DEEP EXPLORATION AS SEQUENTIAL TESTING IN Q-FUNCTION SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Best-policy identification in Markov decision processes (MDPs) is fundamentally an information-acquisition problem: the learner must gather trajectories that resolve uncertainty about which sequences of actions are optimal. We formulate this objective as a sequential testing problem in the space of optimal Q-functions, where each policy induces a region of optimality and exploration seeks to identify the region containing the unknown Bellman fixed point. We introduce TQS , a model-free algorithm that implements this testing problem through a leader–challenger mechanism. At each round, TQS pairs an empirical estimate of $Q^\star$ with a challenger sampled from a Gibbs distribution based on empirical Bellman residuals, and uses the challenger's squared fitted Bellman residual as an adaptive exploration reward. Since this reward changes across rounds, TQS uses an online RL learner as a dedicated exploration controller, steering the agent toward trajectories that maximally separate the competing hypotheses. In linear MDPs, we show that TQS admits a saddle-point interpretation in which the behavior policy maximizes distinguishability against a randomized alternative, and we derive Gibbs-contraction and sample-complexity guarantees governed by an instance-dependent oracle rate. Unlike model-posterior methods, TQS operates directly in value-function space and extends naturally to neural function approximation. We instantiate a value-disagreement surrogate with Bayesian methods, and obtain strong results on large tabular and continuous state benchmarks.

1 INTRODUCTION

We study best-policy identification in discounted infinite-horizon Markov decision processes (MDPs). The learner interacts with an unknown environment during a dedicated exploration phase and must ultimately recommend a policy whose performance is close to optimal. Unlike regret minimization, the rewards collected during exploration are not themselves part of the objective: what matters is how quickly the collected trajectories reduce uncertainty about the optimal policy. The central question of this work is therefore how to design an exploration policy that gathers decision-relevant information as efficiently as possible.

Most exploration methods in reinforcement learning were designed for settings in which the agent must simultaneously learn and collect reward. Optimism, posterior sampling, and exploration bonuses address this trade-off by making uncertain regions attractive to a reward-seeking policy (Brafman & Tennenholtz, 2002; Kearns & Singh, 2002; Jaksch et al., 2010; Osband et al., 2013; 2019). For best-policy identification, however, this coupling is unnecessary and can be highly inefficient. An informative trajectory need not have high estimated return, and reaching a decision-relevant state may require traversing actions that appear suboptimal under every plausible reward-maximizing policy. This distinction has formal consequences: Wagenmaker et al. (2022) construct episodic MDPs on which any low-regret algorithm requires arbitrarily more samples than a dedicated policy-identification algorithm. This motivates designing exploration as an information-acquisition problem, with exploration policies designed explicitly to collect observations that distinguish competing policy hypotheses.

Recent near-optimal approaches to best-policy identification build on optimal experimental design and information-directed posterior sampling to focus exploration on unresolved policy comparisons (Wagenmaker et al., 2022; Wagenmaker & Jamieson, 2022; Kone & Jamieson, 2026). These approaches

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.

construct exploration strategies using either an explicit class of candidate policies or a posterior distribution over the unknown MDP. Both representations can be computationally demanding in large or continuous state spaces. We therefore seek a model-free formulation that carries out policy discrimination directly in Q-function space, without requiring explicit policy enumeration or posterior inference over transition dynamics.

We formulate policy discrimination as sequential testing in Q-function space, guided by a leader-challenger principle. To certify a candidate policy (the leader) as near-optimal, we must rule out the possibility that another policy performs substantially better. The challengers are plausible Q-functions under which such an improvement remains possible.

This makes exploration concrete: collect observations that distinguish the current estimate of $Q^\star$ from these alternatives. Building on this perspective, we introduce TQS , a model-free exploration algorithm based on randomized challengers in the Q-function. In each round, TQS constructs a leader $\hat{Q}_t$, representing its current estimate of $Q^\star$, and an alternative $\widetilde{Q}_t$ which is another plausible candidate for $Q^\star$ that induces a different optimal policy than $\hat{Q}_t$.

To resolve this ambiguity, the learner seeks observations that provide evidence to potential rule out the challenger. Since $Q^\star$ satisfies the Bellman optimality equation, a natural exploration objective is to visit state-action pairs where the challenger violates that equation. This motivates the squared conditional Bellman residual

$$\ell_{\widetilde{Q}_t}(s,a) \coloneqq \tfrac{1}{2}\big(r(s,a) + \gamma P V_{\widetilde{Q}_t}(s,a) - \widetilde{Q}_t(s,a)\big)^2, \qquad V_Q(s) \coloneqq \max_a Q(s,a).$$

Maximizing this reward directs exploration toward trajectories that expose the challenger's inconsistency with the environment.

The learner estimates this residual from stored transitions and uses its squared fitted value as an exploration reward $r_t^{\text{ex.}}$. By optimizing this reward, the agent collects trajectories that are specifically informative to resolve the current policy ambiguity.

As the leader and challenger evolve with the collected data, the exploration reward forms a non-stationary sequence of reward kernels. We optimize this sequence using an online RL algorithm. Its objective determines long-run state-action visitation, whereas the final recommended policy is evaluated under the original discounted criterion. To our knowledge, this is the first use of an online RL learner as an exploration subroutine for discounted infinite-horizon policy identification.

Our analysis for linear MDPs shows that TQS attains an instance-dependent oracle rate for separating $Q^\star$ from alternative Q-functions. This rate governs both posterior contraction and the expected number of interactions required for policy identification. The guarantee follows from a game-theoretic analysis that couples the exploration learner, which seeks informative observations to competing Q-function hypotheses, with the challenger sampler, which identifies alternatives that remain difficult to exclude.

Since TQS operates directly in value-function space, it requires neither an explicit transition model nor posterior sampling over the dynamics. Its leader–challenger design also admits an instantiation with neural function approximation. Our deep RL implementation uses a quadratic surrogate to generate approximately Gaussian challengers and a separate model-free learner to optimize the resulting value-disagreement reward.

We evaluate TQS on DeepSea and Cartpole swingup, spanning tabular problems and continuous-state problems with neural value approximation. The results show improved sample efficiency in policy identification and recommendation quality, with particularly pronounced gains in settings requiring coordinated exploration to reach informative states. These findings support the practical value of combining randomized Q-function challengers with a dedicated exploration controller.

Our contributions are fourfold. First, we formulate best-policy identification as sequential testing in optimal Q-function space and construct a calibrated testing score from cumulative Bellman residuals (§3). Second, we introduce TQS , a leader–challenger algorithm that samples alternative Q functions through Gibbs weighting and uses their fitted Bellman residuals to define rewards for a dedicated online MDP exploration learner. Third, under linear realizability and we establish a game-theoretic characterization of Gibbs contraction, and stopping guarantees governed by an instance-dependent oracle rate (§4). Finally, we develop a deep instantiation using deep Bayesian approximation techniques, and evaluate its sample efficiency on large tabular and continuous-state benchmarks (§5).

1.1 Related work

In this section, we review the literature most closely related to our work. Since our contribution lies at the intersection of best policy identification and deep exploration, and given the rapid expansion of the deep RL literature, we focus on the approaches that are most relevant to our setting.

Randomized value functions and deep exploration. RLSVI randomizes Bellman regression to generate value-function samples (Osband et al., 2019), while Bootstrapped DQN and randomized prior functions implement related ideas with neural ensembles (Osband et al., 2016; 2018). These methods obtain exploration by acting greedily with respect to a sampled value function. Other ensemble methods construct optimistic action values (Chen et al., 2017), while information-directed exploration balances estimated regret against information gain (Nikolov et al., 2019). TQS uses randomization to select a challenger to the current recommendation, then learns an exploration policy that seeks trajectories separating the two hypotheses. Its objective is to resolve policy uncertainty, without trading information acquisition against reward collected during learning.

Intrinsic rewards and entropy-based exploration. Intrinsic rewards encourage exploration through visitation novelty (Bellemare et al., 2016) or prediction error (Burda et al., 2019). More closely related, disagreement-based exploration, MAX, and Plan2Explore use ensembles of dynamics models to seek informative observations (Pathak et al., 2019; Shyam et al., 2019; Sekar et al., 2020). DeRL explicitly separates exploration and exploitation policies (Schäfer et al., 2022). Entropy-based approaches pursue different objectives: Soft Actor-Critic augments task return with action entropy (Haarnoja et al., 2018), while maximizing state-visitation entropy promotes broad coverage (Hazan et al., 2019). Reward-free exploration collects data for reward functions specified after exploration (Jin et al., 2020). TQS ties its reward to a policy-certification test: it samples challengers outside the current recommendation's certification region, favoring alternatives that remain difficult to exclude. For a fixed leader–challenger pair, squared Q-disagreement is the contribution of a visit to the discrepancy underlying this test. The exploration objective is therefore determined by the alternatives that obstruct policy certification.

Policy identification and experimental design. PEDEL combines online experimental design with policy elimination in linear MDPs (Wagenmaker & Jamieson, 2022), but its computational cost depends on an explicit policy class that can be exponentially large. Under generative access, policy identification has also been studied with linear models (Taupin et al., 2023) and kernel value approximation through AE-LSVI (Li et al., 2023). These latter methods can query selected state-action pairs directly; TQS must reach informative states through navigation. Our leader–challenger construction builds on top-two posterior sampling (Russo, 2016) and game-based experimental design (Degenne et al., 2020). In linear bandits, PEPS samples parameters whose optimal arm differs from the current recommendation and updates an online sampling allocation using squared predictive disagreement (Li et al., 2024). In tabular RL, PIPS samples challenger MDPs and trains an exploration subroutine using model KL divergence (Kone & Jamieson, 2026). TQS carries out the comparison directly in optimal Q-function space, replacing challenger MDPs with candidate Q-functions and using their squared disagreement with the current estimate as the exploration reward. This formulation avoids posterior inference over transition dynamics and allows the exploration mechanism to be implemented with neural value estimators.

To collect informative trajectories, our exploration mechanism relies on online learning in infinite-horizon MDPs with changing rewards (Even-Dar et al., 2009). This subroutine optimizes the adaptive residual reward to guide data collection. The policy eventually returned by TQS is greedy with respect to its estimate of $Q^\star$ and is evaluated under the original discounted reward setting.

2 Formal setting

We consider a discounted Markov decision process $M := \langle \mathcal{S}, \mathcal{A}, p, r, \gamma, \nu_0 \rangle$, where $\mathcal{S}$ is a measurable, possibly continuous state space, $\mathcal{A}$ is finite, p is the transition kernel, r is the expected reward function, and $\gamma \in [0, 1)$ is the discount factor. The initial state has distribution ν_0. At round t, the learner selects A_t, observes R_t, and transitions to S_{t+1}. Conditional on the history and action, the observation follows the MDP kernel, with mean reward $r(S_t, A_t)$. We assume $|R_t| \le R_{\max}$ almost surely for a known $R_{\max} > 0$. Write $PV(s, a) := \int V(s')p(\mathrm{d}s' \mid s, a)$. The performance of a stationary policy π

from ν_0 is

$$J(\pi) \coloneqq \mathbb{E}_{\substack{s_1\sim\nu_0\\ \pi,p}}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_t\right]$$

Its value and action-value functions are respectively defined by

$$V^\pi(s) \coloneqq \mathbb{E}_{\pi,p}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_t \,\middle|\, s_1 = s\right], \quad \text{and} \quad Q^\pi(s,a) \coloneqq \mathbb{E}_{\pi,p}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_t \,\middle|\, s_1 = s,\ a_1 = a\right].$$

They satisfy the Bellman equations

$$Q^\pi(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[V^\pi(s')\right], \quad \text{where} \quad V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right].$$

Under standard measurability assumptions, there exists a stationary deterministic optimal policy (Puterman, 1994). Its optimal action-value function $Q^\star$ is the unique solution of the Bellman optimality equation

$$Q^\star(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} Q^\star(s',a')\right].$$

Any policy that is greedy with respect to $Q^\star$ is optimal. The optimal value is denoted by

$$J^\star \coloneqq \sup_\pi J(\pi) = J(\pi^\star).$$

In *best-policy identification*, the learner interacts with the MDP during a dedicated exploration phase and stops at a data-dependent time τ to recommend a policy $\hat{\pi}_\tau$. Its goal is to identify a policy whose discounted performance is within ε of $J^\star$ using as few interactions as possible.

Definition 1 ((ε,δ)-PAC best-policy identification)**.** *An algorithm is (ε,δ)-PAC for best-policy identification if, for every MDP in the considered model class, it stops at an almost surely finite stopping time τ and returns a policy $\hat{\pi}_\tau$ satisfying*

$$\mathbb{P}_M\left(J^\star - J(\hat{\pi}_\tau) > \varepsilon\right) \leqslant \delta.$$

Its expected sample complexity is $\mathbb{E}_M[\tau]$.

Linear MDP. For the theoretical analysis, we assume a linear MDP with known measurable features $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ satisfying $\sup_{s,a}\|\phi(s,a)\|_2 \leq L_\phi < \infty$: there exist an unknown vector $w \in \mathbb{R}^d$ and a vector of finite signed measures μ such that

$$r(s,a) = \phi(s,a)^\top w, \qquad P(\mathrm{d}s' \mid s,a) = \phi(s,a)^\top \mu(\mathrm{d}s').$$

Consequently, the Bellman optimality operator preserves the linear function class, and

$$Q^\star(s,a) = \phi(s,a)^\top w^\star.$$

We consider candidates $Q_\theta = \phi^\top\theta$ with $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^d$ is compact and convex and $w^\star \in \mathrm{int}(\Theta)$.

Further notation. We write $\Delta(\mathcal{X})$ for the set of probability distributions on a measurable space $\mathcal{X}$ and $\|x\|_A^2 \coloneqq x^\top Ax$ for a positive semi-definite matrix A. Let $\mathcal{F}_t$ denote the σ-algebra generated by the history available at the beginning of round t, including $\mathcal{D}_t$ and the learner's past randomization. Unless otherwise specified, $\mathbb{P}$ and $\mathbb{E}$ denote probability and expectation under the joint randomness of the environment and the learning algorithm.

3 Sequential testing in optimal Q-function space

We now present the principle underlying TQS . A sequential test measures the evidence against candidate optimal Q-functions and certifies a policy once its alternatives have been excluded. We construct this evidence from empirical Bellman residuals, then identify the local contributions to its growth. These contributions determine an exploration reward, while the alternatives that remain difficult to exclude determine which rewards the exploration controller should pursue. Algorithm 1, at the end of this section, presents the pseudocode for TQS .

3.1 Policy regions and sequential certification

Let $\mathcal{Q}$ be a class containing $Q^\star$. For a stationary policy π, define its decision gap under $Q \in \mathcal{Q}$ and its certification region by

$$g_Q^\pi(s) := \max_{a \in \mathcal{A}} Q(s,a) - \mathbb{E}_{a \sim \pi(\cdot|s)}[Q(s,a)], \qquad \mathcal{Q}_\pi^\varepsilon := \{Q \in \mathcal{Q} : \|g_Q^\pi\|_\infty \le (1-\gamma)\varepsilon\}.$$

The Bellman equations give $V^\star - V^\pi = (I - \gamma P^\pi)^{-1} g_{Q^\star}^\pi$, hence

$$Q^\star \in \mathcal{Q}_\pi^\varepsilon \quad \Longrightarrow \quad \|V^\star - V^\pi\|_\infty \le \varepsilon \quad \Longrightarrow \quad J^\star - J(\pi) \le \varepsilon.$$

Having $Q^\star \in \mathcal{Q}_\pi^\varepsilon$ certifies that π is ε-optimal. At $\varepsilon = 0$, the region consists of functions for which π selects only maximizing actions. Write $\mathrm{Alt}_\varepsilon(\pi) := \mathcal{Q} \setminus \mathcal{Q}_\pi^\varepsilon$ for the alternatives preventing this certificate.

Suppose a data-dependent statistic $S_t(Q)$ and boundary $\beta_t(\delta)$ satisfy

$$\mathbb{P}\big(\forall t \ge 1 : \; S_t(Q^\star) \le \beta_t(\delta)\big) \ge 1 - \delta.$$

Then $C_t(\delta) := \{Q \in \mathcal{Q} : S_t(Q) \le \beta_t(\delta)\}$ is a confidence sequence for $Q^\star$. At round t, an estimator using the first $t-1$ transitions, stored in $\mathcal{D}_t$, supplies a leader $\hat{Q}_t$ and a recommendation $\hat{\pi}_t \in \mathrm{Greedy}(\hat{Q}_t)$. The evidence supporting this recommendation is limited by its least-rejected alternatives. This motivates the statistic and stopping rule

$$T_t := \inf_{Q \in \mathrm{Alt}_\varepsilon(\hat{\pi}_t)} S_t(Q), \qquad \tau_\delta := \inf\{t \ge 1 : T_t > \beta_t(\delta)\},$$

with $\inf \varnothing = +\infty$. When the test passes, $C_t(\delta) \subseteq \mathcal{Q}_{\hat{\pi}_t}^\varepsilon$. On the event in (3.1), the recommendation therefore satisfies (3.1), giving $\mathbb{P}(\tau_\delta < \infty,\ J^\star - J(\hat{\pi}_{\tau_\delta}) > \varepsilon) \le \delta$. This argument separates confidence calibration from termination: the former controls correctness of the identification, while the latter requires collecting enough evidence against the alternatives.

3.2 Bellman residuals as accumulated evidence

To construct such a score without estimating transition dynamics, consider the observed Bellman residual of a candidate Q:

$$Y_i(Q) := R_i + \gamma V_Q(s_{i+1}) - Q(s_i, a_i), \qquad V_Q(s) := \max_a Q(s,a).$$

Its conditional mean, given the history and the selected state-action pair, is denoted $e_Q(s_i, a_i)$, defined as $e_Q := r + \gamma P V_Q - Q$. The Bellman optimality equation gives $e_{Q^\star} = 0$. Thus the residual is conditionally centered at the optimal Q function, whereas a candidate violating the optimal Bellman equation can produce a systematic drift.

Define $x_i := \phi(s_i, a_i)$, the feature vector of the observed state-action pair, and define $G_t := \sum_{i<t} x_i x_i^\top$ the design matrix and $\Lambda_t := \lambda I_d + G_t$, its regularized counterpart, with $\lambda > 0$. For each candidate Q, we aggregate its Bellman residuals over the observed transitions and define the score

$$m_t(Q) := \sum_{i<t} x_i Y_i(Q), \quad S_t(Q) := \frac{1}{2\sigma_\star^2} \|m_t(Q)\|_{\Lambda_t^{-1}}^2.$$

Each candidate is evaluated on every observed transition. At $Q^\star$, the Bellman residuals are conditionally centered. Thus $m_t(Q^\star)$ is a martingale sum. A self-normalized concentration argument, proved in Appendix A.2, gives

$$\mathbb{P}\big(\forall t \ge 1 : \; S_t(Q^\star) \le \beta_t(\delta)\big) \ge 1 - \delta, \text{ where } \quad \beta_t(\delta) := \log \frac{1}{\delta} + \frac{1}{2} \log \frac{\det \Lambda_t}{\lambda^d},$$

where the score uses the noise bound $\sigma_\star = R_{\max}/(1-\gamma)$. This guarantee holds independently of the leader estimator or sampling rule and justifies using S_t in the stopping rule (3.1).

3.3 From accumulated evidence to an exploration reward

The stopping statistic is limited by alternatives with small scores. Exploration should therefore collect transitions that increase the evidence against these alternatives. To identify where this evidence can be collected, we express the score in terms of fitted Bellman residuals.

For a candidate Q, let $\hat{b}_t(Q)$ be the ridge-regression estimate obtained by regressing its observed residuals $Y_i(Q)$ onto the features x_i:

$$\hat{b}_t(Q) \coloneqq \arg\min_{b\in\mathbb{R}^d} \left\{ \sum_{i<t} \left(Y_i(Q) - x_i^\top b\right)^2 + \lambda \|b\|_2^2 \right\} = \Lambda_t^{-1} m_t(Q),$$

and define $\hat{e}_t(Q; s, a) \coloneqq \phi(s, a)^\top \hat{b}_t(Q)$. Since $m_t(Q) = \Lambda_t \hat{b}_t(Q)$ and $\Lambda_t = \lambda I_d + \sum_{i<t} x_i x_i^\top$,

$$S_t(Q) = \sum_{i<t} \frac{\hat{e}_t(Q; s_i, a_i)^2}{2\sigma_\star^2} + \frac{\lambda}{2\sigma_\star^2} \|\hat{b}_t(Q)\|_2^2.$$

Thus the score aggregates squared fitted Bellman residuals along the observed trajectory, together with a regularization term.

