ |
| 15 | 3 | 206 | 2 | 0 | 206/206/206 | 176/178/179 | 118/120/120 | $6.1 \times 10^{10}$ |
| 15 | 3 | 214 | 1 | 4 | 214/214/214 | 186/188/189 | 125/128/128 | $5.6 \times 10^{10}$ |
| 15 | 3 | 214 | 2 | 4 | 214/214/214 | 190/191/191 | 130/132/132 | $1.1 \times 10^{11}$ |
| 15 | 5 | 206 | 1 | 0 | 206/206/206 | 180/182/183 | 115/117/117 | $3.2 \times 10^{11}$ |
| 15 | 5 | 206 | 2 | 0 | 206/206/206 | 189/190/191 | 130/132/132 | $3.9 \times 10^{11}$ |
| 15 | 5 | 214 | 1 | 4 | 214/214/214 | 182/184/185 | 116/119/119 | $2.9 \times 10^{10}$ |
| 15 | 5 | 214 | 2 | 4 | 214/214/214 | 181/184/185 | 119/122/122 | $1.1 \times 10^{10}$ |

**Dimension law at larger scale.** (`dimlaw_scale.py`) damped best response, $d = 0.5$, 300 rounds, trajectory information with $H = 50$.

| $\|S\|$ | $\|A\|$ | $p$ | seed | settles | predicted $\dim\mathcal{N}$ | $\max(0, p - \|S\|(\|S\|-1))$ | growing $< 1.5\times$ (100→300) |
|---|---|---|---|---|---|---|---|
| 6 | 3 | 26 | 1 | yes | 0 | 0 | 1 |
| 6 | 3 | 26 | 2 | no | 0 | 0 | 0 |
| 6 | 3 | 34 | 1 | yes | 4 | 4 | 8 |
| 6 | 3 | 34 | 2 | yes | 4 | 4 | 4 |
| 6 | 5 | 26 | 1 | yes | 0 | 0 | 1 |
| 6 | 5 | 26 | 2 | yes | 0 | 0 | 1 |
| 6 | 5 | 34 | 1 | no | 4 | 4 | 0 |
| 6 | 5 | 34 | 2 | no | 4 | 4 | 0 |
| 10 | 3 | 86 | 1 | yes | 0 | 0 | 6 |
| 10 | 3 | 86 | 2 | no | 0 | 0 | 2 |
| 10 | 3 | 94 | 1 | yes | 4 | 4 | 8 |
| 10 | 3 | 94 | 2 | yes | 4 | 4 | 12 |
| 10 | 5 | 86 | 1 | yes | 0 | 0 | 10 |
| 10 | 5 | 86 | 2 | yes | 0 | 0 | 7 |
| 10 | 5 | 94 | 1 | no | 4 | 4 | 6 |
| 10 | 5 | 94 | 2 | yes | 4 | 4 | 14 |
| 15 | 3 | 206 | 1 | no | 0 | 0 | 0 |
| 15 | 3 | 206 | 2 | no | 0 | 0 | 8 |
| 15 | 3 | 214 | 1 | yes | 4 | 4 | 26 |
| 15 | 3 | 214 | 2 | yes | 4 | 4 | 31 |
| 15 | 5 | 206 | 1 | yes | 0 | 0 | 15 |
| 15 | 5 | 206 | 2 | yes | 0 | 0 | 19 |
| 15 | 5 | 214 | 1 | yes | 4 | 4 | 27 |
| 15 | 5 | 214 | 2 | yes | 4 | 4 | 25 |

**Base rate on unplanted environments.** Registered before the draws (`pbs_v11.py`, part R2) and run on 500 environments with nothing planted: $|S| = 6$, $|A| = 3$, $\kappa \in \{3, 10\}$, 100 draws at each $p \in \{4, 8, 16, 32, 40\}$, no qualification rule. For each we ran damped best response ($d = 0.5$) for 300 rounds, computed the blind subspace of the retained pairs at its limit under the standard observation model, and applied the decision-relevance rule of `gen_sep`: a blind direction counts if moving the parameter along it changes the optimal action somewhere.

| $p$ | 4 | 8 | 16 | 32 | 40 |
|---|---|---|---|---|---|
| $\dim\mathcal{N} > 0$ | 0/100 | 0/100 | 0/100 | 0/100 | 76/100 |
| decision-relevant | 0/100 | 0/100 | 0/100 | 0/100 | 66/100 |
| median $\dim\mathcal{N}$ | 0 | 0 | 0 | 0 | 4 |

The registration predicted a nontrivial blind subspace in essentially every environment above the threshold $|S|^2 = 36$. It was 76 of 100, not essentially all: damped best response does not always reach a deterministic limit within 300 rounds, and a limit that keeps mixing retains enough pairs to leave nothing blind. No prediction was registered for the decision-relevant share, which is the quantity this part exists to measure; it was 66 of 100, or 87% of the environments whose blind subspace was nontrivial. Below the threshold nothing was blind in 400 draws, which is Corollary 2 with $p < |S|^2$ and no planted structure: at small $p$ the failure this paper studies is not generic, and needs a response that few pairs reveal.

**The regret comparator.** For each of the 48 environments it enumerates all $3^6 = 729$ deterministic plans and computes the performative value of each, giving the exact optimum over deterministic plans, the class against which Theorems 6 and 7 measure regret. The reference used in the runs equals it in 42 environments and exceeds it in 6, where the search found a stochastic policy of higher performative value; it is below it in none, so no reported regret understates regret against the deterministic optimum. A separate, independent search (400 Dirichlet restarts and 200 random deterministic starts, each polished by 120 ascent steps, a different seed and about $20\times$ the original budget) improves on the reference in 4 of 48 environments, by at most $3.4 \times 10^{-3}$ in value against a median $|J|$ of 6.7; over 400 deployments that is at most 1.4 of cumulative regret, against differences between methods of 30 to 450 in Table 1. Method rankings are unchanged.

**Comparing observation models.** With four exploratory deployments chosen once per environment, we change only the observation model (36 environments, pre-registered; Figure 6). Iden-

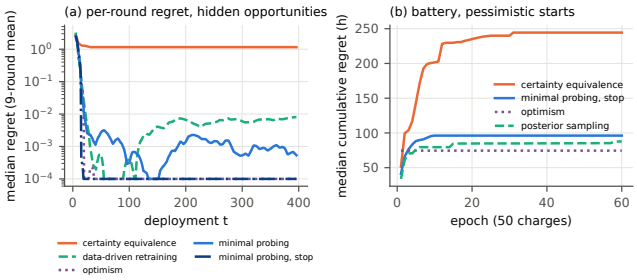


Figure 5: (a) Median per-round regret on the hidden-opportunity class; curves at the bottom edge are zero. (b) Battery charging from pessimistic beliefs: median cumulative regret in hours, over 60 epochs of 50 charges.

tification beats stabilization in 36/36 environments under queries with $\sigma = 0.05$, in 34/36 under generative sampling with 300 draws per pair, and in 36/36 under on-policy trajectories of 54,000 steps. The pre-registered hypothesis that the richest settings would be ordered query $>$ generative $>$ trajectory was not supported (McNemar $p = 1$ and $p = 0.5$). Every observation model identifies the response when its deployments explore; what differs is how many transitions it costs. We report this because an earlier version of this work concluded the opposite from four environments. That conclusion was an artifact of a smoothed-count least-squares estimator whose relative error was 1.9–2.9 where the MLE attains 0.10–0.72, near the Cramér–Rao bound.