Equation (3.3) evaluates all past transitions using the latest fit $\hat{e}_t$. To guide exploration sequentially, we instead seek a representation in which each transition contributes a reward determined from the preceding observations. This leads us to replace the current fit on past transitions with the successive fits $\hat{e}_i$ available at each round. We show (Appendix C.2) that with high probability, uniformly over $\theta \in \Theta$

$$S_t(Q_\theta) \underset{t\to\infty}{\sim} \sum_{i<t} \frac{\hat{e}_i(Q; s_i, a_i)^2}{2\sigma_\star^2}.$$

The above suggests defining the *per step* exploration reward against a fixed candidate Q as

$$\hat{\ell}_{t,Q}(s, a) \coloneqq \frac{\hat{e}_t(Q; s, a)^2}{2\sigma_\star^2}.$$

Consequently, for a fixed candidate, maximizing accumulated exploration rewards maximizes the growth of evidence against it. Collecting this evidence requires learning to reach informative state-action pairs. Moreover, changing challengers produce a sequence of changing reward functions. We therefore use an online MDPs learner to guide exploration. The next subsection specifies how challengers are selected and how their rewards update this learner.

3.4 Randomized challengers and exploration control

To select alternatives against which to collect evidence, TQS samples a challenger from the Gibbs distribution truncated to the alternative set of the current recommendation:

$$\widetilde{Q}_t \sim \Pi_t^{\mathrm{Alt}}, \qquad \Pi_t^{\mathrm{Alt}}(\mathrm{d}Q) \propto \mathbf{1}\{Q \in \mathrm{Alt}_\varepsilon(\hat{\pi}_t)\} e^{-\eta_t S_t(Q)} \Pi_0(\mathrm{d}Q).$$

Here Π_0 is a reference probability measure and $\eta_t > 0$ controls how strongly the distribution favors alternatives with small scores. The restriction ensures that each challenger prevents certification of $\hat{\pi}_t$, while the weighting favors alternatives that remain difficult to exclude.

The exploration reward against the sampled challenger is

$$\hat{r}_t^{\mathrm{ex}}(s, a) \coloneqq \hat{\ell}_{t,\widetilde{Q}_t}(s, a) = \frac{\hat{e}_t(\widetilde{Q}_t; s, a)^2}{2\sigma_\star^2}.$$

Thus the leader provides the policy recommendation, and the challenger determines where further evidence is sought. An online MDP learner Reg learns to reach state-action pairs with large exploration rewards. For a stationary policy μ with stationary state-action distribution d^μ, its gain under the current reward is

$$\bar{r}_t^{\mathrm{ex}}(\mu) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_\mu \left[\sum_{k=1}^{n} \hat{r}_t^{\mathrm{ex}}(S_k, A_k) \right] = \langle d^\mu, \hat{r}_t^{\mathrm{ex}} \rangle.$$

Because the challenger and its fitted residual change across rounds, the explorer must adapt to a changing reward sequence. Its objective is to compete with the best fixed stationary policy for that sequence.

Algorithm 1 combines these components. At each round, the learner updates its recommendation and checks the stopping condition. Reg supplies the exploration policy. The challenger's fitted Bellman residual defines the exploration reward $\hat{r}_t^{\mathrm{ex}}$. After executing the action, the learner appends the observed transition and task reward to $\mathcal{D}_t$ and updates Reg using the exploration reward. The current challenger thus guides subsequent exploration decisions. TQS can also be run without a stopping rule to collect informative samples online. The following section analyzes evidence accumulation and Gibbs contraction under continuing exploration, then derives correctness and sample-complexity guarantees for the stopping rule.

Algorithm 1: TQS : sequential testing in Q-function space

input : Accuracy ε, confidence δ; leader estimator; features ϕ and ridge parameter $\lambda > 0$; prior Π_0 and scales (η_t); online MDP learner Reg.

Initialize $\mathcal{D}_1 \leftarrow \varnothing$ and observe S_1
for $t = 1, 2, \ldots$ **do**
  Estimate $\hat{Q}_t$ from $\mathcal{D}_t$ and choose $\hat{\pi}_t \in \mathrm{Greedy}(\hat{Q}_t)$
  if $T_t > \beta_t(\delta)$ **then**
    return $\hat{\pi}_t$
  Draw $\widetilde{Q}_t$ from (3.4)
  Fit $\hat{e}_t(\widetilde{Q}_t; \cdot, \cdot)$ using $\mathcal{D}_t$ and form $\hat{r}_t^{\mathrm{ex}}$ from (3.4)
  Execute A_t and observe (R_t, S_{t+1})
  Add (S_t, A_t, R_t, S_{t+1}) to $\mathcal{D}_t$ to obtain $\mathcal{D}_{t+1}$
  Pass $\hat{r}_t^{\mathrm{ex}}$ and S_t to Reg and update its proposed action A_{t+1}

4 Theory

We consider exact identification ($\varepsilon = 0$) under the linear representation of Section 2.

Assumption A.1. *Every stationary policy induces an irreducible state chain, and* $\{\phi(s,a) : s \in \mathcal{S}, a \in \mathcal{A}\}$ *spans* $\mathbb{R}^d$. *We further assume the optimal policy* $\pi^\star$ *is unique. We also assume the MDP is* uniformly fast mixing *with coefficient* $\rho \in (0,1)$.

$$\sup_{\pi} \sup_{s \in \mathcal{S}} \|P_\pi^n(s,\cdot) - \nu_\pi\|_{\mathrm{TV}} \le C_{\mathrm{mix}} \rho^n, \qquad n \ge 0.$$

Assume that Θ is compact and convex with nonempty interior, and that Π_0 has a continuous, strictly positive density on Θ (e.g. uniform distribution).

For a stationary policy μ, let d^μ denote its stationary state-action distribution. The set of feasible stationary allocations is

$$\Omega(P) := \Big\{\omega \in \Delta(\mathcal{S} \times \mathcal{A}) : \sum_a \omega(s,a) = \sum_{s',a'} \omega(s',a') P(s \mid s',a') \quad \forall s\Big\}.$$

Recall that $e_\theta(s,a) = r(s,a) + \gamma P V_\theta(s,a) - Q_\theta(s,a)$. Under an allocation ω, the accumulated evidence against Q_θ grows at rate $\sum_{s,a} \omega(s,a) e_\theta(s,a)^2/(2\sigma_\star^2)$. The largest achievable separation from the hardest alternative under this metric is therefore

$$\Gamma^\star := \sup_{\omega \in \Omega(P)} \inf_{\theta \in \mathrm{Alt}_0(\pi^\star)} \frac{1}{2\sigma_\star^2} \sum_{s,a} \omega(s,a) e_\theta(s,a)^2.$$

Here $\mathrm{Alt}_0(\pi^\star) = \{\theta \in \Theta : \pi^\star \notin \mathrm{Greedy}(Q_\theta)\}$. Although its maximizing allocation depends on the unknown environment, TQS attains the corresponding evidence rate.

Theorem 2 (Evidence growth)**.** *Under the preceding conditions, running* TQS *with* $\eta_t = O(t^{-\kappa})$, $\kappa \in (0,1)$, *and without stopping, there exists a deterministic,* $f(t) = o(t)$ *and* $t_0 < \infty$ *such that*

$$\mathbb{P}\big(\mathsf{T}(t) \ge t\Gamma^\star - f(t)\big) \ge 1 - t^{-2} \qquad (t \ge t_0),$$

where $\mathsf{T}(t) = \inf_{\theta \in \mathrm{Alt}_0(\hat{\pi}_t)} S_t(\theta)$, *In particular,* $\liminf_{t\to\infty} \mathsf{T}(t)/t \ge \Gamma^\star$ *almost surely.*

To characterize the converge of TQS consider the uninflated Gibbs distribution

$$\Pi_t(\mathrm{d}\theta) \propto e^{-S_t(\theta)}\Pi_0(\mathrm{d}\theta)\mathbf{1}\{\theta \in \Theta\}$$

The next result identifies its exact contraction rate and shows that no other trajectory-collection rule can improve that rate within this Gibbs construction.

Theorem 3 (Gibbs contraction rate)**.** *Under the conditions of Theorem 2,* TQS *satisfies,*

$$\lim_{t\to\infty} -t^{-1} \cdot \log \Pi_t\big(\mathrm{Alt}_0(\pi^\star)\big) = \Gamma^\star, \quad \textit{almost surely.}$$

Moreover, under the same problem and prior conditions, every nonanticipating algorithm A *satisfies,*

$$\limsup_{t\to\infty} -t^{-1} \cdot \log \Pi_t^{\mathsf{A}}\big(\mathrm{Alt}_0(\pi^\star)\big) \le \Gamma^\star, \quad \textit{almost surely.}$$

The main difficulty is that challengers, fitted residuals, and exploration policies evolve together along a single trajectory. The analysis must control their estimation and sampling errors while accounting for the dynamics that constrain where evidence can be collected. This connects the local rewards of Section 3 to the global separation rate (4). The proofs are given in Appendices B.4 and C.2.

The evidence rate translates directly sample complexity bounds. Consider

$$\tau_\delta := \inf\{t \ge 1 : \mathsf{T}(t) > \beta_t(\delta)\},$$

with the anytime boundary from (3.2). Since $\beta_t(\delta) = \log(1/\delta) + O(\log t)$, the evidence required for certification is reached on the scale $\log(1/\delta)/\Gamma^\star$.

Theorem 4 (Correctness and expected sample complexity)**.** *Under the conditions of Theorem 2, the test-based stopping is δ-correct and satisfies*

$$\limsup_{\delta\downarrow 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \le \frac{1}{\Gamma^\star}.$$

Correctness follows from anytime calibration and the high-probability evidence bound controls the expected stopping time (Appendix C.3).

A computational alternative to (4) avoids minimizing the score over the alternative set (potentially untractable). At round t, draw candidates independently from Π_t until an alternative to $\hat{\pi}_t$ is found or a budget of $M(t,\delta)$ draws is exhausted. In the first case, continue exploration; in the second, stop and return $\hat{\pi}_t$. The first sampled alternative also follows the required challenger law. Writing these candidates as $\vartheta_{t,m}$, the rule is

$$\tau_\delta^{\mathrm{samp}} := \inf\left\{t \ge 1 : \vartheta_{t,m} \notin \mathrm{Alt}_0(\hat{\pi}_t) \text{ for every } m \le M(t,\delta)\right\}.$$

Theorem 5 (Sampling-based stopping)**.** *Under the conditions of Theorem 2, assume additionally that $\theta^\star \in \mathrm{int}(\Theta)$. With the budget $M(t,\delta)$ as defined in Appendix C.4,*

$$\limsup_{\delta\downarrow 0} \mathbb{P}\big(\hat{\pi}_{\tau_\delta^{\mathrm{samp}}} \ne \pi^\star\big)/\delta \le 1, \quad \textit{and} \quad \limsup_{\delta\downarrow 0} \frac{\mathbb{E}[\tau_\delta^{\mathrm{samp}}]}{\log(1/\delta)} \le \frac{1}{\Gamma^\star}.$$

Thus sampling-based stopping retains the leading expected sample-complexity bound with asymptotic δ-correctness, using Gibbs draws and policy comparisons in place of constrained score minimization.

5 Experiments

We evaluate sample efficiency on DeepSea and Cartpole swingup (Osband et al., 2020), varying the grid depth N and swingup threshold h to increase exploration difficulty. DeepSea requires coordinated exploration over N steps; Cartpole swingup tests neural value approximation with continuous observations. Against training transitions, we report DeepSea identification rates over 50 independent seeds, and Cartpole swingup recommendation returns and the fraction exceeding 100. We compare with Bootstrapped DQN (Osband et al., 2016), DQN (Mnih et al., 2015), PPO (Schulman

et al., 2017) and RLSVI. An ablation replaces the exploration controller with the challenger's greedy policy, retaining the same estimator and challenger construction.

Our implementation uses a value-disagreement surrogate motivated by the exact decomposition

$$Y_i(Q) = Y_i(\widehat{Q}_t) + (\widehat{Q}_t - Q)(s_i, a_i) + \gamma(V_Q - V_{\widehat{Q}_t})(s_{i+1}).$$

We note that $Y_i(\widehat{Q}_t) \approx 0$, as $\widehat{Q}_t \to Q^\star$, further the next-state γ-correction, gives $\widehat{e}_t(Q; s, a) \approx (\widehat{Q}_t - Q)(s, a)$, yielding the exploration reward $\hat{r}_t^{\text{ex}}(s, a) = (\widehat{Q}_t - \widetilde{Q}_t)(s, a)^2/(2\sigma_\star^2)$. For a linear last layer with fixed features, the corresponding surrogate score is $\widetilde{S}_t(\theta) = \frac{1}{2\sigma_\star^2}\|\theta - \widehat{\theta}_t\|_{G_t}^2$. With Π_0 uniform, the unrestricted Gibbs distribution Π_t is Gaussian and can be sampled by randomized least squares. We run the algorithm without stopping and evaluate the quality of its policy recommendations as exploration proceeds. Appendix D.1 further describes all implementations details.

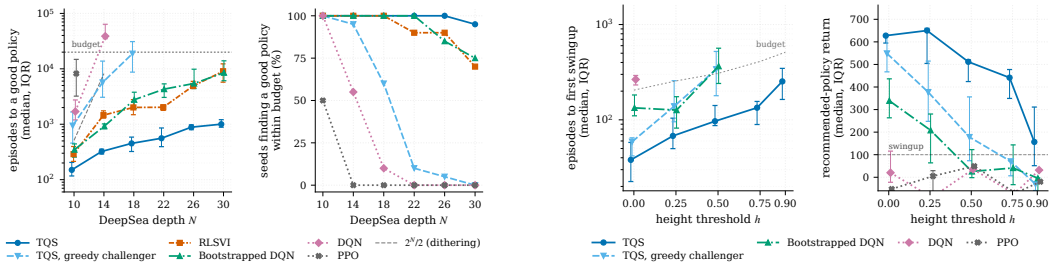

Figure 1: DeepSea instance (left), the average number of episodes to identify a good policy and simialar for Cartpole swingup (right).

Both benchmarks order the agents the same way. TQS identifies a good policy across the whole sweep, the posterior-sampling baselines follow but lose seeds as the task hardens. The ablation isolates the mechanism: replacing the randomised challenger by the greedy alternative, with the leader, the reward and the explorer untouched, degrades exploration sharply, the disagreement signal vanishes wherever the leader is confident, and the explorer reverts to the leader's own behaviour.

6 Discussion and Perspectives

We introduced TQS , a framework for best-policy identification that connects policy certification, Gibbs sampling, and exploration in Q-function space. A confidence sequence provides a stopping certificate, while its underlying score defines a distribution over alternative Q-functions whose disagreement guides data collection. Our analysis establishes guarantees for linear MDPs. The broader construction makes self-normalized confidence sequences for richer function classes a central research objective: they can support both reliable policy recommendations and the exploration needed to obtain them. TQS combines tools from offline, average-reward, and entropy-regularized RL around the task of distinguishing competing policy recommendations. Our experiments on DeepSea and Cartpole swingup illustrate the practical value of this combination. They suggest that pure exploration can provide a useful organizing principle for deep RL algorithm design, connecting statistical uncertainty to an explicit data-collection objective.

Further work should investigate how this approach scales to more demanding benchmarks, including Atari and MuJoCo. Since our experiments involved limited hyperparameter tuning, systematic sensitivity studies are also needed to assess robustness and practical performance. On the theoretical side, deriving tractable confidence sequences for nonlinear value-function classes and understanding how their geometry translates into efficient exploration remain the main challenges.

AI use statement

In this work, we used generative AI tools to polish the manuscript and improve the clarity and readability of our writing. We also used these tools to assist in identifying potentially relevant related

work. We did not use generative AI tools to generate new algorithmic ideas. All AI-assisted material was reviewed by the authors, who retain responsibility for the final content of this work, including its text, references, technical claims, and any artifacts produced with the assistance of generative AI.

Reproducibility statement

The theoretical assumptions are stated in Sections 2 and 4, with proofs and auxiliary results in Appendices A to C. Algorithm 1 and Appendix D.1 describe TQS and its practical implementation.

Code and seeded runs will be made publicly available upon publication as one self-contained notebook per environment, including all agents, experiment configurations, and plotting results. These notebooks will support both rerunning the experiments and regenerating the reported figures from saved results using standard `python` librairies.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 2312–2320, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. POLITEX: Regret bounds for policy iteration using expert prediction. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3692–3702. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lazic19a.html.

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

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

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations (ICLR)*, 2019.

Richard Y. Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. UCB exploration via Q-ensembles, 2017. URL https://arxiv.org/abs/1706.01502.

Rémy Degenne, Pierre Menard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2432–2442. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/degenne20a.html.

Eyal Even-Dar, Sham M. Kakade, and Y. Mansour. Online markov decision processes. *Math. Oper. Res.*, 34:726–736, 2009. URL https://api.semanticscholar.org/CorpusID:17185093.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

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

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. volume 49, pp. 209–232, 2002.

Cyrille Kone and Kevin Jamieson. Optimal posterior sampling for policy identification in tabular markov decision processes. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=DmBeNcPi7n.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *ArXiv*, abs/2110.06169, 2021. URL https://api.semanticscholar.org/CorpusID:238634325.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Xiang Li, Viraj Mehta, Johannes Kirschner, Ian Char, Willie Neiswanger, Jeff Schneider, Andreas Krause, and Ilija Bogunovic. Near-optimal policy identification in active reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3OR2tbtnYC-.

Zhaoqi Li, Kevin Jamieson, and Lalit Jain. Optimal exploration is no harder than Thompson sampling. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533, 2015. doi: 10.1038/NATURE14236.

Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Byx83s09Km.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling, 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4033–4041. Curran Associates Inc., 2016. ISBN 9781510838819.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a7b238ba0f6502e5d6be14424b20ded-Abstract.html.

Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. URL http://jmlr.org/papers/v20/18-339.html.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvári, Satinder Singh, Benjamin Van Roy, Richard Sutton, David Silver, and Hado van Hasselt. Behaviour suite for reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygf-kSYwH.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Martin L. Puterman. Markov decision processes: Discrete stochastic dynamic programming. In *Wiley Series in Probability and Statistics*, 1994. URL https://api.semanticscholar.org/CorpusID:260799408.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *29th Annual Conference on Learning Theory*, Proceedings of Machine Learning Research. PMLR, 2016.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Lukas Schäfer, Filippos Christianos, Josiah P. Hanna, and Stefano V. Albrecht. Decoupled reinforcement learning to stabilise intrinsically-motivated exploration. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1146–1154. International Foundation for Autonomous Agents and Multiagent Systems, 2022. URL https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p1146.pdf.

John D. Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL https://api.semanticscholar.org/CorpusID:28695052.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8583–8592. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sekar20a.html.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Jérôme Taupin, Yassir Jedra, and Alexandre Proutiere. Best policy identification in discounted linear MDPs. In *Sixteenth European Workshop on Reinforcement Learning*, 2023. URL https://openreview.net/forum?id=SOCOgATRQiY.

Andrew Wagenmaker and Kevin G Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*. PMLR, 2022.

A Bellman residuals, calibration, and estimation

The leader proposes a policy, while a separate empirical Bellman moment certifies its recommendation. In particular, the score is not centered at the leader, and its calibration does not rely on the leader's fitted targets. We first establish statistical control, then give sufficient conditions for coverage and eventual correctness of the recommendation.

A.1 Notation and probability conventions

Setting. We work with a time-homogeneous discounted MDP on $\mathcal{S} \times \mathcal{A}$, with discount $\gamma \in [0, 1)$, mean reward r, and transition kernel P. Write $PV(s, a) = \int V(s')P(ds' \mid s, a)$ and assume $|R_t| \leq R_{\max}$ almost surely. The MDP is linear in te fixed feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, let

$$Q_\theta(s, a) = \phi(s, a)^\top \theta, \qquad V_\theta(s) = \max_{a \in \mathcal{A}} Q_\theta(s, a).$$

We assume $\sup_{s,a} \|\phi(s, a)\|_2 \leq L_\phi$. The candidate set $\Theta \subset \mathbb{R}^d$ is compact, convex, and full-dimensional, with $\sup_{\theta \in \Theta} \|\theta\|_2 \leq B$. We assume $Q^\star = Q_{\theta^\star}$ for some $\theta^\star \in \Theta$. The prior Π_0 has a continuous, strictly positive density on Θ with respect to the restriction of Lebesgue measure. These geometric and prior conditions are used for local-mass bounds; the confidence calibration in Appendix A.2 needs only boundedness and realizability. The coverage and occupation-measure arguments below assume finite state and action spaces. Bellman closure and the explorer's regret guarantee are stated at their points of use.

For positive semidefinite M, write $\|v\|_M^2 = v^\top M v$; the supremum norm $\|f\|_\infty$ is taken over the domain of f. We use Greedy(Q) for the deterministic stationary policies that maximize $Q(s, \cdot)$ at every state, with a fixed measurable tie-breaking rule whenever a single policy is required. The leader $\hat{Q}_t$ is a separate estimator; neither its construction nor its accuracy is assumed in calibrating the Bellman residuals.