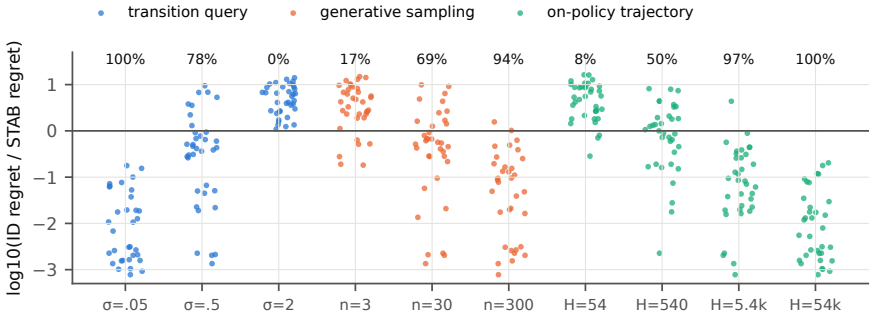


Figure 6: Same 36 environments and deployments; only the observation model changes. Each point is one environment, and points below zero are environments where identification beats stabilization. Numbers above each column: share of environments where identification wins.

**Design under queries (Q).** Greedy A-optimal deployment reduces identification regret against random deployment by 43.3% ($T = 2$, noise at the signal scale, $p = 0.008$), 34.8% ($T = 4$, $p = 0.031$) and 32.9% ($T = 4$, low noise, $p = 0.094$), not at $T = 2$ with low noise ($-2.8\%$, $p = 0.86$), and not at all when data are plentiful.

## E Analysis pipeline

Every choice behind the numbers in §5, in one place.

| | |
|---|---|
| Observational unit | the environment: 100 per class, 200 in total. Its 5 replications differ only in the random seed of the trajectories and of each arm's internal randomness, which is shared across arms. |
| Within-environment aggregation | arithmetic mean over the environment's 5 replications, applied to every reported quantity (cumulative regret, the recovery indicator, per-round regret over the last 50 rounds, wall time), before any test or interval. |
| Table 1 medians | median over the 100 environment-level means of that class, not over the 500 runs. |
| Recovery | per environment, the fraction of its 5 runs whose final plan takes every decision-relevant action of the plan at the true parameter; the table reports the mean of that fraction over the class's 100 environments, which equals the run-level share because every environment contributes exactly 5 runs. On the separation class the hurts/helps columns use the 50 environments (250 runs) of that orientation as the denominator. |
| Paired statistic | the 100 environment-level differences; the Hodges–Lehmann estimate is the median of their pairwise Walsh averages, which is not the difference of the reported medians. |
| Intervals | 95% percentile bootstrap, resampling *environments* (not runs), 20,000 resamples, seed 0. |
| Primary tests | P1 and P2: one-sided paired Wilcoxon signed-rank on the environment-level differences. P3: a non-inferiority test, met when the upper end of the two-sided 95% interval for the paired difference PBS-STOP $-$ OPT is below $+10$ regret units, a margin fixed before the run. |
| Multiplicity | Holm over the three primaries, and only those. The secondary comparisons of Table 4 are uncorrected and reported as effect sizes with intervals rather than as tests. |
| Ties and missing runs | no runs are missing: each of the 11 arms has exactly 1000 completed runs. Exact zero differences are discarded by the signed-rank test; had all differences been zero the $p$-value would have been set to 1. |

# F ASSUMPTIONS, AND WHERE THEY BIND

The results rest on assumptions of three kinds: what the learner knows, what the environment does, and what the constants are worth. The third is the one that matters most in practice, and it is the least visible in a theorem statement, so we give measured values from our own benchmark.

| Assumption | What it buys | Where it binds |
|---|---|---|
| Known response design $B, W, \phi$ | the blind subspace is computable before any data, and its dimension obeys Corollary 2 | identifiability is always relative to a model class (Proposition 7), so this is not removable, only replaceable by a different class. Our battery study has no such design, which is why it runs a numerical analogue rather than the theorem. |
| (B1) margin $r$ | Theorem 6 and certified stopping | $r$ is unknown in practice, and $\pi^\star$ is not locally constant at a decision boundary. Certified stopping needs $r$ and $T$ in advance. |
| (B2) $\Theta$, $c_\Theta$ | the confidence bound of Theorem 5 | $c_\Theta^{-1} \approx 10^{10}$ on our benchmark, so the radius is about ten orders of magnitude larger than the realized error, and the certified stopping rule never fires within our horizons. |
| (B3) $\lvert g_j(\pi)\rvert \geq g_{\mathrm{lo}}$ | Lemma 8, hence the $\sqrt{T}$ rate of Theorem 6 | it holds in our environments but vacuously: enumerating every deterministic plan in 16 of them gives $g_{\mathrm{lo}}$ with median $1.8 \times 10^{-5}$ and worst case $1.8 \times 10^{-6}$. Since $c \propto g_{\mathrm{lo}}^2$ and $t^* \propto 1/c$, the burn-in of Theorem 6 is far beyond any horizon we run. The rate is asymptotic reassurance, not a usable guarantee at this scale. It is also in tension with gated responses, where a trigger genuinely vanishes. |
| (B4) every plan has a cover | probing can restore identifiability at all | no cover exists exactly when a direction is invisible at every pair; then no probing scheme helps, and the test says so. |
| Exact planning and exact posterior | Theorems 6 and 7 | our experiments use sampled approximations. With an $\epsilon_{\mathrm{p}}$-optimal planner every round loses at most $\epsilon_{\mathrm{p}}$ more, so the bounds hold with an added $\epsilon_{\mathrm{p}} T$; that term is linear, so approximate planning is benign only when $\epsilon_{\mathrm{p}} = O(1/\sqrt{T})$. |
| Full-support $\rho$ | the first-step lower bound $\bar{w}_\pi(s,a) \geq \rho(s)\pi(a \mid s)/H$ used in Propositions 4 and 3 | with a fixed start state it fails as stated and must be replaced by a reachability argument under the deployed plan. |
| Converging deployments | Theorem 2(b) and everything downstream | best response can cycle, and does in some of our own environments (Table 2), where the prediction need not hold. |
| Weak response, $\lVert\epsilon\rVert_\infty \leq \bar{\epsilon}$ | the induced-occupancy transfer of Appendix M | the standard formulation is covered only for mild performativity; for a strong response the trigger fixed point need not be unique and we claim nothing. |

Two of these deserve emphasis because they are quantitative rather than structural. The constants in (B2) and (B3) do not invalidate Theorems 5 and 6, but they make them vacuous at the scale we can run: the guarantee is that probing eventually stops paying for exploration, and "eventually" is past any horizon in this paper. What the experiments show is that the *algorithm* works long before its bound does, which is common for confidence-set methods and is why we report the heuristic rule separately.

## G Minimal probing and other exploration approaches

On our benchmark the certified threshold is never reached within the horizon, since the confidence radius is about $10^{10}$ times the empirical estimation error in the corresponding norm, so the certified rule would probe every round. The version we run (Algorithm 1) therefore replaces $\bar{\beta}_T / r^2$ with the smaller heuristic threshold $\eta_0 \log(1+t)$ for a constant $\eta_0$ (Lai and Robbins, 1985; Combes et al., 2017); the guarantee does not transfer to it.

What each method in Table 1 does, how it was configured, and what minimal probing adds and gives up.

**Algorithm 1** Minimal probing with a heuristic stopping rule (PBS-STOP); Corollary 3 gives the certified version. $\mathrm{mix}(\hat{\pi}, \mathcal{Q}, \rho)$ plays $\hat{\pi}$, except that at each state of $\mathcal{Q}$ it puts total probability $\rho$ uniformly on the cover's actions; $k$ is the re-planning period.

$\hat{\epsilon} \leftarrow 0$; $\hat{\pi} \leftarrow \pi^\star(\hat{\epsilon})$; $\mathcal{Q} \leftarrow \emptyset$
**for** $t = 1, 2, \ldots$ **do**
  deploy $\pi_t = \mathrm{mix}(\hat{\pi}, \mathcal{Q}, \alpha_t)$ if probing is active and $\mathcal{Q} \neq \emptyset$, else $\hat{\pi}$;  $\alpha_t = \min(\frac{1}{2}, t^{-1/2})$
  observe a trajectory; every $k$ rounds: $\hat{\epsilon} \leftarrow$ MLE on all data; $\hat{\pi} \leftarrow \pi^\star(\hat{\epsilon})$
  $U \leftarrow$ orthonormal basis of $\mathcal{N}(\hat{\pi})$; $\mathcal{Q} \leftarrow$ greedy cover of $\mathcal{N}(\hat{\pi})$ (Prop. 4a)
  probing active $\iff \lambda_{\min}\big(U^\top \sum_{u \le t} I_O(\pi_u; \hat{\epsilon})\, U\big) < \eta_0 \log(1+t)$
**end for**

Method names, here and in Appendix N: CE (certainty equivalence), RRM (data-driven retraining), UNIF ($\epsilon$-greedy), BOLTZ (Boltzmann dithering), COUNT (count bonus), TS (Thompson sampling), OPT (optimism), IDS (information-directed sampling), DOPT (D-optimal probing), PBS (minimal probing), PBS-STOP (with its stopping rule) and PBS-DR (probing only decision-relevant directions).

Proposition 3 sorts exploration by where it deploys. In the standard model a plan's own deployments reveal its value, so methods whose deployments are always optimal for some plausible parameter (posterior sampling, optimism) are not structurally blind; certainty equivalence is, because it stops exploring. When the reward response is not observed (Appendix N), only methods that deploy outside $\Pi_\Theta$ can escape a separated blind spot. Minimal probing is *structure-directed forced exploration*: it deploys the excluded pairs that fill the current plan's blind subspace, chosen by rank gained per unit cost.

Table 3: Exploration approaches. "Standard": rewards observed; "partial": reward response unobserved.

| Approach | What it explores | Blind under the standard model? | Blind under partial observation? | Relation to minimal probing |
|---|---|---|---|---|
| $\epsilon$-greedy (UNIF), Boltzmann (BOLTZ) | every excluded action, uniformly or by value | no, if exposure diverges | no, if exposure diverges | non-minimal covers; tested |
| Count bonus (COUNT) | rarely visited pairs | only if the bonus is below the gap | only if the bonus is below the gap | structure-blind; tested |
| Thompson sampling (TS) | plans optimal under a posterior sample | not structurally (Prop. 3) | yes, if separated | value-directed; tested with neutral, overconfident and inflated posteriors |
| Optimism (OPT) | plans optimal under an optimistic parameter | not structurally | yes, if separated; fails when the unseen response is harmful | value-directed; tested with a Laplace and a Thm. 5-shaped ellipsoid |
| Information-directed sampling (Russo and Roy, 2018; Kirschner and Krause, 2018) (IDS) | actions trading regret against information about the optimum | no | no: mixtures leave $\Pi_\Theta$ | value of information; tested (performative variant) |
| D-optimal design (Pukelsheim, 2006) (DOPT) | deployments maximizing information about $\epsilon$ | no | no | variance-optimal probes instead of a rank cover; tested |
| Structured minimal exploration (Combes et al., 2017) | rates from a program over confusing parameters | no | no | closest relative; its program is non-convex for performative planning; not run |
| Dual control, Bayes-adaptive RL | the Bayes-optimal trade-off | no | no | intractable here; not run |
| Safe exploration (Berkenkamp et al., 2017) | certified-safe actions only | if $E_v$ is unsafe | if $E_v$ is unsafe | orthogonal objective; not run |

**Specifications.** All arms re-estimate by maximum likelihood and re-plan every five rounds. TS: Laplace posterior, prior $\mathcal{N}(0, I)$, one sample per re-plan. OPT: eight points on the 90% posterior ellipsoid, deploy the plan with the largest value under its own parameter. BOLTZ: softmax of advantage gaps at temperature $0.2/\sqrt{t}$. COUNT: optimistic base reward $0.5\sqrt{\log(t+1)/(n_{sa}+1)}$. IDS: candidates are the MAP plan, the MAP plan mixed with each excluded pair, and plans for four posterior samples; deploy the minimizer of squared expected regret over information gain, with the exact trajectory information of Theorem 1. DOPT: probe the single excluded pair maximizing log-determinant gain per unit advantage gap, at PBS's rate. Inflated TS and the Theorem 5-shaped optimism (Appendix N) use radius $\sqrt{2\log 10 + \log\det(I + H_t)} + \sqrt{p}$ in the observed-information norm, constrained to $\Theta$ in that norm.

**Effect sizes.** Table 4 restates the comparisons of Table 1 as paired differences in cumulative regret, with the environment as the unit: the Hodges–Lehmann estimate and a 95% percentile bootstrap interval over the 100 environments of each class (20,000 resamples; `analyze_pbs_v11.py`). The three primary tests are registered; the remaining rows are the registered secondary comparisons, reported as effect sizes rather than $p$-values.