Interaction and filtration. Rounds start at $t = 1$. The sigma-field $\mathcal{F}_t$ contains S_t, all observations and algorithmic randomization from rounds $i < t$, and the leader and explorer state available before selecting A_t. The explorer draws A_t first. Independently of that draw conditional on $\mathcal{F}_t$, it then draws a challenger $\tilde{\theta}_t$ using fresh randomness from the $\mathcal{F}_t$-measurable distribution $\widehat{q}_t$. Set

$$\mathcal{G}_t = \mathcal{F}_t \vee \sigma(A_t), \qquad \mathcal{H}_t = \mathcal{G}_t \vee \sigma(\tilde{\theta}_t).$$

Thus $\mathcal{L}(\tilde{\theta}_t \mid \mathcal{G}_t) = \widehat{q}_t$. The environment then produces (R_t, S_{t+1}) according to its kernel at (S_t, A_t), conditionally on $\mathcal{H}_t$; the resulting history belongs to $\mathcal{F}_{t+1}$. In particular, A_t cannot depend on the current challenger, although the latter may inform subsequent explorer updates. Fits and sampling laws used in round t depend only on transitions from rounds $i < t$. This order makes $x_t = \phi(S_t, A_t)$ predictable relative to the environment observation and makes challenger sampling errors conditionally centered given $\mathcal{G}_t$. Set $x_i = \phi(S_i, A_i)$, which is $\mathcal{G}_i$-measurable. Define

$$G_t = \sum_{i=1}^{t-1} x_i x_i^\top, \qquad \Lambda_t = \lambda I_d + G_t.$$

$$Y_i(\theta) = R_i + \gamma V_\theta(S_{i+1}) - Q_\theta(S_i, A_i),$$
$$m_t(\theta) = \sum_{i=1}^{t-1} x_i Y_i(\theta).$$

$$S_t(\theta) = \tfrac{1}{2}\|m_t(\theta)\|^2_{\Lambda_t^{-1}}, \qquad \hat{b}_t(\theta) = \Lambda_t^{-1} m_t(\theta).$$

For each θ, $\hat{b}_t(\theta)$ is the ridge-regression fit of $Y_i(\theta)$ on x_i, with penalty $\lambda\|b\|_2^2$. The same candidate is evaluated on *every* stored transition, even if it was chosen using the current history. These quantities are consequently measurable with respect to $\mathcal{F}_t$ as functions of θ; the next state enters the response, never the design vector. By convention, we set $G_1 = 0$ and $S_1(\theta) = 0$.

Probability and asymptotic notation. All logarithms are natural. Unless stated otherwise, constants may depend on the fixed instance, the candidate class and prior, regularization, and algorithmic schedules, but not on t, T, or a displayed failure level. We reserve δ for stopping confidence and use ξ or ζ for auxiliary concentration bounds.

A family of events $(\mathcal{E}_T)_{T\geq 2}$ holds *with polynomially high probability* when, for deterministic $C, T_0 < \infty$,

$$\mathbb{P}(\mathcal{E}_T^c) \leq CT^{-2} \qquad \text{for every integer } T \geq T_0.$$

For deterministic $a_T \geq 0$, we write $X_T \lesssim_{\mathrm{hp}} a_T$ if $X_T \leq Ka_T$ with polynomially high probability for a fixed $K < \infty$. The notation $X_T = o_{\mathrm{hp}}(T)$ means that a deterministic $r(T) = o(T)$ satisfies $|X_T| \leq r(T)$ with polynomially high probability. In inequalities, $X_T \leq Y_T + o_{\mathrm{hp}}(T)$ means $X_T \leq Y_T + r(T)$ on such an event. Finite intersections of these events preserve the rate, with no independence requirement. Summability and Borel-Cantelli give the corresponding eventual almost-sure bounds; in particular $X_T = o_{\mathrm{hp}}(T)$ implies $X_T/T \to 0$ almost surely.

We use the following elementary concentration bound repeatedly. If (ξ_t) is a martingale-difference sequence for a filtration $(\mathcal{J}_t)$ and $|\xi_t| \leq u_t$ for deterministic u_t, then Azuma-Hoeffding yields

$$\left|\sum_{t=1}^{T} \xi_t\right| \lesssim_{\mathrm{hp}} \sqrt{\log(T+1)\sum_{t=1}^{T} u_t^2}.$$

Each application identifies its conditioning sigma-field. If boundedness holds only on a high-probability event, we apply concentration to predictably truncated increments under the original law, then intersect the resulting events. For a fixed $\nu \in (0,1)$, set

$$b_T = \lceil T^\nu \rceil, \qquad I_T = \{b_T, \ldots, T\}.$$

Bounds described as uniform on I_T hold simultaneously at every time in that interval with the stated failure probability; summing pointwise $O(t^{-2})$ errors does not give an $O(T^{-2})$ guarantee for this intersection. Score increments on the late interval are indexed by $b_T \leq t < T$, and the data-collection procedure is not restarted at b_T. An *anytime* assertion instead holds for all $t \geq 1$ on a single event of probability at least $1-\delta$. To analyze stopping, we extend the collection rule beyond a stopping decision; the observed prefix has the same law of the implemented procedure.

A.2 Calibration and fitted Bellman residuals

Proposition 6 (Anytime Bellman-moment calibration)**.** *Under the preceding boundedness and realizability assumptions, put $\sigma_\star = R_{\max}/(1-\gamma)$. For every $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $t \geq 1$,*

$$S_t(\theta^\star) \leq \beta_t(\delta) := \log\frac{1}{\delta} + \frac{1}{2}\log\frac{\det\Lambda_t}{\lambda^d} \leq \bar{\beta}_t(\delta),$$

where

$$\bar{\beta}_t(\delta) = \sigma_\star^2\left[\log\frac{1}{\delta} + \frac{d}{2}\log\left(1 + \frac{(t-1)L_\phi^2}{d\lambda}\right)\right].$$

No assumption on the leader, coverage, or linear transition dynamics is needed for this result.

Proof. The Bellman equation gives $\mathbb{E}[Y_i(\theta^\star) \mid \mathcal{G}_i] = 0$. Conditionally on $\mathcal{G}_i$, the target $R_i + \gamma V^\star(S_{i+1})$ lies in an interval of length at most $2R_{\max}/(1-\gamma)$. Hoeffding's lemma makes its centered version $\sigma_\star$-sub-Gaussian. Applying the Gaussian-mixture self-normalized martingale bound gives (6). For completeness, mixing

$$\exp\left\{v^\top m_t(\theta^\star)/\sigma_\star - \tfrac{1}{2}v^\top G_t v\right\}$$

against $v \sim \mathcal{N}(0, \lambda^{-1}I_d)$ produces the nonnegative supermartingale

$$\frac{\lambda^{d/2}}{\sqrt{\det\Lambda_t}}\exp\{S_t(\theta^\star)/\sigma_\star^2\}.$$

Ville's inequality proves simultaneous coverage; the determinant bound follows from the arithmetic-geometric mean inequality for the eigenvalues. This is the standard self-normalization argument of Abbasi-Yadkori et al. (2011). □

For identification and reward estimation, we additionally impose Bellman closure. A sufficient model is

$$r(s,a) = \phi(s,a)^\top \rho, \qquad P(ds' \mid s,a) = \phi(s,a)^\top \mu(ds'),$$

where the coefficient measures are finite signed measures and the resulting transitions are probability measures. Define

$$b(\theta) = \rho + \gamma \int V_\theta(s')\, \mu(ds') - \theta, \qquad e_\theta(s,a) = \phi(s,a)^\top b(\theta).$$

Then $e_\theta = \mathcal{T} Q_\theta - Q_\theta$. The map b is generally nonlinear in θ, but it is bounded and Lipschitz on Θ. Set $B_b = \sup_\Theta \|b(\theta)\|_2 < \infty$ and

$$E = R_{\max} + (1+\gamma) L_\phi B, \qquad \sigma_\Theta = R_{\max} + \gamma L_\phi B.$$

Both $|Y_i(\theta)|$ and $|e_\theta(s,a)|$ are at most E. The proofs only require (A.2) and its boundedness, rather than the full representation of P.

Proposition 7 (Uniform moment and residual estimation)**.** *Let*

$$N_t(\theta) = m_t(\theta) - G_t b(\theta).$$

For $\xi \in (0,1)$ and $n = t-1 \geq 1$, define the deterministic radius

$$u_t(\xi) = \sigma_\Theta \sqrt{2\left[\log \frac{\pi^2 n^2}{6\xi} + d\log(1+2n) + \frac{d}{2}\log\left(1 + \frac{nL_\phi^2}{d\lambda}\right)\right]} + \frac{2\gamma L_\phi B}{\sqrt{n}}.$$

With probability at least $1-\xi$, simultaneously over $t \geq 2$, $\theta \in \Theta$, and (s,a),

$$\|N_t(\theta)\|_{\Lambda_t^{-1}} \leq u_t(\xi),$$
$$|\phi(s,a)^\top \hat{b}_t(\theta) - e_\theta(s,a)| \leq \|\phi(s,a)\|_{\Lambda_t^{-1}} \left[u_t(\xi) + \sqrt{\lambda} B_b\right].$$

Proof. For fixed θ, write $\zeta_i(\theta) = Y_i(\theta) - e_\theta(S_i, A_i)$. These are conditionally centered and σ_Θ-sub-Gaussian, and $N_t(\theta) = \sum_{i<t} x_i \zeta_i(\theta)$. For each n, take a deterministic B/n-net of Θ, with at most $(1+2n)^d$ points. Apply self-normalization at each net point, allocating total failure probability $6\xi/(\pi^2 n^2)$ to that n, and take a union bound. The centered noise is $2\gamma L_\phi$-Lipschitz in θ. Moreover, for every scalar vector $(c_i)_{i<t}$,

$$\left\| \sum_{i<t} x_i c_i \right\|_{\Lambda_t^{-1}} \leq \left(\sum_{i<t} c_i^2 \right)^{1/2},$$

since the operator norm of $\Lambda_t^{-1/2} X_t^\top$ is at most one, where X_t has rows $x_i^\top$. Extension from the net costs at most $2\gamma L_\phi B/\sqrt{n}$. This proves (7) without substituting a data-dependent candidate into a pointwise concentration inequality. Finally,

$$\hat{b}_t(\theta) - b(\theta) = \Lambda_t^{-1}[N_t(\theta) - \lambda b(\theta)],$$

and Cauchy-Schwarz, with $\lambda \|b(\theta)\|_{\Lambda_t^{-1}} \leq \sqrt{\lambda} B_b$, proves (7). □

Define the oracle and fitted exploration losses by

$$\ell_\theta(s,a) = \tfrac{1}{2} e_\theta(s,a)^2, \qquad \hat{\ell}_{t,\theta}(s,a) = \tfrac{1}{2} \operatorname{clip}_{[-E,E]}\left(\phi(s,a)^\top \hat{b}_t(\theta)\right)^2.$$

Clipping is applied to the fitted residual, not to the testing score. Both losses lie in $[0, L]$, where $L = E^2/2$. If the residual-estimation error is uniformly at most a_t, then

$$\sup_{\theta,s,a} |\hat{\ell}_{t,\theta}(s,a) - \ell_\theta(s,a)| \leq E a_t.$$

Indeed, projection onto $[-E,E]$ cannot increase the error, and $z \mapsto z^2/2$ is E-Lipschitz on that interval. The implemented reward is $\hat{r}_t^{\mathrm{ex}} = \hat{\ell}_{t,\tilde{\theta}_t}$. Squaring an individual $Y_i(\theta)$ would instead introduce its candidate-dependent conditional variance.

A.3 Leader estimation and policy correctness

Uniform estimation of $Q^\star$ fixes the policy recommendation once the error is sufficiently small relative to the optimal action gap. We combine the self-normalized concentration of Appendix A.2 with the coverage guarantee of Appendix A.4 to identify an interval on which the leader is optimal and its alternative set is constant with high probability. We recall that $Q_\theta(s,a) = \phi(s,a)^\top\theta$, $Q_{\theta^\star} = Q^\star$, and $\Lambda_t = \lambda I_d + G_t$ from Appendix A.1. For the score-based construction below, let $\hat{Q}_t = Q_{\hat{\theta}_t}$, where $\hat{\theta}_t \in \Theta$ is $\mathcal{F}_t$-measurable, and let $\hat{\pi}_t \in \mathrm{Greedy}(\hat{Q}_t)$. Let $a^\star(s)$ denote the unique optimal action at state s. We assume a *uniform separation* condition

$$\Delta_{\min} := \inf_{s\in\mathcal{S}} \min_{a\neq a^\star(s)} \big[Q^\star(s,a^\star(s)) - Q^\star(s,a)\big] > 0.$$

For finite state and action spaces, this condition follows from uniqueness of the optimal policy, which means uniqueness of optimal action at every state. For an infinite state space, it is an additional uniform separation requirement. We recall I_T from Appendix A.1 and set $\zeta_T = (12T^2)^{-1}$. For all sufficiently large T, suppose the estimation and coverage results supply events $\mathcal{E}_T^{\mathrm{est}}$ and $\mathcal{E}_T^{\mathrm{cov}}$ with

$$\mathbb{P}\big((\mathcal{E}_T^{\mathrm{est}})^c\big) \le \zeta_T/2, \qquad \mathbb{P}\big((\mathcal{E}_T^{\mathrm{cov}})^c\big) \le \zeta_T/2.$$

On the estimation event, deterministic radii $c_{t,T}, v_{t,T} \ge 0$ satisfy

$$S_t(\hat{\theta}_t) \le c_{t,T}, \qquad \sup_{\theta\in\Theta} \|N_t(\theta)\|_{\Lambda_t^{-1}} \le v_{t,T}, \qquad t \in I_T,$$

where $N_t(\theta) = m_t(\theta) - G_t b(\theta)$ and Bellman closure is assumed as in Appendix A.2. On the coverage event, for deterministic constants $c_{\mathrm{cov}} > 0$ and $k \in (0,1)$,

$$g_t := \lambda_{\min}(G_t) \ge c_{\mathrm{cov}} t^{1-k}, \qquad t \in I_T.$$

Here k is the exponent of the forced-exploration schedule. The conditions used to establish (A.3) are confined to Appendix A.4; the argument below uses only the displayed guarantee. Both inputs concern the actual adaptive trajectory and must hold without assuming that the leader is already optimal.

One sufficient construction is a measurable approximate minimum-score estimator satisfying

$$S_t(\hat{\theta}_t) \le \inf_{\theta\in\Theta} S_t(\theta) + \epsilon_t^{\mathrm{opt}},$$

where $\epsilon_t^{\mathrm{opt}} \ge 0$ is a deterministic optimization tolerance. Continuity and compactness ensure that an exact minimizer exists. Propositions 6 and 7, each applied with failure level $\zeta_T/4$, then give (A.3) with

$$c_{t,T} = \bar{\beta}_t(\zeta_T/4) + \epsilon_t^{\mathrm{opt}}, \qquad v_{t,T} = u_t(\zeta_T/4).$$

Indeed, the minimum score is at most $S_t(\theta^\star)$. This construction requires neither an exact root of the empirical Bellman residuals nor convergence of a fitted-Bellman iteration; computing a certified approximate minimizer remains an optimization requirement.

The second index records confidence calibration at horizon T. The preceding bounds are already simultaneous in time. If only pointwise bounds are available, their failure probabilities must instead be allocated across I_T before applying the argument. Define the deterministic uniform-error envelope

$$e_T := \frac{L_\phi}{1-\gamma} \sup_{t\in I_T} \frac{\sqrt{\lambda + c_{\mathrm{cov}} t^{1-k}}}{c_{\mathrm{cov}} t^{1-k}} \big[\sqrt{2c_{t,T}} + v_{t,T}\big],$$

and choose the schedules and optimization tolerances so that $e_T \to 0$. For an exact minimum-score leader and $b_T = \lceil T^\nu \rceil$, $\nu \in (0,1)$, the radii above give $e_T = O(\sqrt{\log T / b_T^{1-k}})$. For an approximate minimizer, it suffices additionally that $\sup_{t\in I_T} \epsilon_t^{\mathrm{opt}} / t^{1-k} \to 0$.

Proposition 8 (Uniform estimation and policy stabilization)**.** *Under* (A.3)-(A.3)*, with* $e_T \to 0$*, there exists a deterministic* $T_0 < \infty$ *such that, for every* $T \ge T_0$*, the event*

$$\mathcal{A}_T := \mathcal{E}_T^{\mathrm{est}} \cap \mathcal{E}_T^{\mathrm{cov}}$$

has probability at least $1 - \zeta_T$ *and satisfies*

$$\sup_{t\in I_T} \|\hat{Q}_t - Q^\star\|_\infty \le e_T, \qquad \hat{\pi}_t = \pi^\star \quad \textit{for every } t \in I_T.$$

Proof. The probability bound follows from (A.3) by a union bound. Work pathwise on $\mathcal{A}_T$. Put $D_t = G_t \Lambda_t^{-1} G_t$. For every $t \in I_T$ and $\theta \in \Theta$, the moment decomposition and the triangle inequality give

$$\|b(\theta)\|_{D_t} = \|G_t b(\theta)\|_{\Lambda_t^{-1}} \le \|m_t(\theta)\|_{\Lambda_t^{-1}} + \|N_t(\theta)\|_{\Lambda_t^{-1}} \le \sqrt{2S_t(\theta)} + v_{t,T}.$$

Since $\Lambda_t = \lambda I_d + G_t$, the eigenvalues of D_t are $g^2/(\lambda + g)$ as g ranges over the eigenvalues of G_t. Consequently, $\lambda_{\min}(D_t) = g_t^2/(\lambda + g_t)$, and

$$\|e_\theta\|_\infty \le L_\phi \|b(\theta)\|_2 \le L_\phi \frac{\sqrt{\lambda + g_t}}{g_t} \left[\sqrt{2S_t(\theta)} + v_{t,T}\right].$$

Here $e_\theta = \mathcal{T} Q_\theta - Q_\theta$ is the conditional Bellman residual. The contraction property of the Bellman optimality operator yields

$$\|Q_\theta - Q^\star\|_\infty \le \|Q_\theta - \mathcal{T} Q_\theta\|_\infty + \|\mathcal{T} Q_\theta - \mathcal{T} Q^\star\|_\infty \le \|e_\theta\|_\infty + \gamma \|Q_\theta - Q^\star\|_\infty.$$

Thus, simultaneously for all candidates,

$$\|Q_\theta - Q^\star\|_\infty \le \frac{L_\phi}{1-\gamma} \frac{\sqrt{\lambda + g_t}}{g_t} \left[\sqrt{2S_t(\theta)} + v_{t,T}\right].$$

In particular this holds at the data-dependent leader $\hat{\theta}_t$, because the moment bound is uniform in θ. The function $g \mapsto \sqrt{\lambda + g}/g$ is decreasing on $(0, \infty)$. Using (A.3) and (A.3) therefore gives $\|\hat{Q}_t - Q^\star\|_\infty \le e_T$ for every $t \in I_T$.

Since $e_T \to 0$, choose T_0 so that the input guarantees hold and $e_T \le \Delta_{\min}/4$ for all $T \ge T_0$. For any such T, every $t \in I_T$, state s, and action $a \ne a^\star(s)$ satisfy

$$\hat{Q}_t(s, a^\star(s)) - \hat{Q}_t(s, a) \ge \Delta_{\min} - 2\|\hat{Q}_t - Q^\star\|_\infty \ge \Delta_{\min}/2 > 0.$$

Thus $a^\star(s)$ is the unique maximizing action of $\hat{Q}_t$ at every state, and $\hat{\pi}_t = \pi^\star$ throughout I_T. □

On $\mathcal{A}_T$, the alternative set therefore satisfies

$$\mathrm{Alt}_0(\hat{\pi}_t) = \mathrm{Alt}_0(\pi^\star) =: \mathcal{B}, \qquad t \in I_T.$$

Whenever the conditional challenger law is invoked on this interval, it consequently agrees with the Gibbs law restricted to $\mathcal{B}$. This identity permits the fixed-domain analysis in Appendix B.2. All probability statements retain the original law of the algorithm; no conditioning on $\mathcal{A}_T$ is needed.

The preceding proposition supplies the quantitative leader guarantee

$$\mathbb{P}\left(\sup_{t \in I_T} \|\hat{Q}_t - Q^\star\|_\infty > e_T\right) \le \zeta_T$$

for all sufficiently large T. The later analysis uses this guarantee, rather than the particular construction of the leader. A separate history-measurable estimator may therefore be used whenever it satisfies (A.3) under the actual sampling scheme. It need not minimize the Bellman residuals or belong to its confidence set. The action-gap argument and almost-sure corollary then apply unchanged.