Table 4: Paired differences in cumulative regret, minimal probing with heuristic stopping minus each other method (negative favours probing), with 95% bootstrap intervals over the 100 environments of each class.

| | hidden opportunities | | separation | |
|---|---|---|---|---|
| comparison | difference | 95% CI | difference | 95% CI |
| vs certainty equivalence | $-437$ | $[-535, -343]$ | $-216$ | $[-313, -139]$ |
| vs data-driven retraining | $-40$ | $[-103, -18]$ | $-13$ | $[-26, -7]$ |
| vs $\epsilon$-greedy | $-523$ | $[-621, -457]$ | $-451$ | $[-498, -417]$ |
| vs Boltzmann dithering | $-70$ | $[-106, -30]$ | $-36$ | $[-52, -22]$ |
| vs count bonus | $-95$ | $[-245, -61]$ | $-24$ | $[-32, -19]$ |
| vs D-optimal probing | $-41$ | $[-48, -34]$ | $-39$ | $[-47, -34]$ |
| vs minimal probing, no stopping | $-3.0$ | $[-7.6, -0.9]$ | $-41.6$ | $[-60, -28]$ |
| vs Thompson sampling | $-5.9$ | $[-8.6, -3.4]$ | $+0.1$ | $[-1.6, +1.9]$ |
| vs optimism | $+4.3$ | $[+1.7, +7.3]$ | $+4.1$ | $[+2.4, +6.1]$ |
| vs information-directed sampling | $+1.8$ | $[-0.8, +4.9]$ | $+4.3$ | $[+2.7, +6.5]$ |

**What is known, and what is not ours.** Thompson sampling is sensitive to its prior (Liu and Li, 2016; Simchowitz et al., 2021), and frequentist guarantees for linear Thompson sampling need posterior inflation (Abeille and Lazaric, 2017), which is necessary in the worst case (Hamidi and Bayati, 2020); our overconfident-prior experiment (v7) replicates this and is not evidence for probing. That posterior sampling and optimism skip purely informative actions is the observation behind IDS (Russo and Roy, 2018). What is specific to performative MDPs is when that observation applies: never under the standard observation model (Proposition 3), and under an exact, checkable separation condition when the reward response is unobserved (Proposition 10).

**What minimal probing offers, and what it gives up.** It makes the unidentifiable object explicit and computable before any data are collected, comes with a rank guarantee, a regret bound and a certified stopping rule, and chooses probes by cost with Wolsey's guarantee; it is also cheap, a rank computation per re-plan. It gives up decision relevance (it probes blind directions that cannot change the plan; PBS-DR addresses this heuristically), rate ($\sqrt{T}$ forced exploration against the $\log T$ attainable by optimal structured methods on well-separated problems), and information per probe, which optimal design targets directly. In our experiments it did not have lower regret than optimism or IDS where those escape.

## H THE COVERAGE ASSUMPTIONS, READ THROUGH THE TEST

The finite-sample guarantees in performative RL rest on conditions that read like regularity: an overlap ratio between the deployed occupancy and the optimum (Mandal et al., 2023, Asm. 2), a bounded coverage ratio (Mandal and Radanovic, 2025, Asm. 4), positive occupancy of every

pair enforced by mixing in a random policy (Pollatos et al., 2025), or an occupancy lower bound (Chen and Huang, 2025). Our results say what they are doing. They are not discharged by the algorithms those papers analyse: along damped best response the visited set contracts to one pair per state, so coverage fails on the path to the solution concept itself, and the test of §3 decides whether identifiability fails with it. Nor are they a convenience. In the unrestricted class that defines a performative MDP, Proposition 7 makes coverage the *entire* source of identification, so a coverage assumption is an assumption that the learner keeps exploring, which the theory of stable points does not supply. Structure makes coverage sufficient rather than necessary, which is why the grid-world stays identifiable after losing 61 of 256 pairs.

## I REWARD-BIASED MAXIMUM LIKELIHOOD

RBMLE (Kumar and Becker, 1982; Kumar, 1985; Becker et al., 1985) is the classical answer to the identifiability problem: instead of exploring, bias the estimate toward parameters under which the achievable value is higher, $\hat{\epsilon}_t = \arg\max_\epsilon [\ell_t(\epsilon) + \alpha_t J_\epsilon(\pi^\star(\epsilon))]$, and deploy $\pi^\star(\hat{\epsilon}_t)$. We added it after the registered experiments were analysed, so it is exploratory; the schedule $\alpha_t = \sqrt{t}$ was fixed before the first run and no other was tried, the optimiser is a local search seeded at the plain MLE, and everything else matches the registered arms. It ran on the 24-environment version of each class, so it is paired with those runs rather than with Table 1.

| | hidden opportunities | | separation | |
|---|---|---|---|---|
| | median regret | recovery | median regret | recovery |
| RBMLE | 55 | 75% | 37 | 35% |
| certainty equivalence | 474 | 17% | 67 | — |
| minimal probing, heuristic stopping | 35 | 100% | 26 | 100% |
| optimism | 28 | 100% | 20 | 100% |

Paired Hodges–Lehmann differences, RBMLE minus the other arm, with 95% bootstrap intervals over the 24 environments: against CE, $-362$ $[-511, -221]$ on hidden opportunities and $-91$ $[-253, -0.7]$ on separation; against optimism, $+30$ $[+0.2, +127]$ and $+7$ $[-0.1, +44]$; against minimal probing with stopping, $+16$ $[-0.5, +128]$ and $+7$ $[-0.6, +44]$; against IDS, $+19$ $[-0.4, +119]$ and $+10$ $[+0.9, +46]$. It costs about 730 seconds per run, five times optimism and twenty-five times CE, because each estimate needs planning solves. The bias does repair much of the blind spot, which is what Proposition 3 predicts: raising the value of parameters under which some unvisited pair looks good makes the plan optimal for such a parameter, and such plans expose the direction. But the repair is partial in a way exploration is not: in a quarter of hidden-opportunity environments and two thirds of separation environments RBMLE settles without ever taking the decision-relevant action. Biasing an estimator moves where the learner believes the optimum is; it does not guarantee the learner ever looks.

## J SCALE SWEEP

Registered as part R3 of `pbs_v11.py` and amended for compute before any of it ran (the amendment, its reason and the original cells are in the script header). Hidden-opportunity environments at several sizes, 20 per cell and 2 replications, arms CE, minimal probing with its heuristic stopping rule, and optimism; regret against the certified deterministic optimum where $|A|^{|S|} \leq 10^6$.

| $\|S\|$ | $\|A\|$ | $p$ | $p > \|S\|^2$? | CE | probing | optimism | recovery (CE / probing / optimism) | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 3 | 10 | no | 354 | 37 | 35 | 22% | 98% | 95% |
| 6 | 3 | 40 | yes | 284 | 155 | 67 | 65% | 100% | 95% |
| 6 | 5 | 10 | no | 747 | 46 | 45 | 5% | 100% | 100% |
| 6 | 5 | 40 | yes | 438 | 173 | 80 | 48% | 100% | 98% |
| 10 | 3 | 10 | no | 361 | 28 | 26 | 22% | 100% | 98% |
| 10 | 5 | 10 | no | 339 | 32 | 31 | 3% | 100% | 100% |

The pattern of §5 holds at every size: CE recovers the decision-relevant action in 3–65% of runs and the exploring arms in 95–100%, with CE's regret an order of magnitude higher. The registered

prediction was that CE would trail only in the cells above the dimension threshold; that is wrong, and for the reason the base rate already showed. These environments are built to be blind through a planted coordinate, so they are blind at any $p$, and the threshold governs only unplanted draws. The cell at $|S| = 10$, $p = 104$ was still running at submission and is not reported.

## K Extended comparison with classical results

The nullspace picture of §3 has close relatives outside performative RL, and stating the differences precisely is the honest way to place the contribution.

Borkar and Varaiya (1979; 1982) show that certainty-equivalent adaptive control of a finite Markov chain converges to a parameter that is indistinguishable from the truth under the limiting policy, which is our Proposition 2 in a non-performative model: there the environment is fixed and the learner's policy merely fails to excite it, whereas here the policy also *changes* the environment, so the object that goes unidentified is the response itself and the pairs that stop being visited are chosen by the response the learner is trying to learn. Ziemann and Sandberg (2020) characterize uninformative optimal policies in adaptive LQR through a nullspace and prove a $\sqrt{T}$ lower bound by van Trees; their nullspace is of the closed-loop covariance in a linear-quadratic model with a continuum of states, ours of a finite set of response blocks with a computable test attached, and their lower bound is for an informative-regret trade-off rather than for the two-instance blindness of Theorem 4. Bittanti et al. (1990); Lai and Robbins (1981) give the least-squares statement: estimates converge only on the excited subspace. The adaptive-control literature also settled what to do about it: Becker et al. (1985) characterize the limit set of stochastic-approximation estimates, which lies on the manifold of closed-loop-indistinguishable parameters, and Lin et al. (1985) show that self-tuning can fail once the cost criterion is general; Kumar and Becker (1982) bias the likelihood toward parameters with higher optimal reward, which buys average-cost optimality without explicit probing. Minimal probing is the opposite trade: it leaves the estimator alone and pays a measured exploration cost, which is what makes its guarantee hold with the planner retraining already uses. Combes et al. (2017) construct minimal exploration sets for structured bandits by a program over confusing parameters; our cover is the analogous object for a performative MDP, but it is chosen by rank rather than by a confusing-parameter program, which is why it is cheap to compute and why its guarantee is weaker. In pricing, incomplete learning under certainty equivalence and the minimal experimentation that repairs it are classical (Broder and Rusmevichientong, 2012; den Boer and Zwart, 2014; Keskin and Zeevi, 2018), and in LQR certainty equivalence with decaying exploration noise is optimal (Simchowitz and Foster, 2020); neither setting has a response whose support the learner controls.

## L Beyond the exponential tilt

A performative MDP in the sense of Mandal et al. (2023) is any map $\pi \mapsto (P_\pi, r_\pi)$ that is Lipschitz in $\pi$; §2 studies a parametric family inside that class. This appendix separates what the parametric form is doing from what it is not. We establish two statements. The information geometry of §3 is not special to the tilt and holds for any smooth family (L.1); and without a family there is nothing to identify at all (L.2), which is why a test of this kind has to be relative to a model class.

### L.1 Any smooth family

Let $\Theta \subseteq \mathbb{R}^p$ and let $(\pi, \epsilon) \mapsto (P^\epsilon_\pi, r^\epsilon_\pi)$ be a response family with $P^\epsilon_\pi(\cdot \mid s, a)$ bounded below by $c_0 > 0$ and $C^1$ in $\epsilon$. Define the *block at a pair* to be the Jacobian of what a visit to that pair reveals,

$$A_{sa}(\pi, \epsilon) = \Big[\partial_\epsilon \log P^\epsilon_\pi(\cdot \mid s, a);\ \sigma_r^{-1} \partial_\epsilon r^\epsilon_\pi(s, a)\Big] \in \mathbb{R}^{(|S|+1)\times p},$$

the first $|S|$ rows being the score of the successor and the last the score of the reward.

**Proposition 5** (Factorization, general form)**.** *Under (O), the Fisher information about $\epsilon$ in one deployment of $\pi$ is*

$$I_O(\pi; \epsilon) = H \sum_{s,a} \bar{w}_\pi(s, a)\, A_{sa}(\pi, \epsilon)^\top \Lambda_{sa}(\pi, \epsilon) A_{sa}(\pi, \epsilon), \qquad \Lambda_{sa} = \begin{bmatrix} F(P^\epsilon_\pi(\cdot \mid s, a)) & 0 \\ 0 & 1 \end{bmatrix},$$

*and under (G) the same sum with $H\bar{w}_\pi$ replaced by $n$.*

*Proof.* Identical to the proof of Theorem 1 with $B_{sa}D_g$ replaced by $\partial_\epsilon \log P^\epsilon_\pi(\cdot \mid s,a)$ and $-\kappa D_g w_{sa}$ by $\partial_\epsilon r^\epsilon_\pi(s,a)$: the transition score at a visited pair has mean zero and covariance $(\partial_\epsilon \log P)^\top F(P)(\partial_\epsilon \log P)$, the reward score contributes $\sigma_r^{-2}(\partial_\epsilon r)(\partial_\epsilon r)^\top$, the two are conditionally uncorrelated, and the martingale argument over the trajectory is unchanged, since it uses only that each step's score has zero conditional mean. □

**Proposition 6** (Dichotomy, general form)**.** *Fix $\epsilon$, let $\pi_t \to \pi_\infty$, let $\mathcal{R}$ be the pairs with infinite cumulative expected visitation, and suppose $\sum_t \|A_{sa}(\pi_t,\epsilon) - A_{sa}(\pi_\infty,\epsilon)\|^2_{\mathrm{op}} < \infty$ for every $(s,a) \in \mathcal{R}$ and $\sup_{t,s,a} \|A_{sa}(\pi_t,\epsilon)\|_{\mathrm{op}} < \infty$. Then Theorem 2(b) holds with*

$$\mathcal{N}(\pi_\infty,\epsilon) = \{v : A_{sa}(\pi_\infty,\epsilon)v = 0 \text{ for all } (s,a) \in \mathcal{R}\},$$

*and Proposition 1 holds with $I_O(\pi_\infty;\epsilon)$ from Proposition 5.*

*Proof.* The proof of Theorem 2(b) uses the trigger only through the two steps isolated in the proof of Proposition 9, and both go through verbatim with $A_{sa}$ in place of $O_{sa}D_{g(\pi_t)}$. □

Our model is the case $A_{sa}(\pi,\epsilon) = O_{sa}D_{g(\pi)}$, and the induced-occupancy model of Appendix M the case $A_{sa}(\pi,\epsilon) = O_{sa}J(\pi,\epsilon)$. What the tilt buys is not the geometry but its *computability*: there the block splits into a known $\epsilon$-free matrix and a trigger, so the blind subspace can be computed from a candidate policy before any data exist, its dimension obeys the law of Corollary 2, and the score is linear in $\epsilon$, which is what Theorem 5 uses. For a general smooth family, $A_{sa}$ depends on $\epsilon$, so the test becomes local: it is run at an estimate and certifies a neighbourhood, the dimension count is a rank condition rather than $p - |S|^2$, and Theorems 5, 6 and 7, which rest on the linear score and on covers of $\epsilon$-free blocks, do not follow from Proposition 6. We state them only for the tilt.