For the score-based construction, the same event $\mathcal{A}_T$ also controls the fitted residuals. Specifically, (7) and its proof give

$$\sup_{\theta,s,a} |\phi(s,a)^\top \hat{b}_t(\theta) - e_\theta(s,a)| \le \frac{L_\phi \left[v_{t,T} + \sqrt{\lambda} B_b\right]}{\sqrt{\lambda + c_{\mathrm{cov}} t^{1-k}}}, \qquad t \in I_T.$$

With $v_{t,T} = u_t(\zeta_T/4)$, these bounds have the same orders as the envelopes in (10)-(10). They supply the residual-estimation errors used in the challenger and explorer analyses jointly with policy correctness, at total failure probability at most ζ_T. The Bellman residuals is evaluated directly from empirical residuals, so this joint control requires no bound on successive leader updates.

A.4 Coverage and regret of the explorer

This subsection establishes the two properties required of the explorer: feature coverage for estimating the empirical quantities and sublinear regret for the online exploration reward. We implement forced exploration through action randomization

$$\mu_t(\cdot \mid S_t) = (1-\epsilon_t)\mu_t^{\text{base}}(\cdot \mid S_t) + \epsilon_t \,\text{Unif}(\mathcal{A}), \qquad \epsilon_t = t^{-k}, \quad 0 < k < 1,$$

and μ_t^{base} will be later provided by the online learner. Equivalently, at round t, a fresh Bernoulli variable B_t with mean ε_t determines whether the action is drawn uniformly or adaptively from the explorer's subroutine. The mixture acts on action selection, while the subroutine optimizes the exploration reward. We first establish coverage uniformly over mixing with any adaptive exploration, and then specialize the exploration in the online-learning analysis. We note hat exploration operates mainly throughout learning and adaptivity, with only a sublinear expected number $\sum_{t\le T}\epsilon_t = O(T^{1-k})$ of forced exploration actions.

A.4.1 Coverage under forced exploration.

We prove that in a fast mixing linear MDP with finite actions, randomization over the actions is sufficient to increase the eigenvalues of the design matrix, under any adaptive sampling rule, even when $\mathcal{S}$ is continuous. Such result appears in Al Marjani et al. (2021) for tabular MDPs.

Let $\mathcal{S}$ be Borel space and $\mathcal{A}$ a finite action space. Assume that $\phi(\cdot, a)$ is continuous for every action, $\sup_{s,a}\|\phi(s,a)\|_2 \le L_\phi$, and

$$P(\mathrm{d}s' \mid s, a) = \sum_{\ell=1}^{d} \phi_\ell(s,a) M_\ell(\mathrm{d}s'),$$

where the M_ℓ are finite signed measures. Every stationary randomized policy π is assumed irreducible with respect to a common probability measure ψ and admits a stationary law ν_π. Thus every measurable set of positive ψ-measure is reachable with positive probability from every initial state under π. For constants $C_{\text{mix}} < \infty$ and $\rho \in (0,1)$, assume

$$\sup_{\pi}\sup_{s\in\mathcal{S}} \|P_\pi^n(s,\cdot) - \nu_\pi\|_{\text{TV}} \le C_{\text{mix}}\rho^n, \qquad n \ge 0.$$

Finally, we require

$$\text{span}\{\phi(s,a) : s \in \text{supp}(\psi),\ a \in \mathcal{A}\} = \mathbb{R}^d.$$

When ψ has full support, it reduces to ordinary spanning over $\mathcal{S}\times\mathcal{A}$.

Proposition 9 (Forced coverage with continuous features)**.** *Under (A.4), there exist instance constants $c_0, c_1, c_2 > 0$ such that, for every adaptive base explorer and all sufficiently large t,*

$$\mathbb{P}\big(\lambda_{\min}(G_t) < c_0 t^{1-k}\big) \le c_1 e^{-c_2 t^{1-k}}.$$

Consequently, for $I_T = \{b_T, \dots, T\}$ with $b_T \to \infty$,

$$\mathbb{P}\left(\exists t \in I_T : \lambda_{\min}(G_t) < c_0 t^{1-k}\right) \le c_1 T e^{-c_2 b_T^{1-k}}$$

for all sufficiently large T.

Proof. We first identify informative neighborhoods, then control their visitation under adaptive actions, and finally count the visits followed by forced actions.

Choose d state-action pairs (s_j, a_j), with $s_j \in \text{supp}(\psi)$, whose features are linearly independent. By continuity, there are open neighborhoods U_j of s_j and a constant $\lambda_\star > 0$ such that

$$\sum_{j=1}^{d} \phi(x_j,a_j)\phi(x_j,a_j)^\top \succeq \lambda_\star I_d, \qquad x_j \in U_j.$$

The neighborhoods can be chosen so that the state-action regions $U_j \times \{a_j\}$ are pairwise disjoint. Since $s_j \in \text{supp}(\psi)$, each U_j has positive ψ-measure.

We claim that

$$p := \min_j \inf_\pi \nu_\pi(U_j) > 0.$$

To establish uniformity over policies, consider any sequence (π_n) and define its stationary state-action laws $q_n(\mathrm{d}s, a) = \nu_{\pi_n}(\mathrm{d}s)\pi_n(a \mid s)$. Writing $z_n = \int \phi \, \mathrm{d}q_n$, stationarity gives

$$\nu_{\pi_n} = \sum_{\ell=1}^{d} z_{n,\ell} M_\ell, \qquad \|z_n\|_2 \le L_\phi.$$

Along a subsequence, $z_n \to z$, and hence $\nu_{\pi_n} \to \nu := \sum_\ell z_\ell M_\ell$ and the convergence happens in total variation. The limiting dsitribution ν is a probability measure.

The state marginals therefore form a tight sequence. Since $\mathcal{A}$ is finite, (q_n) is tight as well. By Prokhorov's theorem, a further subsequence converges weakly to a probability measure q with state marginal ν. Write $q(\mathrm{d}s, a) = \nu(\mathrm{d}s)\bar{\pi}(a \mid s)$. Boundedness and continuity of ϕ imply $z = \int \phi \, \mathrm{d}q$, so

$$\nu P_{\bar{\pi}} = \sum_{\ell=1}^{d} \left(\int \phi_\ell \, \mathrm{d}q \right) M_\ell = \nu.$$

Thus $\nu = \nu_{\bar{\pi}}$, proving that the family of stationary laws is compact in total variation.

Irreducibility gives $\nu_\pi(U_j) > 0$ for every π and j. Indeed, $\nu_\pi(U_j) = 0$ would imply $P_\pi^n(s, U_j) = 0$ for ν_π-almost every s, simultaneously for all $n \ge 1$, contradicting irreducibility. Since $\nu \mapsto \nu(U_j)$ is continuous in total variation, compactness proves (A.4.1).

Fix j and an auxiliary discount factor $\beta \in (0, 1)$. Let V_j be the bounded solution of the contracting map defined as

$$V_j(s) = \mathbf{1}_{U_j}(s) + \beta \min_{a \in \mathcal{A}} PV_j(s, a).$$

Applying (A.4.1) to that policy gives

$$(1 - \beta)V_j(s) \ge p - \frac{C_{\mathrm{mix}}(1 - \beta)}{1 - \beta\rho}.$$

Choose a fixed β sufficiently close to one that the right-hand side is at least $p/2$, simultaneously for all j. Then $h_j := V_j - \inf_s V_j(s)$ is bounded and nonnegative, and the Bellman inequality implies

$$\mathbf{1}_{U_j}(s) \ge p/2 + h_j(s) - Ph_j(s, a), \qquad (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Indeed, writing $m_j = \inf_s V_j(s)$,

$$V_j(s) - \beta PV_j(s, a) = (1 - \beta)m_j + h_j(s) - \beta Ph_j(s, a) \ge p/2 + h_j(s) - Ph_j(s, a).$$

This inequality holds for every action, so it applies along any adaptive trajectory. With

$$N_j(n) = \sum_{i=1}^{n} \mathbf{1}_{U_j}(S_i), \qquad M_{j,n} = \sum_{i=1}^{n} \big[h_j(S_{i+1}) - Ph_j(S_i, A_i) \big],$$

summing (A.4.1) yields

$$N_j(n) \ge \frac{p}{2} n + h_j(S_1) - h_j(S_{n+1}) + M_{j,n}.$$

$(M_{j,n})$ is a martingale with bounded increments. Azuma-Hoeffding and a union bound therefore give a constant $c_{\mathrm{vis}} > 0$ such that, for all sufficiently large n,

$$\mathbb{P}\left(\min_j N_j(n) < \frac{p}{4} n \right) \le d \exp(-c_{\mathrm{vis}} n).$$

Finally we can control the eigenvalues of the design matrix. Let $\mathcal{H}_i$ denote the history immediately before drawing the forced-exploration coin and the action, including S_i. Define

$$Z_{i,j} = \mathbf{1}\{S_i \in U_j,\ B_i = 1,\ A_i = a_j\}, \qquad q_{i,j} = \mathbb{E}[Z_{i,j} \mid \mathcal{H}_i] = \frac{i^{-k}}{|\mathcal{A}|} \mathbf{1}_{U_j}(S_i),$$

and let $F_j(n) = \sum_{i=1}^n Z_{i,j}$ and $Q_j(n) = \sum_{i=1}^n q_{i,j}$. On the visitation event from (A.4.1),

$$Q_j(n) \geq \frac{n^{-k}}{|\mathcal{A}|} N_j(n) \geq \alpha n^{1-k}, \qquad \alpha := \frac{p}{4|\mathcal{A}|}.$$

The conditional Bernoulli identity

$$\mathbb{E}[e^{-(\log 2)Z_{i,j}} \mid \mathcal{H}_i] = 1 - \frac{q_{i,j}}{2} \leq e^{-q_{i,j}/2}$$

implies $\mathbb{E}[e^{-(\log 2)F_j(n)+Q_j(n)/2}] \leq 1$. Thus, with $\kappa = (1 - \log 2)/2$,

$$\mathbb{P}\left(F_j(n) \leq \tfrac{1}{2}Q_j(n),\ Q_j(n) \geq v\right) \leq e^{-\kappa v}.$$

Combining this bound with (A.4.1) gives

$$\mathbb{P}\left(\min_j F_j(n) < \frac{\alpha}{2} n^{1-k}\right) \leq de^{-c_{\mathrm{vis}} n} + de^{-\kappa \alpha n^{1-k}}.$$

On the complementary event, group the counted observations into $\min_j F_j(n)$ tuples, each containing one observation from every region. The regions are disjoint, so no observation is reused. By (A.4.1),

$$G_{n+1} \geq \lambda_\star \min_j F_j(n)\, I_d \geq \frac{\lambda_\star \alpha}{2} n^{1-k} I_d.$$

Setting $n = t - 1$ and adjusting instance constants proves the fixed-time bound. A union bound over $b_T \leq t \leq T$ gives the simultaneous claim. □

A.4.2 An online explorer with sublinear regret

We now construct an explorer for the information rewards $g_t := \hat{r}_t^{\mathrm{ex}}$. The construction combines the exponential policy updates of Politex (Abbasi-Yadkori et al., 2019) with fitted policy evaluation using the accumulated transitions. Every round uses the forced-action mixture

$$\mu_t(\cdot \mid s) = (1 - t^{-k})\pi_t(\cdot \mid s) + t^{-k}\,\mathrm{Unif}(\mathcal{A}), \qquad 0 < k < 1.$$

The challenger retains its (unrealted) global schedule $\eta_t = t^{-\kappa}$, where $0 < \kappa < 1$.

Partition time into doubling epochs $\mathcal{E}_m = \{2^m, \ldots, 2^{m+1} - 1\}$. Within epoch m, use

$$N_m = 2^m, \qquad \alpha_m = N_m^{-1/2}, \qquad H_m = \left\lceil \log^2(2 + N_m) \right\rceil.$$

Only the policy weights are reset between epochs; the trajectory, replay data, residual fits, and challenger statistics are retained. Write

$$C_t = \frac{1}{2}\hat{b}_t(\tilde{\theta}_t)\hat{b}_t(\tilde{\theta}_t)^\top, \qquad g_t(s,a) = \phi(s,a)^\top C_t \phi(s,a).$$

Here C_t uses the residual fit available when the reward is issued.

The explorer's critic evaluates the average information reward. For the current pair (π_t, g_t), put $L_t = L_\phi^2 \|C_t\|_{\mathrm{op}}$ and $V_0 = 0$. Using all transitions through round t, perform

$$\begin{aligned} w_h &= \Lambda_{t+1}^{-1} \sum_{i=1}^{t} x_i V_{h-1}(S_{i+1}), \\ \widehat{Q}_h(s,a) &= g_t(s,a) + \phi(s,a)^\top w_h, \\ V_h(s) &= \left[\sum_a \pi_t(a \mid s)\widehat{Q}_h(s,a)\right]_{[0,hL_t]}, \qquad h = 1, \ldots, H_m. \end{aligned}$$

Set $\widehat{Q}_t = \widehat{Q}_{H_m}$. The projection bounds the critic's continuation values. The quadratic immediate reward is retained exactly, while linearity of the transition kernel makes its continuation value linear in ϕ. Consequently, both the critic and the policy updates have finite representations: the policy uses a matrix U and a vector v.

Algorithm 2: Replay Politex with forced exploration

for $m = 0, 1, \ldots$ **do**
  Set α_m, H_m as above; initialize $U = 0$, $v = 0$
  for $t = 2^m, \ldots, 2^{m+1} - 1$ **do**
    Define $\pi_t(a \mid s) \propto \exp\{\phi(s,a)^\top U \phi(s,a) + \phi(s,a)^\top v\}$
    Draw $A_t \sim (1 - t^{-k})\pi_t(\cdot \mid S_t) + t^{-k}\,\mathrm{Unif}(\mathcal{A})$
    Receive $\tilde{\theta}_t \sim \widehat{q}_t$ independently of A_t conditional on $\mathcal{F}_t$; form C_t
    Observe (R_t, S_{t+1}) and append the transition to the replay data
    Compute w_{H_m} by replay evaluation below
    $U \leftarrow U + \alpha_m C_t, \quad v \leftarrow v + \alpha_m w_{H_m}$
    Update the residual fits, leader, and challenger statistics using the new transition

Proposition 10 (Regret of the concrete explorer)**.** *Assume the conditions established above for coverage and minimum-score leader stabilization, and use exact Gibbs sampling with $\eta_t = t^{-\kappa}$. For $\zeta_T = (12T^2)^{-1}$ and all sufficiently large T, Algorithm 2 satisfies, with probability at least $1 - \zeta_T$,*

$$\sup_{\mu} \sum_{t=1}^{T-1} \langle d^\mu, g_t \rangle - \sum_{t=1}^{T-1} g_t(S_t, A_t) \le \widetilde{O}\Big(T^{(1+k)/2} + T^{1-k} + T^{1-\kappa}\Big).$$

Constants may depend on the fixed instance and the schedules. In particular, $k = 1/3$ and $\kappa = 1/2$ give $\widetilde{O}(T^{2/3})$ regret.

Proof sketch. Fix T, and let b be the first doubling-epoch boundary at least $T^{1/4}$. Allocate fixed fractions of ζ_T to the statistical and martingale bounds below. The preceding results ensure, uniformly for $b \le t < T$, that

$$\lambda_{\min}(\Lambda_t) \gtrsim t^{1-k}, \qquad \hat{\pi}_t = \pi^\star, \qquad \sup_{\theta} \|\hat{b}_t(\theta)\|_2 \le B$$

for an instance constant B. In particular, the issued rewards are uniformly bounded on this interval. Before b, the ridge-regression contraction gives $\|g_t\|_\infty \le Ct$. Since $g_t \ge 0$, this prefix contributes at most $O(b^2) = O(\sqrt{T})$ to regret. The cutoff b is used only in the analysis.

Write $J(\pi, g) = \langle d^\pi, g \rangle$, and let $Q^{\pi,g}$ denote a differential action-value function. For a round in epoch m, replay evaluation satisfies

$$\begin{aligned} e_t &\coloneqq \inf_{c \in \mathbb{R}} \|\widehat{Q}_t - Q^{\pi_t, g_t} - c\|_\infty \\ &\le C H_m^2 \sqrt{\frac{d^2 \log(T/\zeta_T)}{\lambda_{\min}(\Lambda_{t+1})}} + C\rho^{H_m - 1} = \widetilde{O}(t^{-(1-k)/2}), \end{aligned}$$

where $\rho < 1$ is the mixing factor.

We justify this estimate despite replay and adaptive targets, by observing that every regression target in (10) is a clipped value function parameterized by (U, v, C, w), with $O(d^2)$ parameters. Ridge regularization and the bounded residual responses give deterministic polynomial bounds on their coefficient norms over a horizon T. Restricting C to its late-interval bound, this class has envelope $O(H_m)$ and covering entropy $O(d^2 \log(T/\varepsilon))$ in the uniform norm. The same covering and self-normalized argument used for the residual fits therefore controls the regressions uniformly over their data-dependent targets. Accumulating the backup errors gives the first term in (10); clipping is nonexpansive. Uniform mixing bounds the difference between the exact H_m-step action value and $H_m J(\pi_t, g_t) + Q^{\pi_t, g_t}$ by the second term.

For every stationary comparator μ, the performance-difference identity gives

$$J(\mu, g_t) - J(\pi_t, g_t) = \int \nu_\mu(ds)\, \langle \mu(\cdot \mid s) - \pi_t(\cdot \mid s), Q^{\pi_t, g_t}(s, \cdot) \rangle.$$

The policy update is exponential weights with payoff $\widehat{Q}_t(s, \cdot)$ at each state. Uniform mixing bounds the action span of the differential values, and (10) bounds the estimated spans on the late interval.

Applying the exponential-weights inequality pointwise in s, integrating against ν_μ, and summing the doubling epochs yields

$$\sup_\mu \sum_{t=b}^{T-1} [J(\mu, g_t) - J(\pi_t, g_t)] \le O(\sqrt{T}) + 2 \sum_{t=b}^{T-1} e_t = \widetilde{O}(T^{(1+k)/2}).$$

This comparison is pathwise and permits the adaptive rewards g_t.

For analysis, define the predictable mean reward

$$f_t(s,a) = \int \hat{\ell}_{t,\theta}(s,a) \widehat{q}_t(d\theta).$$

After leader stabilization, the Gibbs restriction is fixed. The recursive ridge update gives $\sup_\theta \|\hat{b}_{t+1}(\theta) - \hat{b}_t(\theta)\|_2 = O(t^{k-1})$. The bounded score increments and $S_t = O(t)$ also give $\|\eta_{t+1} S_{t+1} - \eta_t S_t\|_\infty = O(t^{-\kappa})$. Consequently,

$$\|f_{t+1} - f_t\|_\infty = O(t^{k-1} + t^{-\kappa}).$$

The bounded critic spans similarly imply $\sup_s \|\pi_{t+1}(\cdot \mid s) - \pi_t(\cdot \mid s)\|_{\mathrm{TV}} = O(t^{-1/2})$ within epochs; epoch resets contribute $O(\log T)$ to total policy variation.

For uniformly bounded rewards, uniform mixing gives bounded Poisson biases $h_{\pi,f}$. Normalize them at a common reference state. Subtracting their Poisson equations and summing the centered transition kernels gives

$$\|h_{\pi,f} - h_{\pi',f'}\|_\infty \le C \left(\|f - f'\|_\infty + \sup_s \|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|_{\mathrm{TV}} \right).$$

Conditional independence of the challenger and action permits replacement of g_t by f_t in the learner's stationary and realized rewards, at a martingale cost $\widetilde{O}(\sqrt{T})$. The Poisson equations then telescope along the actual trajectory. Changing biases incurs their total variation; the forced-action mixture incurs $O(\sum_t t^{-k})$. Thus

$$\sum_{t=b}^{T-1} [J(\pi_t, g_t) - g_t(S_t, A_t)] \le \widetilde{O}\left(T^{1-k} + T^k + T^{1-\kappa} + \sqrt{T} \right).$$

Which e follows by martingale bounds on the respective statistical events.

Combining the prefix, stationary comparison, and trajectory bounds, and using $T^k + \sqrt{T} = O(T^{(1+k)/2})$, proves the claim. □

Consequences for residual estimation and the leader. The simultaneous coverage failure probability is eventually smaller than ζ_T whenever $\log(T/\zeta_T) = o(b_T^{1-k})$. In particular, this holds for every polynomial confidence schedule and every fixed $\nu \in (0,1)$. The cutoff b_T is used only in the analysis; the exploration rule operates from the first round.

Allocate a fixed fraction ξ_T of the prescribed failure budget ζ_T to each of the finitely many confidence events below. On the coverage event and the event of Proposition 7 at confidence level ξ_T, put

$$a_{t,T} = \frac{L_\phi}{\sqrt{\lambda + c_0 t^{1-k}}} \left[u_t(\xi_T) + \sqrt{\lambda} B_b \right], \qquad t \in I_T.$$

For polynomial confidence schedules, $a_{t,T} = O(\sqrt{\log T / t^{1-k}})$ uniformly on I_T, and therefore

$$A_1(T) \coloneqq \sum_{t=b_T}^{T-1} a_{t,T} = O\left(T^{(1+k)/2} \sqrt{\log T} \right),$$

$$A_2(T) \coloneqq \sum_{t=b_T}^{T-1} a_{t,T}^2 = O(T^k \log T).$$

Both terms are sublinear because $0 < k < 1$. Combining the same coverage event with Propositions 6 and 8 supplies (A.3) for a minimum-score leader, with

$$e_T = O\left(\sqrt{\frac{\log T}{b_T^{1-k}}} \right).$$

Choosing the component failure budgets to sum to at most ζ_T gives the prescribed simultaneous probability guarantee.

B Information accumulation

Once the leader stabilizes, the conditional Gibbs sampler ranges over a fixed alternative set. Its regret guarantee compares accumulated Bellman residual losses with the terminal score of the best alternative Q-function. The exploration learner supplies a complementary comparison with feasible stationary allocations. We combine these two guarantees to establish growth of the test statistic. The analysis uses the score function S_t and losses ℓ_θ of Appendix A.2.

B.1 Quadratic geometry and the separation game

We work with finite state and action spaces and we recall the bounded linear representation, compact candidate domain, and prior assumptions of Appendix A.1. Assume Bellman closure as in (A.2), so that

$$e_\theta(s,a) = r(s,a) + \gamma P V_\theta(s,a) - Q_\theta(s,a) = \phi(s,a)^\top b(\theta).$$

Write $a^\star(s) = \pi^\star(s)$ for the unique optimal action and retain the positive gap $\Delta_{\min}$ from (A.3). For each terminal time T, let

$$\mathcal{E}_T^{\mathrm{stab}} = \{\hat{\pi}_t = \pi^\star \text{ for every } t \in I_T\}.$$

Appendix A.3 supplies this event with polynomially high probability. We use it to identify the alternative set along the trajectory; all probabilistic arguments remain under the original law.

Let $\mathcal{B} = \mathrm{Alt}_0(\pi^\star)$ and assume $\mathcal{B} \neq \varnothing$. The linear representation gives

$$\mathcal{B} = \bigcup_{s\in\mathcal{S}} \bigcup_{a\neq a^\star(s)} \big\{\theta \in \Theta : [\phi(s,a) - \phi(s,a^\star(s))]^\top \theta > 0\big\}.$$

Indeed, $\pi^\star$ fails to be greedy precisely when another action has a strictly larger value at some state. Ties alone do not invalidate $\pi^\star$. We write $\overline{\mathcal{B}}$ for the Euclidean closure of the actual alternative set: a tie belongs to this closure only if it can be approached by alternatives. On $\mathcal{E}_T^{\mathrm{stab}}$, $\mathrm{Alt}_0(\hat{\pi}_t) = \mathcal{B}$ for every $t \in I_T$.

The oracle loss is quadratic in the conditional residual,

$$\ell_\theta(s,a) = \tfrac{1}{2} e_\theta(s,a)^2 = \tfrac{1}{2}[\phi(s,a)^\top b(\theta)]^2.$$

Its cumulative value and the empirical score satisfy

$$\sum_{i<t} \ell_\theta(S_i, A_i) = \tfrac{1}{2} b(\theta)^\top G_t b(\theta),$$

$$S_t(\theta) = \tfrac{1}{2}\|G_t b(\theta) + N_t(\theta)\|^2_{\Lambda_t^{-1}},$$

where $N_t(\theta) = m_t(\theta) - G_t b(\theta)$. The corresponding noise-free quadratic form is

$$\tfrac{1}{2} b(\theta)^\top G_t \Lambda_t^{-1} G_t b(\theta).$$

The map b need not be linear, so these expressions need not be globally quadratic or convex as functions of θ. In particular, the empirical score is continuous on Θ. Define the algorithm's exact testing statistic by

$$\mathsf{T}(t) = \inf_{\theta \in \mathrm{Alt}_0(\hat{\pi}_t)} S_t(\theta).$$

On $\mathcal{E}_T^{\mathrm{stab}}$, for $t \in I_T$,

$$\mathsf{T}(t) = \inf_{\theta\in\mathcal{B}} S_t(\theta) = \min_{\theta\in\overline{\mathcal{B}}} S_t(\theta),$$

by continuity and compactness. The infimum over an empty alternative set is $+\infty$; such a recommendation is greedy for every candidate, including $\theta^\star$, and may be returned immediately. For the continued-process analysis, define any history-measurable challenger law on rounds with an empty alternative set. Its bounded feedback extends the process and does not affect the late-interval argument.

Let $\Omega(P)$ be the polytope of stationary state-action occupations:

$$\Omega(P) = \left\{ \omega \geq 0 : \sum_{s,a} \omega(s,a) = 1, \quad \sum_a \omega(s,a) = \sum_{x,a} \omega(x,a) P(s \mid x, a) \text{ for every } s \right\}.$$

For $\omega \in \Omega(P)$, set

$$\Sigma(\omega) = \sum_{s,a} \omega(s,a)\phi(s,a)\phi(s,a)^\top,$$
$$\mathcal{L}(\omega,\theta) = \sum_{s,a} \omega(s,a)\ell_\theta(s,a) = \tfrac{1}{2} b(\theta)^\top \Sigma(\omega) b(\theta),$$
$$\Gamma^\star = \sup_{\omega\in\Omega(P)} \inf_{\theta\in\mathcal{B}} \mathcal{L}(\omega,\theta).$$

The allocation player maximizes this payoff, while the challenger minimizes it over $\mathcal{B}$. To attain this benchmark through the exploration regret guarantee, we require the comparison class to satisfy

$$\overline{\mathrm{conv}}\{d^\mu : \mu \in \mathcal{M}\} = \Omega(P),$$

where d^μ denotes the stationary occupation of policy μ. For a probability measure q on $\mathcal{B}$, write $\mathcal{L}(\omega, q) = \int \mathcal{L}(\omega,\theta) q(d\theta)$. These integrals are finite because $0 \le \ell_\theta \le L = E^2/2$.

Lemma 11. *Under the conditions above, for any finite index set I and probability measures $(q_t)_{t\in I}$ on $\mathcal{B}$,*

$$\sup_\mu \sum_{t\in I} \mathcal{L}(d^\mu, q_t) \ge |I|\Gamma^\star.$$

Moreover, if $\Omega(P)$ contains a full-support occupation $\bar\omega$, then, with $m = \min_{s,a} \bar\omega(s,a) > 0$,

$$0 < \frac{m(1-\gamma)^2\Delta_{\min}^2}{8} \le \Gamma^\star \le L.$$

Proof. The function $\omega \mapsto \inf_{\theta\in\mathcal{B}} \mathcal{L}(\omega,\theta)$ is upper semicontinuous, being an infimum of continuous linear functions. It attains its maximum at some $\omega^\star$ on the compact polytope $\Omega(P)$. For every $\theta \in \mathcal{B}$, $\mathcal{L}(\omega^\star,\theta) \ge \Gamma^\star$. Integrating and summing gives

$$\sum_{t\in I} \mathcal{L}(\omega^\star, q_t) \ge |I|\Gamma^\star.$$

By linearity in ω and (B.1), maximizing the left-hand side over $\Omega(P)$ is equivalent to taking the supremum over d^μ. This proves (11).

For positivity, fix $\theta \in \mathcal{B}$. Some state-action pair satisfies $Q_\theta(s,a) > Q_\theta(s,a^\star(s))$, and hence

$$2\|Q_\theta - Q^\star\|_\infty > Q^\star(s,a^\star(s)) - Q^\star(s,a) \ge \Delta_{\min}.$$

The contraction property of the Bellman optimality operator yields

$$\|e_\theta\|_\infty \ge (1-\gamma)\|Q_\theta - Q^\star\|_\infty.$$

Consequently,

$$\mathcal{L}(\bar\omega,\theta) \ge \tfrac{m}{2}\|e_\theta\|_\infty^2 > \frac{m(1-\gamma)^2\Delta_{\min}^2}{8}.$$

Taking the infimum over alternatives proves the lower bound. The upper bound follows from $\ell_\theta \le L$ and nonemptiness of $\mathcal{B}$. □

The comparison (11) is deterministic and therefore also applies pathwise to random challenger laws, including point masses at sampled alternatives. The next subsections relate these oracle payoffs to fitted exploration rewards and to the terminal testing score.

B.2 Conditional Gibbs sampling

We compare the cumulative loss averaged under the challenger law with the terminal testing score. The argument follows the conditional exponential-weights analysis and techniques inspired from Kone & Jamieson (2026). The proof controls the evolution of a log-partition potential, and compare the terminal potential with the best alternative.

The nonempty alternative set $\mathcal{B}$ has positive prior mass, as also follows from Lemma 14. Define

$$\rho_0 = \Pi_0(\,\cdot \mid \mathcal{B}), \qquad Z_t(\eta) = \int_{\mathcal{B}} e^{-\eta S_t(\theta)} \rho_0(d\theta), \qquad \Phi_t(\eta) = -\frac{1}{\eta}\log Z_t(\eta).$$

Let $(\eta_t)_{t\ge 1}$ be deterministic, positive, and nonincreasing, and set

$$q_t(d\theta) = \frac{e^{-\eta_t S_t(\theta)}}{Z_t(\eta_t)}\rho_0(d\theta), \qquad \ell_t(\theta) = \ell_\theta(S_t, A_t).$$

The auxiliary laws q_t are $\mathcal{F}_t$-measurable. Whenever the algorithm samples exactly from the Gibbs law restricted to its current alternatives, q_t matches the challenger law on $\mathcal{E}_T^{\mathrm{stab}}$ for $t \in I_T$. We further introduce $\langle q, f\rangle := \int f\, dq$ and $D_\Theta := \mathrm{diam}(\Theta)$.

For each candidate, abbreviate

$$p_t(\theta) = x_t^\top \hat{b}_t(\theta), \qquad h_t = x_t^\top \Lambda_t^{-1} x_t,$$
$$e_t(\theta) = e_\theta(S_t, A_t), \qquad \zeta_t(\theta) = Y_t(\theta) - e_t(\theta), \qquad \Delta_t(\theta) = S_{t+1}(\theta) - S_t(\theta).$$

Thus p_t is the residual at the selected pair, $h_t \ge 0$ is the information accumulated in the direction x_t, and $\zeta_t(\theta)$ is centered conditionally on $\mathcal{G}_t$.

Lemma 12 (Score and potential evolution)**.** *For every $t \ge 1$ and $\theta \in \Theta$,*

$$\Delta_t(\theta) = \ell_t(\theta) - \tfrac{1}{2}[p_t(\theta) - e_t(\theta)]^2 + p_t(\theta)\zeta_t(\theta) + \frac{h_t}{2(1+h_t)}[Y_t(\theta) - p_t(\theta)]^2.$$

Let $C_\Delta = (2E+1)^2/2$ and define the correction

$$\psi_t = \tfrac{1}{2}\langle q_t, (p_t - e_t)^2\rangle - \langle q_t, p_t\zeta_t\rangle.$$

For integers $1 \le b < T$, on every history satisfying $\sup_{\theta\in\Theta}|p_t(\theta)| \le E+1$ for $b \le t < T$,

$$\sum_{t=b}^{T-1}\langle q_t, \ell_t\rangle \le \Phi_T(\eta_T) - \Phi_b(\eta_b) + \frac{C_\Delta^2}{2}\sum_{t=b}^{T-1}\eta_t + \sum_{t=b}^{T-1}\psi_t.$$

Proof. For each fixed candidate, design updates are simply $m_{t+1}(\theta) = m_t(\theta) + x_t Y_t(\theta)$ and $\Lambda_{t+1} = \Lambda_t + x_t x_t^\top$. Applying Sherman-Morrison formula therefore gives after simplification

$$\Delta_t(\theta) = \frac{2p_tY_t + h_tY_t^2 - p_t^2}{2(1+h_t)} = p_tY_t - \tfrac{1}{2}p_t^2 + \frac{h_t}{2(1+h_t)}(Y_t - p_t)^2.$$

Substituting $Y_t = e_t + \zeta_t$ in the first two terms and completing the square proves (12). This calculation evaluates the same candidate on the full dataset at both times.

On the stated histories, $|p_t| \le E+1$ and $|Y_t| \le E$. The first expression above implies

$$|\Delta_t(\theta)| \le E(E+1) + \tfrac{1}{2}E^2 + \tfrac{1}{2}(E+1)^2 = C_\Delta.$$

The partition-function identity and Hoeffding's lemma, applied under q_t after fixing the observed transition, yield

$$\log\frac{Z_{t+1}(\eta_t)}{Z_t(\eta_t)} = \log\langle q_t, e^{-\eta_t\Delta_t}\rangle \le -\eta_t\langle q_t, \Delta_t\rangle + \tfrac{1}{2}\eta_t^2 C_\Delta^2.$$

For $0 < \eta' \le \eta$, concavity of $z \mapsto z^{\eta'/\eta}$ (cf Kone & Jamieson (2026)) gives $Z_t(\eta') \le Z_t(\eta)^{\eta'/\eta}$ and hence $\Phi_t(\eta) \le \Phi_t(\eta')$. Since $\eta_{t+1} \le \eta_t$,

$$\langle q_t, \Delta_t\rangle \le \Phi_{t+1}(\eta_{t+1}) - \Phi_t(\eta_t) + \tfrac{1}{2}C_\Delta^2\eta_t.$$

Finally, the last term in (12) is nonnegative, so $\langle q_t, \ell_t\rangle \le \langle q_t, \Delta_t\rangle + \psi_t$. Summing telescopes and proves (12). □

Controlling the moving center. The Bellman residuals does not depend on the changing leader. Its update therefore requires no bound on successive leader estimates. The correction in (12) instead consists of a squared residual-estimation error and a conditionally centered fluctuation. The following lemma controls these terms without conditioning on the statistical event that bounds the fitted residuals.

Lemma 13 (Control of residual-estimation corrections)**.** *Fix deterministic integers $1 \le b < T$, numbers $a_{t,T} \in [0,1]$ for $b \le t < T$, and $\zeta \in (0,1)$. Let*

$$\mathcal{D}_{b,T} = \left\{ \sup_{\theta,s,a} |\phi(s,a)^\top \hat{b}_t(\theta) - e_\theta(s,a)| \le a_{t,T} \text{ for every } b \le t < T \right\}.$$

There is an event $C_{b,T}(\zeta)$ with failure probability at most ζ such that, on $\mathcal{D}_{b,T} \cap C_{b,T}(\zeta)$,

$$\sum_{t=b}^{T-1} \psi_t \le \tfrac{1}{2} \sum_{t=b}^{T-1} a_{t,T}^2 + D_{\mathrm{N}} \sqrt{2(T-b)\log(1/\zeta)}, \qquad D_{\mathrm{N}} = 2E(E+1).$$

Proof. We introduce quantities

$$J_t = \mathbf{1}\{\sup_{\theta\in\Theta} |p_t(\theta)| \le E+1\}, \qquad W_t = J_t \int_{\mathcal{B}} p_t(\theta)\zeta_t(\theta) q_t(d\theta).$$

The quantities J_t, p_t, and q_t are $\mathcal{G}_t$-measurable. For each fixed candidate, $\mathbb{E}[\zeta_t(\theta) \mid \mathcal{G}_t] = 0$ and $|\zeta_t(\theta)| \le 2E$. Conditional integration therefore gives

$$\mathbb{E}[W_t \mid \mathcal{G}_t] = 0, \qquad |W_t| \le D_{\mathrm{N}}.$$

Since W_t is $\mathcal{F}_{t+1}$-measurable and $\mathcal{F}_t \subseteq \mathcal{G}_t$, the tower property makes (W_t) a martingale-difference sequence for the round filtration. Azuma-Hoeffding implies

$$\mathbb{P}\left(-\sum_{t=b}^{T-1} W_t > D_{\mathrm{N}} \sqrt{2(T-b)\log(1/\zeta)} \right) \le \zeta.$$

Let $C_{b,T}(\zeta)$ be the complementary event. On $\mathcal{D}_{b,T}$, $J_t = 1$ throughout the interval and $\langle q_t, (p_t - e_t)^2 \rangle \le a_{t,T}^2$. Summing (12) proves the result. □

It remains to upper-bound the terminal potential.

Lemma 14 (Local volume and terminal potential)**.** *There exists a deterministic $c \in (0,1]$ such that, for every $\theta \in \overline{\mathcal{B}}$ and $0 < r \le 1$,*

$$\rho_0(\mathcal{B} \cap B(\theta, r)) \ge c r^d.$$

Consequently, for every $T \ge 2$ and $\eta > 0$,

$$\Phi_T(\eta) \le \inf_{\theta\in\mathcal{B}} S_T(\theta) + K_0 + \frac{d\log T + K_1}{\eta},$$

where $K_0 = E(1+\gamma)L_\phi$ and $K_1 = \log(1/c)$.

Proof. Index the nonempty sets in (B.1) by $\mathcal{B}_j = \Theta \cap \{\theta : w_j^\top \theta > 0\}$. Each contains a Euclidean ball $B(z_j, r_j)$ with $r_j > 0$: mix a strictly feasible point with an interior point of Θ, retaining strict feasibility, and then take a sufficiently small ball. In particular, $\Pi_0(\mathcal{B}) > 0$. Since there are finitely many components, $\overline{\mathcal{B}} = \bigcup_j \overline{\mathcal{B}_j}$. For $\theta \in \overline{\mathcal{B}_j}$ and $0 < r \le 1$, put $\alpha_r = r/(D_\Theta + 1)$. Convexity and the strict linear inequality on $B(z_j, r_j)$ imply

$$(1-\alpha_r)\theta + \alpha_r B(z_j, r_j) \subseteq \mathcal{B}_j \cap B(\theta, r).$$

Indeed, every point on the left satisfies the inequality strictly and lies within distance $\alpha_r D_\Theta < r$ of θ. This set is a ball of radius $\alpha_r r_j$. The positive lower bound on the prior density and $\min_j r_j > 0$ give (14), after decreasing c to at most one.

Let $n = T - 1$. The matrix inequality in the proof of Proposition 7 gives

$$\|m_T(\theta)\|_{\Lambda_T^{-1}} \le E\sqrt{n},$$
$$\|m_T(\theta) - m_T(\theta')\|_{\Lambda_T^{-1}} \le (1+\gamma)L_\phi \sqrt{n}\, \|\theta - \theta'\|_2.$$

The second inequality uses the $(1+\gamma)L_\phi$-Lipschitz property of each observed residual $Y_i(\cdot)$. Taking a difference of squared norms therefore yields

$$|S_T(\theta) - S_T(\theta')| \le K_0(T-1)\|\theta - \theta'\|_2.$$

Choose a minimizer $\theta_T \in \overline{\mathcal{B}}$ (it exists by compactness of $\overline{\mathcal{B}}$). Its score equals $\inf_{\mathcal{B}} S_T$ by continuity, and $S_T(\theta) \le S_T(\theta_T) + K_0$ for $\theta \in \Theta \cap B(\theta_T, T^{-1})$. Restricting $Z_T(\eta)$ to this ball and applying (14) gives

$$Z_T(\eta) \ge cT^{-d} e^{-\eta[S_T(\theta_T)+K_0]}.$$

Taking logarithms on both sides proves (14). □

We now combine the three lemmas. The statistical input is uniform estimation of the conditional residuals on a deterministic late interval. Policy stabilization is used only to to fix the truncation set for Gibbs laws with the algorithm's actual challenger laws.

Proposition 15 (Regret of the conditional Gibbs sampler)**.** *Under the setup of this subsection, fix $1 \le b < T$, deterministic $a_{t,T} \in [0,1]$, and $\zeta \in (0,1)$. On $\mathcal{D}_{b,T} \cap C_{b,T}(\zeta)$,*

$$\sum_{t=b}^{T-1} \langle q_t, \ell_t \rangle \le \inf_{\theta \in \mathcal{B}} S_T(\theta) + r_{\mathcal{G}}(T, b, \zeta),$$

where

$$r_{\mathcal{G}}(T,b,\zeta) = K_0 + \frac{d\log T + K_1}{\eta_T} + \frac{C_\Delta^2}{2}\sum_{t=b}^{T-1}\eta_t + \tfrac{1}{2}\sum_{t=b}^{T-1} a_{t,T}^2 + D_{\mathrm{N}}\sqrt{2(T-b)\log(1/\zeta)}.$$

Moreover, for $b = b_T$, the additional event $\mathcal{E}_T^{\mathrm{stab}}$ identifies $\inf_{\theta\in\mathcal{B}} S_T(\theta)$ with $\mathsf{T}(T)$ and q_t with the actual challenger laws on the the associated interval.