`general_family_check.py` verifies Propositions 5 and 6 on a family that is not a tilt: a mixture kernel $(1-h)P_0 + hQ$ with $h$ a squashed logistic of the trigger, and a saturating reward response. The information from the blocks matches a direct finite-difference computation of the trajectory Fisher information to $10^{-10}$, and the kernel of the information equals the common kernel of the blocks at visited pairs (0 at $p = 6$ and 20 at $p = 24$ with a deterministic limit). The same script confirms that the blocks move with $\epsilon$ in this family, unlike the tilt.

### L.2 Without a family, nothing is identified

The literature's performative MDP assumes only that $\pi \mapsto (P_\pi, r_\pi)$ is Lipschitz (Mandal et al., 2023; Rank et al., 2024; Mandal and Radanovic, 2025; Pollatos et al., 2025). In that class the question this paper asks has a short and discouraging answer.

**Proposition 7** (Model-free blindness)**.** *Let $\mathcal{M}$ be the class of response maps that are $L$-Lipschitz in $\pi$ and have transition probabilities at least $c_0$. Let an algorithm deploy $\pi_1, \pi_2, \dots$ and suppose $(s^*,a^*)$ is a pair with $\sum_t \bar{w}_{\pi_t}(s^*,a^*) < \infty$ almost surely under $M \in \mathcal{M}$. Then for every $\eta > 0$ there is $M' \in \mathcal{M}$ agreeing with $M$ at every pair other than $(s^*,a^*)$, with $\sup_\pi \|(P^M_\pi - P^{M'}_\pi, r^M_\pi - r^{M'}_\pi)\| \geq \eta$ at $(s^*,a^*)$, and the laws of the algorithm's data under $M$ and $M'$ have total variation at most $C\eta \sum_t \bar{w}_{\pi_t}(s^*,a^*)$ for a constant $C = C(c_0,\sigma_r,H)$. Consequently no algorithm can distinguish them, and for $\eta$ large enough that the two optimal policies differ, one of the two instances forces linear regret.*

*Proof.* Build $M'$ from $M$ as follows. Fix a $C^\infty$ bump $\chi : \Pi \to [0,1]$ with $\chi(\pi_\infty) = 1$, supported in the ball of radius $\varrho$ around $\pi_\infty$, with $\|\nabla\chi\|_\infty \leq 2/\varrho$. Put $r^{M'}_\pi(s,a) = r^M_\pi(s,a) + \eta\chi(\pi)\mathbf{1}\{(s,a) = (s^*,a^*)\}$, and tilt the kernel at that pair only, $P^{M'}_\pi(\cdot \mid s^*,a^*) \propto P^M_\pi(\cdot \mid s^*,a^*)\exp(\eta\chi(\pi)b)$ for a fixed $b \in \mathbb{R}^{|S|}$ with $\mathbf{1}^\top b = 0$, $\|b\|_\infty = 1$, leaving every other pair untouched. Then $M'$ is $(L + C\eta/\varrho)$-Lipschitz with $C$ depending only on $c_0$, so for fixed $\varrho$ and $\eta \leq \varrho(L'-L)/C$ it lies in the class with constant $L' > L$, and its transition probabilities stay above $c_0 e^{-\eta}$. Only visits to $(s^*,a^*)$ can distinguish the two. The two laws differ only through steps taken at $(s^*,a^*)$, whose expected count is $H\sum_t \bar{w}_{\pi_t}(s^*,a^*)$; the per-visit total variation is at most $C'\eta$ by Pinsker and Lemma 7, and the chain rule for the total variation of the trajectory laws gives the bound. The regret statement follows as in Theorem 4, by the same two-instance argument. □

The content of Proposition 7 is that in the unrestricted class, *every* pair a learner stops visiting is a blind spot whose size the data do not bound at all, beyond the class-wide Lipschitz constant. Nothing propagates from the pairs that are still visited to the pairs that are not. A test that distinguishes safe from unsafe cases therefore cannot exist for the class as a whole: it must be relative to a model that ties the response at unvisited pairs to the response elsewhere, which is exactly what the parametric family does through the shared coordinates $\epsilon$. Seen this way our setting is the favourable case rather than a convenient one: structure can rescue identifiability, and the paper measures when it does and when it does not. The grid-world of Mandal et al. (2023) is an instance where it does; the classes of §5 are instances where it does not.

## M INDUCED-OCCUPANCY TRIGGERS

§2 drives the response with $g_j(\pi) = \langle \phi_j, d^0_\pi \rangle$, the occupancy under the base kernel. Mandal et al. (2023) and its successors drive it with the occupancy induced by the response itself. This appendix shows that everything in §3 carries over to that formulation, with one matrix replaced.

Write the induced model in the trigger variables. For a fixed deployed $\pi$, the environment of (1) depends on $\epsilon$ only through $\theta \in \mathbb{R}^p$, $\theta_j = \epsilon_j g_j$:

$$P_\theta(\cdot \mid s,a) = \mathrm{softmax}\Big( \log P_0(\cdot \mid s,a) + \textstyle\sum_j \theta_j B_j(s,a,\cdot) \Big), \qquad r_\theta = r_0 - \kappa \textstyle\sum_j \theta_j W_j.$$

In our model $\theta_j = \epsilon_j g_j(\pi)$ with $g$ known. In the induced-occupancy model $\theta$ solves the fixed point

$$\theta_j = \epsilon_j \langle \phi_j, d^\theta_\pi \rangle, \qquad j = 1, \dots, p, \tag{2}$$

where $d^\theta_\pi$ is the discounted occupancy of $\pi$ under $P_\theta$.