Proof. On $\mathcal{D}_{b,T}$, $\sup_\theta |p_t(\theta)| \le E+1$. Apply Lemma 12. Since ρ_0 is a probability measure and $S_b \ge 0$, $Z_b(\eta_b) \le 1$ and $\Phi_b(\eta_b) \ge 0$, so the initial potential can be discarded. Lemmas 13 and 14 then give the displayed bound. The final identification follows from exact conditional Gibbs sampling and (A.3). □

For $b = b_T$ and $\zeta = \zeta_T = (12T^2)^{-1}$, the remainder is sublinear provided

$$\sum_{t=1}^{T}\eta_t = o(T), \qquad \frac{\log(T+1)}{\eta_T} = o(T), \qquad \sum_{t=b_T}^{T-1} a_{t,T}^2 = o(T).$$

For example, $\eta_t = t^{-\kappa}$ with $0 < \kappa < 1$ gives potential and bounded-increment contributions $O(T^\kappa \log T + T^{1-\kappa})$; the residual-error control of (10) satisfy the remaining condition.

The result above averages the losses at realized state-action pairs. The next subsection relates them to sampled challenger losses.

B.3 Exploration and trajectory control

We relate stationary comparator rewards to the Gibbs averages controlled in Appendix B.2. The exploration learner receives the online rewards $\hat{r}_t^{\mathrm{ex}} = \hat{\ell}_{t,\tilde{\theta}_t}$ from (A.2). For the analysis, write $r_t^{\mathrm{ex}} = \ell_{\tilde{\theta}_t}$ for the corresponding oracle rewards. Both rewards lie in some interval $[0, L]$. We prove that uniform residual estimation transfers the regret bound between these two reward sequences, and a martingale argument replaces sampled oracle losses by their conditional averages.

Assumption A.2 (Exploration regret)**.** *For every $T \ge 1$ and $\zeta \in (0,1)$, there is a deterministic, nonnegative bound $R(T,\zeta)$ such that*

$$\mathbb{P}\left(\sup_{\mu}\sum_{t=1}^{T-1}\langle d^\mu, \hat{r}_t^{\mathrm{ex}}\rangle - \sum_{t=1}^{T-1}\hat{r}_t^{\mathrm{ex}}(S_t, A_t) \le R(T,\zeta)\right) \ge 1-\zeta.$$

This guarantee holds for the actual explorer, including its forced-action mixture, and for the adaptive rewards generated under the interaction protocol of Appendix A.1.

The comparator is the best fixed stationary policy in hindsight.

Write $\widehat{q}_t$ for the actual conditional challenger law given $\mathcal{F}_t$, and define the gain

$$\bar{r}_t(s,a) = \int_\Theta \ell_\theta(s,a)\widehat{q}_t(d\theta).$$

Then $\bar{r}_t(S_t, A_t) = \langle \widehat{q}_t, \ell_t \rangle$.

Proposition 16 (Exploration and challenger averages)**.** *Fix $1 \le b < T$ and $\zeta \in (0,1)$, and put*

$$M(T,b,\zeta) = L\sqrt{2(T-b)\log(1/\zeta)}.$$

Let $\mathcal{R}_T(\zeta)$ be the event in (A.2). There is an event $\mathcal{M}_{b,T}(\zeta)$, with failure probability at most ζ, on which

$$\sum_{t=b}^{T-1}\left[r_t^{\mathrm{ex}}(S_t,A_t) - \langle \widehat{q}_t, \ell_t\rangle\right] \le M(T,b,\zeta).$$

Moreover, on $\mathcal{R}_T(\zeta) \cap \mathcal{M}_{b,T}(\zeta) \cap \mathcal{D}_{b,T}$,

$$\sup_\mu \sum_{t=b}^{T-1}\langle d^\mu, r_t^{\mathrm{ex}}\rangle \le \sum_{t=b}^{T-1}\langle \widehat{q}_t, \ell_t\rangle + R(T,\zeta) + L(b-1) + M(T,b,\zeta) + 2E\sum_{t=b}^{T-1} a_{t,T}.$$

If $b = b_T$ and $\mathcal{E}_T^{\mathrm{stab}}$ also holds, then

$$(T-b_T)\Gamma^\star \le \sum_{t=b_T}^{T-1}\langle q_t, \ell_t\rangle + R(T,\zeta) + L(b_T-1) + M(T,b_T,\zeta) + 2E\sum_{t=b_T}^{T-1} a_{t,T}.$$

Proof. The interaction protocol gives $\mathcal{L}(\tilde{\theta}_t \mid \mathcal{G}_t) = \widehat{q}_t$. For each candidate, $\ell_t(\theta)$ is $\mathcal{G}_t$-measurable. Therefore

$$\xi_t = \ell_t(\tilde{\theta}_t) - \langle \widehat{q}_t, \ell_t\rangle$$

satisfies $\mathbb{E}[\xi_t \mid \mathcal{G}_t] = 0$ and $|\xi_t| \le L$. The tower property gives $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$ thus ξ_t is $\mathcal{F}_{t+1}$-measurable. Azuma-Hoeffding on the interval $\{b, \dots, T-1\}$ proves (16).

For every fixed comparator μ, boundedness gives

$$\sum_{t=1}^{b-1}\left[\langle d^\mu, \hat{r}_t^{\mathrm{ex}}\rangle - \hat{r}_t^{\mathrm{ex}}(S_t,A_t)\right] \ge -L(b-1).$$

Subtracting this prefix from the full-interval regret bound and taking the supremum over μ shows that, on $\mathcal{R}_T(\zeta)$,

$$\sup_\mu \sum_{t=b}^{T-1}\langle d^\mu, \hat{r}_t^{\mathrm{ex}}\rangle \le \sum_{t=b}^{T-1}\hat{r}_t^{\mathrm{ex}}(S_t,A_t) + R(T,\zeta) + L(b-1).$$

On $\mathcal{D}_{b,T}$, (A.2) gives $\|\hat{r}_t^{\mathrm{ex}} - r_t^{\mathrm{ex}}\|_\infty \le E a_{t,T}$. Replacing the comparator rewards and then the realized rewards by their oracle values costs at most $2E\sum_{t=b}^{T-1} a_{t,T}$. Combining with (B.3) proves (16).

On $\mathcal{E}_T^{\mathrm{stab}}$, the sampled challengers belong to $\mathcal{B}$ on the late interval, almost surely. Applying Lemma 11 gives

$$\sup_\mu \sum_{t=b_T}^{T-1}\langle d^\mu, r_t^{\mathrm{ex}}\rangle \ge (T-b_T)\Gamma^\star.$$

The identification $\widehat{q}_t = q_t$ then proves the last assertion. □

To calirate the probabilities of the events considered in the proof above, take $\zeta_T = (12T^2)^{-1}$, and define

$$R(T) = R(T,\zeta_T), \qquad M(T) = M(T,b_T,\zeta_T), \qquad r_{\mathcal{G}}^{\mathrm{cal}}(T) = r_{\mathcal{G}}(T,b_T,\zeta_T).$$

The events $\mathcal{R}_T = \mathcal{R}_T(\zeta_T)$ and $\mathcal{M}_T = \mathcal{M}_{b_T,T}(\zeta_T)$ each have failure probability at most ζ_T. We note $M(T) = O(\sqrt{T\log T})$. We require $R(T) = o(T)$ (with is satisfied by the regret bounds of the explorer). The residual estimation costs is $2EA_1(T)$, with $A_1(T)$ from (10), and is also sublinear.

B.4 Evidence growth

We combine the two regret bounds. Once the leader is optimal, every sampled challenger is an alternative to $\pi^\star$. Combining with our previous results, we prove that this implies saddle-point convergence to the payoff of the information game, thus the statistical testing evidence gowing at a linear rate.

We reuse the assumptions and notation of Appendices B.1 to B.3. At time T, use $\zeta_T = (12T^2)^{-1}$. The leader convergence guarantee (A.3), the coverage guarantee, and Proposition 7 have a combined failure probability $3\zeta_T$. Denote by $\mathcal{A}_T$ their intersection, under this event, for all sufficiently large T,

$$\hat{\pi}_t = \pi^\star \quad (t \in I_T),$$
$$\sup_{\theta,s,a} |\phi(s,a)^\top \hat{b}_t(\theta) - e_\theta(s,a)| \le a_{t,T} \quad (t \in I_T), \qquad \sup_{t \in I_T} a_{t,T} \le 1.$$

Indeed, $e_T \to 0$ yields policy correctness through the action gap, and $a_{t,T} = O(\sqrt{\log T / t^{1-k}})$ uniformly on I_T tends uniformly to zero when $b_T = \lceil T^\nu \rceil$, $0 < \nu < 1$. For a minimum-score leader, the joint event constructed in Appendix A.3 may be used directly with the radii specified there; their orders are unchanged.

The exploration and sampling events from the preceding subsection give

$$\sup_\mu \sum_{t=1}^{T-1} \langle d^\mu, \hat{r}_t^{\mathrm{ex}} \rangle - \sum_{t=1}^{T-1} \hat{r}_t^{\mathrm{ex}}(S_t, A_t) \le R(T)$$

and

$$\sum_{t=b_T}^{T-1} \left[r_t^{\mathrm{ex}}(S_t, A_t) - \langle \widehat{q}_t, \ell_t \rangle \right] \le M(T),$$

respectively, each with failure probability at most ζ_T.

Theorem 17 (Evidence growth)**.** *Suppose the algorithm samples the challenger exactly from the Gibbs law restricted to* $\mathrm{Alt}_0(\hat{\pi}_t)$*, with the interaction order of Appendix A.1. Let* $\eta_t = t^{-\kappa}$ *and* $b_T = \lceil T^\nu \rceil$*, where* $0 < \kappa, \nu < 1$*, and let* $0 < k < 1$ *be the forced-exploration exponent. There is a deterministic* $f(T) = o(T)$*, such that for every sufficiently large* T*,*

$$\mathbb{P}\left(\mathsf{T}(T) \ge T\Gamma^\star - f(T)\right) \ge 1 - T^{-2}.$$

One may take the explicit remainder in (B.4)*.*

Proof. Fix a sufficiently large T for which $b_T < T$ and the statistical guarantees hold. Define

$$\mathcal{E}_T = \mathcal{A}_T^{\mathcal{B}} \cap \mathcal{C}_T \cap \mathcal{R}_T \cap \mathcal{M}_T.$$

The union bound gives

$$\mathbb{P}(\mathcal{E}_T^c) \le 3\zeta_T + \zeta_T + \zeta_T + \zeta_T = \frac{1}{2T^2} \le T^{-2}.$$

We work pathwise on this intersection for the rest of the proof.

For $b_T \le t < T$, the sampled challenger belongs to $\mathcal{B}$, the actual law equals q_t, and the residual error is at most $a_{t,T}$. Applying theexploration comparison, and challenger regret successively gives

$$\begin{aligned}
(T - b_T)\Gamma^\star_{\mathcal{B}} &\le \sup_\mu \sum_{t=b_T}^{T-1} \langle d^\mu, r_t^{\mathrm{ex}} \rangle \\
&\le \sum_{t=b_T}^{T-1} \langle q_t, \ell_t \rangle + R(T) + L(b_T - 1) + M(T) + 2EA_1(T) \\
&\le \mathsf{T}_{\mathcal{B}}(T) + r_{\mathcal{G}}^{\mathrm{cal}}(T) + R(T) + L(b_T - 1) + M(T) + 2EA_1(T).
\end{aligned}$$

Here the last inequality uses $\mathsf{T}(T) = \inf_{\theta \in \mathcal{B}} S_T(\theta)$. Rearranging proves the claim with

$$\begin{aligned}
f(T) = b_T\Gamma^\star + L(b_T - 1) + 2EA_1(T) \\
+ R(T) + M(T) + r_{\mathrm{G}}^{\mathrm{cal}}(T).
\end{aligned}$$

Every term is deterministic and direct derivation shows that their sum is $o(T)$. □

More generally, with deterministic positive nonincreasing (η_t), it suffices that the calibrated statistical inputs hold, $b_T \to \infty$, $b_T = o(T)$, and

$$\sum_{t=1}^{T} \eta_t = o(T), \qquad \frac{\log(T+1)}{\eta_T} = o(T), \qquad A_1(T) + A_2(T) + R(T) = o(T).$$

The martingale terms are $O(\sqrt{T \log T})$. From (10)

$$f(T) = O\big(T^{\nu} + R(T) + T^{\kappa} \log T + T^{1-\kappa} + T^{(1+k)/2}\sqrt{\log T} + T^{k} \log T + \sqrt{T \log T}\big).$$

Every term is sublinear for $0 < k, \kappa, \nu < 1$ and $R(T) = o(T)$.

Corollarry 18 (Almost-sure evidence growth)**.** *Under the assumptions of Theorem 17, for the algorithm run without stopping,*

$$\liminf_{T\to\infty} \frac{\mathsf{T}(T)}{T} \geq \Gamma^{\star} \qquad \textit{almost surely}.$$

Proof. Since $\sum_{T \geq T_0} \mathbb{P}(\mathcal{E}_T^c) \leq \sum_{T \geq T_0} T^{-2} < \infty$, the Borel-Cantelli lemma implies that $\mathcal{E}_T$ holds for all sufficiently large T, almost surely. On these sample paths, divide the evidence bound by T and use $f(T)/T \to 0$. □

C Gibbs contraction and stopping guarantees

We first relate Gibbs mass to our testing score. We finally analyze the projection-based and sampling-based stopping rules.

C.1 From quadratic scores to Gibbs mass

The Bellman score is quadratic in the empirical moment, $S_T(\theta) = \frac{1}{2\sigma^{\star}} \|m_T(\theta)\|^2_{\Lambda_T^{-1}}$, although it need not be globally quadratic in θ. The Gibbs mass of a fixed alternative set has the same exponential scale as the difference between its minimum score and the unrestricted minimum. We establish a pathwise comparison with a logarithmic remainder by controlling the alternative integral and the normalizing integral separately.

We the finite state and action spaces, bounded linear representation, and fixed nonempty alternative set $\mathcal{B} = \mathrm{Alt}_0(\pi^{\star})$ from Appendix B.1. Assume that $\Theta \subset \mathbb{R}^d$ is compact and convex with nonempty interior, and that Π_0 has a continuous positive density p_0 on Θ. The regularization parameter $\lambda > 0$ is fixed. Write $D_\Theta = \mathrm{diam}(\Theta)$ and assume $\sigma_\star = R_{\max}/(1-\gamma) > 0$. For contraction and stopping, we use the fixed, noise-normalized Gibbs law

$$\Pi_T(\mathrm{d}\theta) = \frac{\exp\big(-S_T(\theta)\big)\Pi_0(\mathrm{d}\theta)}{\int_\Theta \exp\big(-S_T(u)\big)\Pi_0(\mathrm{d}u)}.$$

This is the unit-temperature law for the normalized score S_T, it is distinct from the conditional challenger law with decreasing inverse temperature η_T. Set

$$v_T^{\mathcal{B}} = \inf_{\theta \in \mathcal{B}} S_T(\theta), \qquad v_T^{\Theta} = \min_{\theta \in \Theta} S_T(\theta).$$

The unrestricted minimum exists by continuity and compactness. Neither minimum involves the leader estimator.

Proposition 19 (Gibbs mass and minimum score)**.** *There exist deterministic constants $C_-, C_+ \geq 0$, depending only on the fixed domain, prior, alternative geometry, and boundedness parameters, such that for every $T \geq 2$ and every bounded-data history,*

$$v_T^{\mathcal{B}} - v_T^{\Theta} - d\log T - C_- \leq -\log \Pi_T(\mathcal{B}) \leq v_T^{\mathcal{B}} - v_T^{\Theta} + d\log T + C_+.$$

Equivalently, with a deterministic remainder uniform over histories,

$$-\log \Pi_T(\mathcal{B}) = v_T^{\mathcal{B}} - v_T^{\Theta} + O(\log T).$$

Under the calibration assumptions of Proposition 6, almost surely,

$$-\log \Pi_T(\mathcal{B}) = v_T^{\mathcal{B}} + O(\log T) \quad \textit{for all sufficiently large } T.$$

Proof. Define

$$Z_T^{\mathcal{B}} = \int_{\mathcal{B}} e^{-S_T(\theta)}\Pi_0(\mathrm{d}\theta), \qquad Z_T^{\Theta} = \int_{\Theta} e^{-S_T(\theta)}\Pi_0(\mathrm{d}\theta).$$

Lemma 14 ensures that $p_{\mathcal{B}} = \Pi_0(\mathcal{B}) > 0$. Both integrals are positive, and

$$-\log \Pi_T(\mathcal{B}) = -\log Z_T^{\mathcal{B}} + \log Z_T^{\Theta}.$$

For the alternative integral, the minimum score gives $Z_T^{\mathcal{B}} \le p_{\mathcal{B}} e^{-v_T^{\mathcal{B}}}$. For the reverse bound, use $\rho_0 = \Pi_0(\,\cdot \mid \mathcal{B})$ and the restricted potential from Appendix B.2. Since $-\log Z_T^{\mathcal{B}} = \Phi_T(1) - \log p_{\mathcal{B}}$, equation (14) yields

$$v_T^{\mathcal{B}} - \log p_{\mathcal{B}} \le -\log Z_T^{\mathcal{B}} \le v_T^{\mathcal{B}} + d\log T + K_0 + K_1 - \log p_{\mathcal{B}}.$$

To control the normalizer, let $p_{\min} = \min_{\Theta} p_0 > 0$ and fix a ball $B(z, r_0) \subseteq \Theta$, $r_0 > 0$. For every $u \in \Theta$ and $0 < r \le 1$, set $\alpha_r = r/(D_\Theta + 1)$. Convexity gives

$$(1-\alpha_r)u + \alpha_r B(z, r_0) \subseteq \Theta \cap B(u, r).$$

The left-hand side is a ball of radius $\alpha_r r_0$. If v_d is the volume of the Euclidean unit ball, it follows that

$$\Pi_0\big(\Theta \cap B(u,r)\big) \ge c_\Theta r^d, \qquad c_\Theta = \min\left\{1, p_{\min} v_d \left(\frac{r_0}{D_\Theta + 1}\right)^d\right\} > 0.$$

This bound is uniform over u, including boundary points.

Choose $u_T \in \arg\min_\Theta S_T$. The deterministic Lipschitz bound (B.2) implies

$$S_T(\theta) \le v_T^{\Theta} + K_0 \quad \text{for } \theta \in \Theta \cap B(u_T, T^{-1}).$$

Restricting the normalizing integral to this ball gives

$$c_\Theta T^{-d} e^{-(v_T^{\Theta} + K_0)} \le Z_T^{\Theta} \le e^{-v_T^{\Theta}}.$$

Consequently,

$$-v_T^{\Theta} - d\log T - C_- \le \log Z_T^{\Theta} \le -v_T^{\Theta}, \qquad C_- = K_0 - \log c_\Theta.$$

Combining (C.1) and (C.1) in (C.1) proves (19), taking $C_+ = K_0 + K_1 - \log p_{\mathcal{B}}$.

Unlike a quadratic score centered at an estimator, the Bellman score need not attain zero. However, $0 \le v_T^{\Theta} \le S_T(\theta^\star)$. Proposition 6, applied at failure level T^{-3}, and the Borel–Cantelli lemma give

$$S_T(\theta^\star) \le \bar{\beta}_T(T^{-3}) = O(\log T) \quad \text{eventually almost surely.}$$

Substitution proves (19). □

C.2 Contraction converse and attained rate

Every limiting allocation generated by a trajectory satisfies the stationary flow constraints in the underlying MDP. We use this fact to bound the contraction rate of any adaptive algorithm, we use this obtain the rate attained by our algorithm.

We reuse the assumptions of Appendix C.1, and we recall $B_b = \sup_{\theta\in\Theta} \|b(\theta)\|_2 < \infty$. Let $N_T(\theta) = m_T(\theta) - G_T b(\theta)$ and let $u_T(\xi)$ be noise radius from (7). Proposition 19 gives

$$-\log \Pi_T(\mathcal{B}) = \sigma_\star^{-2} \inf_{\theta\in\mathcal{B}} S_T(\theta) + o(T) \qquad \text{almost surely.}$$

We prove the result below.

Lemma 20 (Uniform comparison with oracle residual losses)**.** *For $T \ge 2$, suppose*

$$\sup_{\theta\in\Theta} \|N_T(\theta)\|_{\Lambda_T^{-1}} \le u_T(\xi).$$

Then

$$\sup_{\theta\in\Theta}\left|S_T(\theta) - \sum_{t<T} \ell_\theta(S_t, A_t)\right| \le \frac{\lambda}{2}B_b^2 + E\sqrt{T-1}\,u_T(\xi) + \frac{1}{2}u_T(\xi)^2.$$

Under any nonanticipating sampling scheme, the left-hand side is $O(\sqrt{T\log T} + \log T) = o(T)$ eventually almost surely.

Proof. Fix θ and abbreviate $b = b(\theta)$ and $N = N_T(\theta)$. Expanding the score and using $\sum_{t<T} \ell_\theta(S_t, A_t) = \frac{1}{2} b^\top G_T b$ gives

$$S_T(\theta) - \sum_{t<T} \ell_\theta(S_t, A_t) = -\frac{\lambda}{2} b^\top G_T \Lambda_T^{-1} b + b^\top G_T \Lambda_T^{-1} N + \frac{1}{2} \|N\|^2_{\Lambda_T^{-1}}.$$

Here we used $G_T - G_T \Lambda_T^{-1} G_T = \lambda G_T \Lambda_T^{-1}$. Since $0 \preceq G_T \Lambda_T^{-1} \preceq I_d$, the first term has absolute value at most $\lambda B_b^2/2$. Moreover,

$$\|G_T b\|^2_{\Lambda_T^{-1}} \le b^\top G_T b = \sum_{t<T} e_\theta(S_t, A_t)^2 \le E^2 (T-1).$$

Cauchy-Schwarz bounds the cross term by $E\sqrt{T-1}\, u_T(\xi)$, and the final term is at most $u_T(\xi)^2/2$. These bounds are uniform in θ. Proposition 7 with $\xi = T^{-3}$ gives $u_T(T^{-3}) = O(\sqrt{\log T})$ and a summable sequence of failure probabilities. Borel-Cantelli proves the almost-sure conclusion. □

Let $\Delta(\mathcal{S} \times \mathcal{A})$ denote the probability simplex. Extend $\mathcal{L}(\omega, \theta)$ to this simplex and define

$$F(\omega) = \inf_{\theta \in \mathcal{B}} \mathcal{L}(\omega, \theta), \qquad \Gamma^\star = \sup_{\omega \in \Omega(P)} F(\omega).$$

For $T \ge 2$, the empirical occupation measure is

$$\widehat{\omega}_T(s, a) = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{1}\{S_t = s, A_t = a\}.$$

We use the stationary flow characterization

$$\Omega(P) = \left\{ \omega \in \Delta(\mathcal{S} \times \mathcal{A}) : \sum_a \omega(x, a) = \sum_{s,a} \omega(s, a) P(x \mid s, a) \text{ for every } x \in \mathcal{S} \right\}.$$

Indeed, given such an ω, its state marginal $\nu(s) = \sum_a \omega(s, a)$ is invariant for the stationary policy $\mu(a \mid s) = \omega(s, a)/\nu(s)$ when $\nu(s) > 0$, with arbitrary choices at zero-mass states. Conversely, every stationary occupation satisfies these constraints.

Lemma 21 (Feasibility of empirical occupation limits)**.** *For any nonanticipating adaptive algorithm collecting observations along a single trajectory of P, almost surely every accumulation point of $(\widehat{\omega}_T)_{T \ge 2}$ belongs to $\Omega(P)$.*

Proof. For each $x \in \mathcal{S}$, define

$$\xi_t(x) = \mathbf{1}\{S_{t+1} = x\} - P(x \mid S_t, A_t).$$

With the filtration of Appendix A.1, $\mathbb{E}[\xi_t(x) \mid \mathcal{G}_t] = 0$, and hence $\mathbb{E}[\xi_t(x) \mid \mathcal{F}_t] = 0$ by the tower property. Moreover, $\xi_t(x)$ is $\mathcal{F}_{t+1}$-measurable and bounded in absolute value by one. Azuma-Hoeffding and a union bound give

$$\mathbb{P}\left(\max_{x \in \mathcal{S}} \left| \sum_{t=1}^{T-1} \xi_t(x) \right| > \sqrt{6(T-1)\log T} \right) \le 2|\mathcal{S}| T^{-3}.$$

By Borel-Cantelli, the normalized sums converge to zero almost surely, simultaneously for every x. The exact identity

$$\sum_a \widehat{\omega}_T(x, a) - \sum_{s,a} \widehat{\omega}_T(s, a) P(x \mid s, a) = \frac{\mathbf{1}\{S_1 = x\} - \mathbf{1}\{S_T = x\}}{T-1} + \frac{1}{T-1} \sum_{t=1}^{T-1} \xi_t(x)$$

therefore shows that flow errors vanish with probability one. Every subsequential limit lies in the simplex and satisfies (C.2), proving the claim. □

Theorem 22 (Lower bound for Gibbs distribution contraction)**.** *Under the preceding boundedness, and prior assumptions, consider any nonanticipating adaptive algorithm* A *collecting observations along a single trajectory and forming* (C.1) *with the same domain and prior. Then, almost surely,*

$$\limsup_{T\to\infty} -\frac{1}{T}\log \Pi_T^{\mathsf{A}}(\mathcal{B}) \le \frac{\Gamma^\star}{\sigma_\star^2}.$$

Proof. Fix such an algorithm and suppress its superscript. Define the oracle score by

$$S_T^\star(\theta) = \sum_{t<T} \ell_\theta(S_t, A_t) = (T-1)\mathcal{L}(\widehat{\omega}_T, \theta).$$

Lemma 20 gives

$$\sup_{\theta\in\Theta} |S_T(\theta) - S_T^\star(\theta)| = o(T) \qquad \text{almost surely.}$$

This uniform bound permits taking infima over the same set $\mathcal{B}$. Since $0 \le F \le L = E^2/2$,

$$\frac{1}{T}\inf_{\theta\in\mathcal{B}} S_T(\theta) = F(\widehat{\omega}_T) + o(1) \qquad \text{almost surely.}$$

Together with (C.2), this yields

$$-\frac{1}{T}\log \Pi_T(\mathcal{B}) = \sigma_\star^{-2} F(\widehat{\omega}_T) + o(1) \qquad \text{almost surely.}$$

For any two probability vectors ω, ω', boundedness gives

$$\sup_{\theta\in\mathcal{B}} |\mathcal{L}(\omega,\theta) - \mathcal{L}(\omega',\theta)| \le L\|\omega - \omega'\|_1,$$

and therefore $|F(\omega) - F(\omega')| \le L\|\omega - \omega'\|_1$. In particular, F is continuous even though $\mathcal{B}$ is defined by strict inequalities. Work on a sample path where (C.2) and Lemma 21 hold. Choose a subsequence along which $F(\widehat{\omega}_T)$ converges to its limsup. Compactness supplies a further subsequence $\widehat{\omega}_{T_k} \to \omega$. The lemma gives $\omega \in \Omega(P)$, so continuity implies

$$\limsup_{T\to\infty} F(\widehat{\omega}_T) = F(\omega) \le \Gamma^\star.$$

Equation (C.2) proves the theorem. □

Theorem 23 (Gibbs contraction rate)**.** *Under the assumptions of Theorem 17, almost surely,*

$$\lim_{T\to\infty} -\frac{1}{T}\log \Pi_T(\mathcal{B}) = \frac{\Gamma^\star}{\sigma_\star^2}.$$

Proof. Apply the converse to our algorithm. For the reverse inequality, use eventual leader correctness from Appendix A.3: almost surely, $\hat{\pi}_T = \pi^\star$ for every sufficiently large T. Consequently,

$$\mathsf{T}(T) = \inf_{\theta\in\mathcal{B}} S_T(\theta) \quad \text{for all sufficiently large } T.$$

Corollary 18 and (C.2) then give

$$\liminf_{T\to\infty} -\frac{1}{T}\log \Pi_T(\mathcal{B}) = \sigma_\star^{-2} \liminf_{T\to\infty} \frac{\mathsf{T}(T)}{T} \ge \frac{\Gamma^\star}{\sigma_\star^2}.$$

Combining this inequality with the converse proves the result. □

These results characterize the contraction rate of the specified Bellman Gibbs law. They do not identify this rate with a universal Kullback-Leibler lower bound for policy identification.

C.3 TEST-BASED STOPPING

The test score minimizes the Bellman accumulated residuals e over the recommendation's alternatives and certifies the recommendation through an anytime confidence bound. Its expected duration is controlled by the time after which the deterministic evidence lower bound exceeds the stopping boundary. The summable failure probabilities in Theorem 17 then control the remaining tail. As in Appendix A.1, define all statistics along the algorithm run without stopping, coupled with the stopped procedure up to

$$\tau_\delta = \inf\{t \geq 1 : \mathsf{T}(t) > \beta_t(\delta)\}, \qquad L_\delta = \log(1/\delta).$$

We assume $0 < \Gamma^\star < \infty$, as established under the conditions of Lemma 11. The confidence calibration required from Appendix A.2 is

$$\mathbb{P}(C_\delta) \geq 1 - \delta, \qquad C_\delta = \{S_t(\theta^\star) \leq \beta_t(\delta) \text{ for every } t \geq 1\}.$$

For the duration bound, the self-normalized boundary in (6) supplies a deterministic $0 \leq K < \infty$ such that

$$\beta_t(\delta) \leq \overline{\beta}_t(\delta) \leq \sigma_\star^2 L_\delta + K\big(1 + \log(t+1) + \log(L_\delta + 1)\big),$$

for every $t \geq 1$ and $\delta \in (0,1)$. The upper envelope $\overline{\beta}_t(\delta)$ is deterministic, although $\beta_t(\delta)$ depends on the observed design. The constant K, the deterministic validity time t_0 of the evidence bound, and its nonnegative remainder f are independent of δ.

Theorem 4 (Correctness and expected sample complexity)**.** *Under the conditions of Theorem 2, the test-based stopping is δ-correct and satisfies*

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \frac{1}{\Gamma^\star}.$$

Proof. We first prove the correctness of the stopping rule. We work on C_δ. If $\hat{\pi}_t \neq \pi^\star$, uniqueness of the optimal policy and $Q_{\theta^\star} = Q^\star$ imply $\theta^\star \in \mathrm{Alt}_0(\hat{\pi}_t)$. Hence

$$\mathsf{T}(t) = \inf_{\theta \in \mathrm{Alt}_0(\hat{\pi}_t)} S_t(\theta) \leq S_t(\theta^\star) \leq \beta_t(\delta).$$

Thus a wrong recommendation cannot satisfy the stopping condition, so

$$\mathbb{P}\big(\tau_\delta < \infty, \hat{\pi}_{\tau_\delta} \neq \pi^\star\big) \leq \delta.$$

We now control the expected stopping time.

For a fixed $\delta \in (0,1)$, define

$$n_\delta = \min\left\{ n \in \mathbb{N} : \begin{array}{l} n \geq \max\{t_0, 2\}, \\ t\Gamma^\star - f(t) > \overline{\beta}_t(\delta) \text{ for every integer } t \geq n \end{array} \right\}.$$

This integer is finite: for fixed δ, both $f(t)$ and $\overline{\beta}_t(\delta)$ are $o(t)$, whereas $\Gamma^\star > 0$. For every $t \geq n_\delta$, the event $\mathcal{E}_t$ of Theorem 17 implies $\mathsf{T}(t) > \beta_t(\delta)$, and therefore $\tau_\delta \leq t$. Consequently,

$$\mathbb{P}(\tau_\delta > t) \leq \mathbb{P}(\mathcal{E}_t^c) \leq t^{-2}, \qquad t \geq n_\delta.$$

This comparison is under the original probability law, without conditioning on C_δ. Summing the tail yields

$$\mathbb{E}[\tau_\delta] = \sum_{t=0}^{\infty} \mathbb{P}(\tau_\delta > t) \leq n_\delta + \sum_{t=n_\delta}^{\infty} t^{-2} \leq n_\delta + \frac{1}{n_\delta - 1} < \infty.$$

In particular, $\tau_\delta < \infty$ almost surely, which completes the δ-correctness argument.

Fix $u > 0$ and set

$$N_\delta(u) = \max\left\{ \lceil t_0 \rceil, 2, \left\lceil \frac{(1+u)\sigma_\star^2 L_\delta}{\Gamma^\star} \right\rceil \right\}.$$

We show that $n_\delta \leq N_\delta(u)$ for all sufficiently small δ. Abbreviate $N = N_\delta(u)$. Since $f(t)/t \to 0$, $N \sim (1+u)\sigma_\star^2 L_\delta/\Gamma^\star$, and $t \mapsto \log(t+1)/t$ is decreasing on $(0,\infty)$,

$$\begin{aligned} &\sup_{t \geq N} \frac{f(t) + K\,(1 + \log(t+1) + \log(L_\delta + 1))}{t} \\ &\qquad \leq \sup_{t \geq N} \frac{f(t)}{t} + \frac{K\,(1 + \log(N+1) + \log(L_\delta + 1))}{N} \longrightarrow 0 \qquad (\delta \downarrow 0). \end{aligned}$$

All suprema here are over integer times. For every $t \geq N$, $\sigma_\star^2 L_\delta / t \leq \Gamma^\star/(1+u)$. Combining this fact with (C.3) and (C.3) gives

$$\inf_{t \geq N_\delta(u)} \frac{t\Gamma^\star - f(t) - \overline{\beta}_t(\delta)}{t} \geq \Gamma^\star - \frac{\Gamma^\star}{1+u} - o(1) > 0$$

for sufficiently small δ. This proves $n_\delta \leq N_\delta(u)$ without any monotonicity requirement on f, $\beta_t(\delta)$, or $\mathsf{T}(t)$. Since $n_\delta \geq 2$, (C.3) now implies

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \lim_{\delta \downarrow 0} \frac{N_\delta(u)+1}{L_\delta} = \frac{(1+u)\sigma_\star^2}{\Gamma^\star}.$$

Letting $u \downarrow 0$ proves the claimed bound. □

The score at round t uses $t-1$ transitions. Thus the number of collected transitions is $\tau_\delta - 1$ and satisfies the same asymptotic upper bound.

An implementation of this test needs a certified lower bound on the constrained minimum. Evaluating a single alternative provides an upper bound and cannot certify the stopping condition.

C.4 Sampling-based stopping

Evaluating the projection statistic requires minimizing S_t over the leader's alternatives. We replace this optimization with a search for a sampled alternative. The resulting test uses Gibbs sampling and greedy-policy comparisons, with a sampling budget calibrated to preserve asymptotic correctness and the leading sample-complexity bound. The argument separates confidence failure from failure to sample an alternative under an incorrect recommendation.

The sampling test. we recall that the score, bounded features, compact convex parameter domain, and continuous positive prior density of Appendix C.1. Assume also that $\theta^\star \in \mathrm{int}(\Theta)$ and that the uniform action gap (A.3) is positive. Write

$$\mathcal{B}_t = \mathrm{Alt}_0(\hat{\pi}_t), \qquad p_t = \Pi_t(\mathcal{B}_t),$$

where Π_t is the unrestricted law in (C.1). For a deterministic integer budget $M(t,\delta) \geq 1$, generate fresh conditionally independent candidates $\vartheta_{t,1}, \ldots, \vartheta_{t,M(t,\delta)}$ from Π_t and set

$$\tau_\delta^{\mathrm{samp}} = \inf\left\{ t \geq 1 : \vartheta_{t,m} \notin \mathcal{B}_t \text{ for every } 1 \leq m \leq M(t,\delta) \right\}.$$

The recommendation at stopping is $\hat{\pi}_{\tau_\delta^{\mathrm{samp}}}$. At each round, sampling can end as soon as an alternative is found, in which case data collection continues. Acceptance requires all $M(t,\delta)$ candidates to pass. The candidates come from the unconditioned, unit-temperature law for the normalized score and are used only for stopping; they do not enter the estimator or exploration updates. For analysis, continue the observation process and these auxiliary draws indefinitely, and enlarge $\mathcal{F}_t$ to include the preceding stopping draws. Conditional on this pre-check history, the current draws are independent and have law Π_t. The evidence-growth guarantees are unchanged under this coupling. Next, we bound the

Gibbs mass near $\theta^\star$. The score is quadratic there, because the optimal greedy policy is unchanged in a sufficiently small neighborhood. This permits a half-ball argument for the alternative integral. The unrestricted normalizer, however, is not a single Gaussian integral, so the volume factor must be retained.

Lemma 24 (Alternative mass under an incorrect recommendation)**.** *Under the preceding assumptions, there exists a deterministic $a_\star \in (0,1]$, independent of t and δ, such that on every bounded-data history and for every $t \geq 1$,*

$$Z_t^\Theta \geq a_\star (t+1)^{-d/2} e^{-S_t(\theta^\star)/\sigma_\star^2}.$$

Moreover,

$$\hat{\pi}_t \neq \pi^\star \quad \Longrightarrow \quad p_t \geq a_\star (t+1)^{-d/2} e^{-S_t(\theta^\star)/\sigma_\star^2}.$$

The constant may depend on the instance, but is not needed to implement the sampling budget below.

Proof. Interiority and the positive uniform gap allow us to choose $r_\star > 0$ such that

$$B(\theta^\star, r_\star) \subseteq \Theta, \qquad 2L_\phi r_\star < \Delta_{\min}.$$

For every θ in this ball, state s, and action $a \neq a^\star(s)$,

$$Q_\theta(s, a^\star(s)) - Q_\theta(s, a) \geq \Delta_{\min} - 2L_\phi r_\star > 0.$$

Thus every parameter in the ball has unique greedy policy $\pi^\star$. On a history with $\hat{\pi}_t \neq \pi^\star$, the whole ball lies in $\mathcal{B}_t$.

Fix a bounded-data history. For $i < t$, define

$$z_i = \gamma\phi(S_{i+1}, \pi^\star(S_{i+1})) - x_i, \qquad D_t = \sum_{i<t} x_i z_i^\top, \qquad C_z = (1+\gamma)L_\phi.$$

For $\|u\|_2 < r_\star$, the fixed greedy policy on this ball gives the exact identities

$$Y_i(\theta^\star + u) = Y_i(\theta^\star) + z_i^\top u, \qquad m_t(\theta^\star + u) = m_t(\theta^\star) + D_t u.$$

Hence, with $g_t = D_t^\top \Lambda_t^{-1} m_t(\theta^\star)$ and $H_t = D_t^\top \Lambda_t^{-1} D_t$,

$$S_t(\theta^\star + u) = S_t(\theta^\star) + g_t^\top u + \tfrac{1}{2} u^\top H_t u.$$

The design contraction inequality used in Proposition 7 yields, for every u,

$$u^\top H_t u = \left\| \sum_{i<t} x_i z_i^\top u \right\|^2_{\Lambda_t^{-1}} \leq \sum_{i<t} (z_i^\top u)^2 \leq (t-1) C_z^2 \|u\|_2^2.$$

Restrict to the half-ball

$$\mathcal{U}_t = \left\{ u : \|u\|_2 < \frac{r_\star}{\sqrt{t+1}}, \quad g_t^\top u \leq 0 \right\}.$$

On this set, the linear term is nonpositive and the quadratic term is at most $C_z^2 r_\star^2/2$. The set has volume at least $v_d r_\star^d (t+1)^{-d/2}/2$, including when $g_t = 0$. Writing $p_{\min} = \min_\Theta p_0$, we obtain

$$\int_{\theta^\star + \mathcal{U}_t} e^{-S_t(\theta)/\sigma_\star^2} \Pi_0(\mathrm{d}\theta) \geq \frac{p_{\min} v_d r_\star^d}{2} e^{-C_z^2 r_\star^2/(2\sigma_\star^2)} (t+1)^{-d/2} e^{-S_t(\theta^\star)/\sigma_\star^2}.$$

Taking $a_\star$ to be the prefactor, decreased to at most one, proves (24). If the leader is wrong, the same lower bound applies to the alternative integral. Since $Z_t^\Theta \leq 1$, division proves (24). Neither Bellman closure nor feature coverage is used in this local calculation. □

Test calibration. Assume the anytime coverage (C.3) and use the deterministic envelope from (6). Define

$$L_\delta = \log(1/\delta), \qquad b_t(\delta) = \frac{\bar{\beta}_t(\delta)}{\sigma_\star^2}, \qquad h_\delta = \log(e + L_\delta),$$

and choose

$$M(t, \delta) = \left\lceil h_\delta (t+1)^{d/2} e^{b_t(\delta)} \log \frac{t(t+1)}{\delta^2} \right\rceil.$$

The polynomial factor compensates for the local volume in (24). The factor $h_\delta \to \infty$ absorbs the unknown constant $a_\star$ as $\delta \downarrow 0$, while its logarithm contributes only a lower-order term to the effective stopping boundary.