**Lemma 9** (The trigger is well posed and smooth)**.** *There is $\bar\epsilon > 0$ such that for $\|\epsilon\|_\infty \le \bar\epsilon$ and every $\pi$,* (2) *has a unique solution $\theta(\pi,\epsilon)$ in $\{\|\theta\|_\infty \le 2\bar\epsilon\bar\phi/(1-\gamma)\}$, and $\theta$ is $C^\infty$ in $\epsilon$ with $\partial\theta/\partial\epsilon \to D_g$ as $\epsilon \to 0$.*

*Proof.* $\theta \mapsto d^\theta_\pi$ is smooth with a derivative bounded by a constant $C_d$ depending only on $\gamma$ and $|S|$, since $d^\theta_\pi = (1-\gamma)\rho^\top (I - \gamma P^\pi_\theta)^{-1}$ and the resolvent is smooth with $\|(I-\gamma P^\pi_\theta)^{-1}\| \le 1/(1-\gamma)$, while $\theta \mapsto P_\theta$ is smooth with bounded derivative. The right-hand side of (2) therefore has Jacobian in $\theta$ of norm at most $\|\epsilon\|_\infty \bar\phi C_d$, which is below $\frac{1}{2}$ once $\bar\epsilon \le 1/(2\bar\phi C_d)$; it also maps the stated ball into itself, because $|\langle \phi_j, d^\theta_\pi \rangle| \le \bar\phi/(1-\gamma)$. Banach's fixed point theorem gives existence, uniqueness and, with the implicit function theorem, smoothness. At $\epsilon = 0$ the fixed point is $\theta = 0$, and differentiating (2) there gives $\partial\theta_j/\partial\epsilon_k = \delta_{jk} g_j(\pi)$. □

**Proposition 8** (Fixed-policy transfer)**.** *Fix $\pi$ and $\|\epsilon\|_\infty \le \bar\epsilon$, and let $J(\pi,\epsilon) = \partial\theta(\pi,\epsilon)/\partial\epsilon$. Under any of the observation models (O), (G), (Q), the Fisher information of a deployment of $\pi$ in the induced-occupancy model is*

$$I^{\mathrm{ind}}_O(\pi;\epsilon) = J(\pi,\epsilon)^\top \Big[ \textstyle\sum_{s,a} \nu_O \bar w_\pi(s,a)\, O_{sa}^\top \Lambda_O O_{sa} \Big] J(\pi,\epsilon) = J(\pi,\epsilon)^\top I_\theta(\pi;\theta) J(\pi,\epsilon),$$

*that is, Theorem 1 with $D_{g(\pi)}$ replaced by $J(\pi,\epsilon)$. Hence* $\ker I^{\mathrm{ind}}_O(\pi;\epsilon) = \{v : J(\pi,\epsilon)v \in \mathcal{N}_\theta(\pi)\}$, *which is $J^{-1}\mathcal{N}_\theta(\pi)$ when $J$ is invertible, where $\mathcal{N}_\theta(\pi) = \{u : O_{sa}u = 0$ for all $(s,a) \in \mathcal{R}(\pi)\}$, and then $\dim\ker I^{\mathrm{ind}}_O = \dim\mathcal{N}_\theta$, the dimension of Corollary 2.*

*Proof.* The likelihood of a deployment of $\pi$ depends on $\epsilon$ only through $\theta(\pi,\epsilon)$: the successor at a visited pair is drawn from $P_\theta(\cdot \mid s,a)$ and its reward has mean $r_\theta(s,a)$, and the pair itself is chosen by $\pi$, whose law does not involve $\epsilon$. Writing $\ell$ for the log-likelihood in the $\theta$ parameterization, the chain rule gives $\nabla_\epsilon \ell = J^\top \nabla_\theta \ell$ and hence $I^{\mathrm{ind}} = J^\top I_\theta J$. The bracketed matrix is $I_\theta$: it is Theorem 1 applied with trigger matrix the identity, since in the $\theta$ parameterization the tilt is $\sum_j \theta_j B_j$ and the reward shift is $-\kappa\sum_j \theta_j W_j$. Finally $\ker(J^\top I_\theta J) = J^{-1}\ker I_\theta$ for invertible $J$, and $\ker I_\theta = \mathcal{N}_\theta(\pi)$ by Corollary 1. □

Proposition 8 is a statement about one deployment of one policy. Transferring the asymptotic results of §3 needs one more hypothesis, because those results hold the trigger matrix fixed along the sequence while $J$ varies with $\pi_t$.

**Proposition 9** (Asymptotic transfer). *Fix $\epsilon$ with $\|\epsilon\|_\infty \le \bar{\epsilon}$ and let $\pi_t \to \pi_\infty$ with*

$$\sum_t \|J(\pi_t, \epsilon) - J(\pi_\infty, \epsilon)\|_{\mathrm{op}}^2 < \infty. \tag{3}$$

*Then Theorem 2(b) holds in the induced-occupancy model with $D_{g(\pi_\infty)}$ replaced by $J(\pi_\infty, \epsilon)$, and so do Corollaries 1 and 2 and Proposition 1. Condition* (3) *holds whenever $\sum_t \|\pi_t - \pi_\infty\|^2 < \infty$, which is the hypothesis of Corollary 2: $\pi \mapsto J(\pi, \epsilon)$ is Lipschitz on the policy simplex.*

*Proof.* The proof of Theorem 2(b) uses the trigger only twice: to split $O_{sa} D_{g(\pi_t)} v = O_{sa} D_{g(\pi_\infty)} v + O_{sa}(D_{g(\pi_t)} - D_{g(\pi_\infty)})v$ at retained pairs, where the second term contributes $\sum_t \|g(\pi_t) - g(\pi_\infty)\|_\infty^2$, and to bound the blocks at non-retained pairs. Replacing $D_{g(\pi_t)}$ by $J(\pi_t, \epsilon)$ and $\|g(\pi_t) - g(\pi_\infty)\|_\infty$ by $\|J(\pi_t, \epsilon) - J(\pi_\infty, \epsilon)\|_{\mathrm{op}}$ leaves both steps intact, the first by (3) and the second because $\sup_\pi \|J(\pi, \epsilon)\|_{\mathrm{op}} < \infty$ by Lemma 9. For the last claim, $\theta(\cdot, \epsilon)$ extends $C^\infty$ to a neighbourhood of the policy simplex by Lemma 9, so $J$ is $C^1$ on a compact neighbourhood and hence Lipschitz on the simplex, boundary included, giving $\|J(\pi_t, \epsilon) - J(\pi_\infty, \epsilon)\|_{\mathrm{op}} \le L_J \|\pi_t - \pi_\infty\|$. □

It remains to say when $J$ is invertible. Write $q_j(\theta) = \langle \phi_j, d_\pi^\theta \rangle$ for the *induced* trigger, so that (2) reads $\theta_j = \epsilon_j q_j(\theta)$ and $q_j(0) = g_j(\pi)$.

**Lemma 10** (Invertibility). *Write $J_q(\theta) = \partial q / \partial \theta$ for the Jacobian of $q$ and $D_{q(\theta)} = \mathrm{diag}(q(\theta))$, $D_\epsilon = \mathrm{diag}(\epsilon)$, two different matrices. For $\|\epsilon\|_\infty \le \bar{\epsilon}$, $J = (I - D_\epsilon J_q(\theta))^{-1} D_{q(\theta)}$, the inverse existing because $\|D_\epsilon J_q\| < 1$ there. Hence $J$ is invertible if and only if the* diagonal *factor $D_{q(\theta)}$ is, that is if and only if $q_j(\theta(\pi, \epsilon)) \neq 0$ for every $j$. In particular, if no $g_j(\pi)$ vanishes then $J$ is invertible for all small enough $\|\epsilon\|_\infty$; the converse fails, since a coordinate with $g_j(\pi) = 0$ can have $q_j(\theta) \neq 0$ once the response moves the occupancy.*

*Proof.* Differentiating $\theta_j = \epsilon_j q_j(\theta)$ in $\epsilon_k$ gives $J_{jk} = \delta_{jk} q_j(\theta) + \epsilon_j \sum_l (J_q)_{jl}(\theta) J_{lk}$, that is $J = D_{q(\theta)} + D_\epsilon J_q \, J$, so $(I - D_\epsilon J_q) J = D_{q(\theta)}$. The bound $\|D_\epsilon J_q\| \le \|\epsilon\|_\infty \bar{\phi} C_d < 1$ from the proof of Lemma 9 makes $I - D_\epsilon J_q$ invertible, and then $J$ is invertible exactly when the diagonal matrix $D_{q(\theta)}$ is, i.e. when no $q_j(\theta)$ vanishes. Since $\theta \to 0$ and $q_j(\theta) \to g_j(\pi)$ as $\epsilon \to 0$, $g_j(\pi) \neq 0$ for all $j$ gives invertibility for small $\epsilon$. □

`induced_check.py` and `induced_jinv.py` verify these statements numerically: the information identity of Proposition 8 to $10^{-11}$, the closed form of Lemma 10 against a finite-difference Jacobian to $10^{-11}$, $J \to D_g$ as $\epsilon \to 0$, and the blind dimensions of the two models equal (for instance 8 and 14 at $p = 24$ and 30) with pullback residual $10^{-16}$. They also exhibit the failure of the converse in Lemma 10: choosing $\phi_j \perp d_\pi^0$ gives $g_j(\pi) = 0$ exactly while $q_j(\theta) = -1.7 \times 10^{-3}$ and $J$ stays invertible.

Two consequences. First, the test of §3 applies to the standard formulation: compute $\mathcal{N}_\theta$ from the retained pairs, then pull it back through the Jacobian $J$, which Lemma 10 gives in closed form at the cost of one linear solve. Second, under (3) the dimension counts are unchanged, so Corollary 2 and Proposition 1 transfer. What does change is the identity of the blind directions: a direction blind in our model is blind in the induced model only after the pullback, so the two models can disagree about *which* parameters are hidden while agreeing on *how many*. The restriction to small $\|\epsilon\|_\infty$ is what buys a well-defined trigger; for a large response the fixed point (2) need not be unique, and we make no claim there.

## N Partial observation: when the reward response is not observed

**Model.** Suppose the learner observes transitions but not the performative part of the reward, for example because the cost of the response is borne outside the trajectory or realized later. Formally $\sigma_r = \infty$: the observation block is $O_{sa} = B_{sa}$, and every result of §3–4 holds with $B_{sa}$ in place of $O_{sa}$, with dimension law $\dim \mathcal{N} \ge p - |S|(|S| - 1)$ and greedy guarantee $\mathsf{H}(|S| - 1)$. What changes is Proposition 3: a plan's value still depends on the reward response at its own pairs, but its deployments no longer reveal it.

**Separation.** With $\Pi_\Theta$ and $E_v = \{(s,a) : B_{sa} D_{g(\pi)} v \neq 0 \text{ for some } \pi\}$ as in §4, call $v$ *separated* on $\Theta$ if some $\epsilon^0, \epsilon^0 + hv \in \Theta$ have no common optimal policy, with $\|v\| = 1$, while no plan in $\Pi_\Theta$ puts mass on $E_v$.

**Proposition 10** (Plausibly optimal deployments stay blind)**.** *Under partial observation, if $v$ is separated on $\Theta$, every algorithm that deploys only plans in $\Pi_\Theta$ has $\mathcal{V}_T(v) = 0$, and hence regret at least $\Delta T/8$ on one of the two instances. This includes certainty equivalence, Thompson sampling with any prior and any posterior inflation whose samples lie in $\Theta$, and optimism over any confidence set contained in $\Theta$.*

*Proof.* A trajectory visits only pairs in the support of the deployed plan, so every term of $\mathcal{V}_T(v)$ vanishes, and Theorem 4 applies. □ □

This is the observation behind information-directed sampling (Russo and Roy, 2018). Separation arises when a coordinate acts on the reward at one pair and on the dynamics only at another pair that no plausibly optimal plan uses. Optimism then deploys the plan that is optimal when the unseen response is benign, so its regret is linear exactly when the response is harmful. Under the standard model the same coordinate is revealed by the reward at the decision pair, and Proposition 3 rules separation out.

## O THE GRID-WORLD BENCHMARK OF MANDAL ET AL.

We re-implemented the two-agent grid-world used to evaluate repeated retraining by Mandal et al. (2023) and Rank et al. (2024), from the authors' released code: an $8 \times 8$ grid with cell rewards $-0.01$, $-0.02$ and $-0.5$ and $+1$ at the goal, a principal choosing among four moves, followers who may override it at a cost of $0.05$, a Boltzmann response ($\beta = 5$) to the followers' averaged optimal $Q$ functions on privately perturbed grids ($\epsilon = 0.3$, 50 followers), $\gamma = 0.9$ and regularization $\lambda = 1$. For identifiability the unknown is the cell-cost grid seen by a representative follower ($\theta \in \mathbb{R}^{64}$, true value the mean perturbed grid), and the observation Jacobian at each pair is computed exactly through the follower's greedy policy (checked against finite differences to $5 \times 10^{-9}$). This is a check of the theory's predictions; no learner is compared.

*Coverage fails.* Repeated policy optimization, their algorithm, converges within about ten iterations to a policy that puts zero occupancy on 61 of the 256 state–action pairs; damped best response puts zero occupancy on 192. Their overlap assumption holds along the path (the ratio of the next iterate to the current occupancy tends to one).

*Decision-relevant identifiability survives.* The follower responds to states, not to the principal's action, and its Boltzmann response keeps every state visited. The observation Jacobian over the visited pairs has rank 63 of 64 along both algorithms and for every $\beta \in \{5, 20, 50, 200\}$; the one blind direction, a joint shift of the goal cell and a uniform shift elsewhere, leaves the response unchanged and cannot affect any decision. After 40 rounds of 20,000 steps, one to six directions have information below 100. Theorem 2 predicts both facts: every direction that matters is revealed by some retained pair. Blind spots require response components that only the abandoned actions reveal, as in the classes of §5; on this benchmark, retraining toward stability loses coverage but not identifiability.