Theorem 25 (Asymptotic guarantees for sampling-based stopping)**.** *Under the preceding assumptions, suppose Theorem 17 holds with $0 < \Gamma^\star < \infty$ and deterministic remainder $f(t) = o(t)$ independent of δ. With exact, conditionally independent Gibbs samples, the rule (C.4) with budget (C.4) satisfies $\mathbb{E}[\tau_\delta^{\mathrm{samp}}] < \infty$ for every $\delta \in (0,1)$. For each fixed instance,*

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{P}\left(\tau_\delta^{\mathrm{samp}} < \infty, \hat{\pi}_{\tau_\delta^{\mathrm{samp}}} \neq \pi^\star\right)}{\delta} \leq 1,$$

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}[\tau_\delta^{\mathrm{samp}}]}{\log(1/\delta)} \leq \frac{\sigma_\star^2}{\Gamma^\star}.$$

Proof. Fix $\delta \in (0,1)$ and let $\mathcal{Z}_t$ be the event that all $M(t,\delta)$ candidates at round t lie outside $\mathcal{B}_t$. Conditional independence gives

$$\mathbb{P}(\mathcal{Z}_t \mid \mathcal{F}_t) = (1-p_t)^{M(t,\delta)} \le e^{-M(t,\delta)p_t}.$$

Let C_δ be the anytime confidence event in (C.3) and define

$$\mathcal{D}_{t,\delta} = \{\hat{\pi}_t \neq \pi^\star,\ S_t(\theta^\star) \le \bar{\beta}_t(\delta)\} \in \mathcal{F}_t.$$

On C_δ, an incorrect stop at round t implies $\mathcal{D}_{t,\delta} \cap \mathcal{Z}_t$. A union bound followed by conditional expectation therefore yields

$$\begin{aligned}\mathbb{P}\big(\tau_\delta^{\mathrm{samp}} < \infty, \hat{\pi}_{\tau_\delta^{\mathrm{samp}}} \neq \pi^\star\big) \\ \le \delta + \sum_{t\ge 1} \mathbb{E}\Big[\mathbf{1}_{\mathcal{D}_{t,\delta}} e^{-M(t,\delta)p_t}\Big].\end{aligned}$$

The conditional calculation uses the measurable local event $\mathcal{D}_{t,\delta}$; the global event C_δ is used only to separate confidence failure.

By Lemma 24, on $\mathcal{D}_{t,\delta}$,

$$p_t \ge a_\star (t+1)^{-d/2} e^{-b_t(\delta)}, \qquad M(t,\delta)p_t \ge a_\star h_\delta \log\frac{t(t+1)}{\delta^2}.$$

For each fixed instance, choose $\delta_\star > 0$ so that $a_\star h_\delta \ge 1$ whenever $0 < \delta \le \delta_\star$. For these confidence levels, the summand in (C.4) is at most $\delta^2/[t(t+1)]$. Since $\sum_{t\ge1}[t(t+1)]^{-1} = 1$,

$$\mathbb{P}\big(\tau_\delta^{\mathrm{samp}} < \infty, \hat{\pi}_{\tau_\delta^{\mathrm{samp}}} \neq \pi^\star\big) \le \delta + \delta^2, \qquad 0 < \delta \le \delta_\star.$$

This proves (25).

Termination and expected duration. Since $S_t \ge \mathsf{T}(t)$ on $\mathcal{B}_t$, the normalizer lower bound (24) gives, on every history,

$$p_t \le a_\star^{-1}(t+1)^{d/2} \exp\left(-\frac{\mathsf{T}(t) - S_t(\theta^\star)}{\sigma_\star^2}\right).$$

This also holds when $\mathcal{B}_t$ is empty, with $\mathsf{T}(t) = +\infty$ and $p_t = 0$. For $t \ge 2$, let

$$Q_t = \{S_t(\theta^\star) \le \bar{\beta}_t(t^{-3})\}.$$

Calibration gives $\mathbb{P}(Q_t^c) \le t^{-3}$. Define the deterministic effective boundary

$$\begin{aligned}B_t^{\mathrm{samp}}(\delta) = \bar{\beta}_t(t^{-3}) \\ + \sigma_\star^2\left[\log M(t,\delta) + \log(1/a_\star) + \left(\frac{d}{2}+2\right)\log(t+1)\right].\end{aligned}$$

On every history in Q_t where $\mathsf{T}(t) > B_t^{\mathrm{samp}}(\delta)$,

$$\mathbb{P}(\mathcal{Z}_t^c \mid \mathcal{F}_t) = 1 - (1-p_t)^{M(t,\delta)} \le M(t,\delta)p_t \le (t+1)^{-2}.$$

The constant $a_\star$ enters this analytical boundary only, not the implemented rule.

We next verify the boundary's growth. Writing $g_{t,\delta} = \log\big(t(t+1)/\delta^2\big)$, we have $\log 2 \le g_{t,\delta} \le 2(L_\delta + \log(t+1))$. Also $1 \le h_\delta \le 1 + L_\delta$ and $b_t(\delta) \ge 0$. The ceiling in (C.4) consequently gives

$$\log M(t,\delta) \le b_t(\delta) + \frac{d}{2}\log(t+1) + \log h_\delta + \log g_{t,\delta} + \log 3.$$

Since $\bar{\beta}_t(t^{-3}) = O(1 + \log(t+1))$, these inequalities yield a deterministic $K_{\mathrm{samp}} < \infty$, independent of t, δ, such that for $t \ge 2$,

$$B_t^{\mathrm{samp}}(\delta) \le \sigma_\star^2 L_\delta + K_{\mathrm{samp}}\left(1 + \log(t+1) + \log(L_\delta + 1)\right).$$

Let n_δ^{samp} be the smallest integer at least $\max\{t_0, 2\}$ such that $t\Gamma^\star - f(t) > B_t^{\mathrm{samp}}(\delta)$ for every $t \ge n_\delta^{\mathrm{samp}}$. It is finite by (C.4) and $f(t) = o(t)$. For $t \ge n_\delta^{\mathrm{samp}}$, evidence growth implies $\mathbb{P}(\mathsf{T}(t) \le$

$B_t^{\mathrm{samp}}(\delta)) \le t^{-2}$. Since $\{\tau_\delta^{\mathrm{samp}} > t\} \subseteq \mathcal{Z}_t^c$, splitting according to this score event and $\mathcal{Q}_t$, and then using (C.4), gives

$$\mathbb{P}(\tau_\delta^{\mathrm{samp}} > t) \le t^{-2} + t^{-3} + (t+1)^{-2}, \qquad t \ge n_\delta^{\mathrm{samp}}.$$

Summing this tail yields

$$\mathbb{E}[\tau_\delta^{\mathrm{samp}}] \le n_\delta^{\mathrm{samp}} + \frac{3}{n_\delta^{\mathrm{samp}} - 1} < \infty.$$

Finally, (C.4) has the same form as (C.3). The uniform crossing argument in Appendix C.3 therefore gives, for each $u > 0$ and all sufficiently small δ,

$$n_\delta^{\mathrm{samp}} \le \left\lceil \frac{(1+u)\sigma_\star^2 L_\delta}{\Gamma^\star} \right\rceil.$$

Divide (C.4) by L_δ and let $\delta \downarrow 0$, then $u \downarrow 0$, to obtain (25). □

Computational cost. The cost of the sampling test increases as evidence accumulates and alternatives become rare. After leader stabilization, Theorem 23 gives, almost surely,

$$p_t = \exp(-t\Gamma^\star/\sigma_\star^2 + o(t)).$$

At a fixed history, the mean number of independent Gibbs draws needed to find an alternative is therefore

$$p_t^{-1} = \exp(t\Gamma^\star/\sigma_\star^2 + o(t)).$$

The relevant scale is reached when $M(t,\delta)p_t$ is of order one. The budget (C.4) satisfies

$$\log M(t,\delta) = L_\delta + O(1 + \log(t+1) + \log(L_\delta + 1)),$$

so balancing these quantities locates the transition at the first-order scale

$$t_\delta \sim \frac{\sigma_\star^2 L_\delta}{\Gamma^\star}.$$

More precisely, for any fixed $u \in (0,1)$, set $t_\delta^\pm = \lceil (1 \pm u)\sigma_\star^2 L_\delta/\Gamma^\star \rceil$. On the probability-one event of stabilization and contraction,

$$\log\big(M(t_\delta^\pm, \delta)p_{t_\delta^\pm}\big) = \mp u L_\delta + o(L_\delta).$$

Thus the conditional probability of accepting the test tends to zero at t_δ^- and to one at t_δ^+. At the latter time, the full budget is examined with probability tending to one. These statements concern checks on the continued process, whose collection rule does not depend on δ.

Substituting (C.4) into the budget expresses the cost in terms of the confidence level. With $c_\lambda = L_\phi^2/(d\lambda) > 0$, the chosen envelope gives

$$e^{b_t(\delta)} = \delta^{-1}\,(1 + c_\lambda(t-1))^{d/2}\,.$$

Since $h_\delta \sim \log L_\delta$ and $\log\left(t_\delta(t_\delta+1)/\delta^2\right) \sim 2L_\delta$, every deterministic integer sequence satisfying (C.4) obeys

$$M(t_\delta, \delta) \sim 2c_\lambda^{d/2}\left(\frac{\sigma_\star^2}{\Gamma^\star}\right)^d \frac{L_\delta^{d+1}\log L_\delta}{\delta}.$$

Consequently, a completed check at the stopping scale uses an inverse-δ number of draws, up to logarithmic factors, while the leading number of environment transitions is logarithmic in $1/\delta$. The budget bounds the number of draws per check, not the cost of producing each draw or the total work over all checks. The Bellman score is only piecewise quadratic; Gaussian proposals are not in general exact draws from (C.1). Approximate sampling and approximate alternative-membership tests require additional error control before the theorem applies.

D Experiments and implementation details

D.1 Algorithm implementation

The implementation has three components: a leader estimating the optimal action-value function, a challenger representing an alternative policy, and an exploration policy. The Bellman formulation changes the statistical link between them: candidates are scored by their empirical Bellman residuals, and exploration is driven by their fitted conditional Bellman residuals. The leader can therefore be chosen independently of the testing score. We describe these components in the order used by the algorithm, retaining tractable neural proposals where exact Gibbs sampling is unavailable.

D.1.1 Leader estimation and interaction protocol

The leader is learned by applying an offline RL algorithm to the growing transition dataset $\mathcal{D}_t$. This choice reflects the separation between data collection and policy identification: the exploration explorer generates the observations, while the leader must learn from data collected by a sequence of different behavior policies. We have tested implicit Q-learning (IQL; Kostrikov et al. (2021)) and conservative Q-learning (CQL; Kumar et al. (2020)) as leader estimators, with both yielding satisfactory preliminary behavior.

Let $\hat{Q}_t = Q_{w_t}$ denote the resulting critic. The common regression component uses a frozen bootstrap value V_t^{boot}:

$$y_{i,t} = r_i + \gamma(1 - d_i)V_t^{\mathrm{boot}}(s_i'), \qquad \widehat{L}_t(Q;\mathcal{B}) = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\big(Q(s_i,a_i) - y_{i,t}\big)^2,$$

where d_i indicates a terminal transition of the task being evaluated. A time-limit truncation suppresses bootstrapping only when the time limit is part of that task; terminal successors have zero continuation value. The bootstrap value and additional objectives are supplied by the chosen offline learner.

Implicit Q-learning. IQL fits a state-value function by upper-expectile regression of the critic values at dataset actions. For an expectile level $q \in (1/2, 1)$, its value objective has the form

$$\mathbb{E}_{(s,a)\sim\mathcal{D}_t}\left[\ell_q\big(Q_t^-(s,a) - V(s)\big)\right], \qquad \ell_q(u) = \big|q - \mathbf{1}\{u < 0\}\big|u^2.$$

The fitted value supplies V_t^{boot} in (D.1.1), avoiding explicit maximization over unobserved actions in the critic update. When a separate recommendation actor is used, IQL extracts it through advantage-weighted regression on dataset actions. The expectile level, weighting temperature, and any weight clipping are part of the estimator configuration.

Conservative Q-learning. CQL augments Bellman regression with a penalty discouraging large values for actions poorly represented in the dataset. For finite actions, a standard form is

$$\widehat{L}_t(Q;\mathcal{B}) + \alpha_{\mathrm{CQL}}\left[\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\sum_a e^{Q(s_i,a)} - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}Q(s_i,a_i)\right].$$

For scoring a candidate Q, the responses instead use that candidate's own continuation values:

$$Y_i(Q) = r_i + \gamma(1 - d_i)\max_a Q(s_i', a) - Q(s_i, a).$$

Thus the leader's frozen regression targets in (D.1.1) are not substituted into the testing score. Let $\psi(s,a) \in \mathbb{R}^{d_\psi}$ be a fixed instrument vector, and define

$$\Lambda_t^\psi = \lambda I + \sum_{i<t}\psi_i\psi_i^\top, \qquad m_t^\psi(Q) = \sum_{i<t}\psi_i Y_i(Q),$$
$$\widehat{S}_t(Q) = \tfrac{1}{2}\|m_t^\psi(Q)\|^2_{(\Lambda_t^\psi)^{-1}}, \qquad \hat{b}_t^\psi(Q) = (\Lambda_t^\psi)^{-1}m_t^\psi(Q),$$

where $\psi_i = \psi(s_i, a_i)$. The residual predictor is $\hat{g}_{t,Q}(s,a) = \psi(s,a)^\top\hat{b}_t^\psi(Q)$. With $\psi = \phi$ and $Q = Q_\theta$, these are exactly the estimators of Appendix A.2. The same matrix is shared by all candidates; only the residual responses change.

A nonlinear residual critic can instead fit the conditional mean of $Y_i(Q)$ by regression on pre-transition replay:

$$\hat{g}_{t,Q} \in \arg\min_{g\in\mathcal{G}}\frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t}\big(Y_i(Q) - g(s_i,a_i)\big)^2 + \mathrm{pen}(g).$$

The reward is the square of this fitted mean. Squaring an individual response would also reward conditional target variance. Approximation and optimization errors of the nonlinear fit are additional to the regression error controlled by Proposition 7. Its squared predictions do not automatically define the moment score in (D.1.1).

The interaction order follows Appendix A.1. At the start of round t, the explorer selects a_t using the available history. A challenger $\widetilde{Q}_t$ is subsequently constructed using fresh randomness, independently

of this action conditional on the history. Its conditional Bellman residual is fitted on pre-transition replay, and the fitted predictor is frozen to define exploration feedback. After observing the environment reward and successor state, the transition is added to replay, the explorer receives its exploration feedback, and the networks are updated for subsequent decisions. In particular, the candidate pool, its evaluation batches, and all residual fits use information available before the new transition and do not depend on the current action draw.

The pseudocode in Algorithm 3 summarizes this implementation. The leader and explorer keep their parameters between updates; invoking an offline learner means taking the prescribed replay updates, rather than retraining from initialization at each round.

Algorithm 3: Deep implementation of TQS

input : Interaction budget T, offline learner OffRL, explorer Reg, candidate generator SampleQ, residual fitter FitResidual, pool size K, clipping level E.

Initialize replay $\mathcal{D}_1$, leader $\hat{Q}_1$, recommendation $\hat{\pi}_1$, and exploration policy μ_1;
for $t = 1, \dots, T$ **do**
  Select $A_t \sim \mu_t(\cdot \mid S_t)$ before generating the challenger;
  Construct probes and scoring/fitting batches from pre-transition information;
  Draw $Q^{(1:K)} \leftarrow \mathsf{SampleQ}(\hat{Q}_t, \mathcal{D}_t, K)$ with fresh randomness independent of the current action draw;
  Select $\widetilde{Q}_t$ by Bellman scoring, screening, and the fallback rule in Appendix D.1.3;
  Fit $\hat{g}_{t,\widetilde{Q}_t} \leftarrow \mathsf{FitResidual}(\widetilde{Q}_t, \mathcal{D}_t)$ and freeze $\hat{r}_t^{\mathrm{ex}} = \frac{1}{2}[\mathrm{clip}_{[-E,E]}(\hat{g}_{t,\widetilde{Q}_t})]^2$;
  Execute a_t; observe (r_t, s_t', d_t) and append $(s_t, a_t, r_t, s_t', d_t)$ to replay, giving $\mathcal{D}_{t+1}$;
  Update Reg with the frozen residual reward to obtain μ_{t+1};
  Update $(\hat{Q}_{t+1}, \hat{\pi}_{t+1}) \leftarrow \mathsf{OffRL}(\mathcal{D}_{t+1})$ using environment rewards;
  Set $S_{t+1} = S_t'$, or reset after an episodic termination;
return *the recommendation* $\hat{\pi}_{T+1}$;

D.1.2 Generating candidate value functions

Write P_t for the candidate proposal distribution, conditional on the data and current networks. The ideal construction uses the unrestricted Gibbs law associated with the testing score, followed by conditioning on the alternative set. The constructions below provide tractable proposals with different uncertainty representations.

Gaussian last-layer sampling. For a network with a linear output head, write

$$Q_{w,\theta}(s,a) = \phi_w(s,a)^\top \theta.$$

Freeze the representation at its current value and construct

$$A_t^{\mathrm{head}} = \lambda_{\mathrm{head}} I + \sum_{i \in \mathcal{I}_t} \phi_{w_t}(s_i,a_i)\phi_{w_t}(s_i,a_i)^\top, \qquad \theta^{(k)} \sim \mathcal{N}\Big(\hat{\theta}_t, \sigma_t^2 (A_t^{\mathrm{head}})^{-1}\Big),$$

where $\mathcal{I}_t$ indexes the replay design used to construct the covariance and $\lambda_{\mathrm{head}} > 0$ is the proposal regularizer. Each draw defines a candidate $Q^{(k)} = Q_{w_t,\theta^{(k)}}$. This is a Gaussian proposal centered at the leader. As noted in Russo (2019) sampling does not require explicitly inverting A_t^{head}. Let Φ_t have rows $\phi_{w_t}(s_i,a_i)^\top$, $i \in \mathcal{I}_t$, and draw independent standard Gaussian vectors $\xi^{(k)} \in \mathbb{R}^{|\mathcal{I}_t|}$ and $z^{(k)} \in \mathbb{R}^{\dim(\theta)}$. Solve the randomized least-squares problem

$$u^{(k)} \in \arg\min_u \frac{1}{2}\|\Phi_t u - \sigma_t \xi^{(k)}\|_2^2 + \frac{\lambda_{\mathrm{head}}}{2}\left\|u - \frac{\sigma_t}{\sqrt{\lambda_{\mathrm{head}}}} z^{(k)}\right\|_2^2,$$
$$\theta^{(k)} = \hat{\theta}_t + u^{(k)}.$$

The normal equations give $A_t^{\mathrm{head}} u^{(k)} = \sigma_t(\Phi_t^\top \xi^{(k)} + \sqrt{\lambda_{\mathrm{head}}} z^{(k)})$. The right-hand side has covariance $\sigma_t^2 A_t^{\mathrm{head}}$, so an exact solve yields precisely the law in (D.1.2). This permits least-squares solvers or iterative optimization using feature products without constructing an inverse. Both perturbations are required for this covariance identity.

Neural tangent kernel sampling. A richer local model allows perturbations of all network parameters. Define the tangent features and their kernel by

$$\psi_t(s,a) = \nabla_w Q_{w_t}(s,a), \qquad k_t\big((s,a),(s',a')\big) = \psi_t(s,a)^\top \psi_t(s',a').$$

The linearized candidate family is

$$Q_t^{(k)}(s,a) = \hat{Q}_t(s,a) + \psi_t(s,a)^\top u^{(k)}, \qquad u^{(k)} \sim \mathcal{N}\Big(0, \sigma_t^2 (A_t^{\mathrm{NTK}})^{-1}\Big),$$

with $A_t^{\mathrm{NTK}} = \lambda_{\mathrm{NTK}} I + \sum_{i \in \mathcal{I}_t} \psi_t(s_i,a_i)\psi_t(s_i,a_i)^\top$. This defines a Gaussian law over linearized functions, with covariance $\sigma_t^2 \psi_t(s,a)^\top (A_t^{\mathrm{NTK}})^{-1} \psi_t(s',a')$. It extends last-layer sampling by including uncertainty in the local representation directions. A kernel formulation or a low-rank approximation avoids explicitly forming a matrix over all parameters; these choices must preserve joint sampling across the states and actions used to evaluate each candidate. The linearized function in (D.1.2) is the sampled object. Evaluating the perturbed network $Q_{w_t + u^{(k)}}$ instead introduces an additional linearization approximation.

Randomized ensembles. An alternative is to maintain several heads or networks trained with bootstrap weights, randomized prior functions, or perturbed regression targets (Russo, 2019; Osband et al., 2019; 2016). Their outputs form the candidate pool. Shared representations reduce the cost, while separate networks allow representation variability. The randomization should persist over multiple updates so that the pool represents coherent value hypotheses. These proposals can be used with the same selection procedure as Gaussian candidates, without requiring an explicit density over network parameters.

D.1.3 Selecting a challenger

Challenger selection combines compatibility with the observations and disagreement with the recommendation. Compatibility is measured by the candidate's Bellman score, independently of the leader's regression loss. Disagreement is checked using the candidate's own action values.

Exact checks in the linear case. For a deterministic recommendation and a finite state-action space, define $v_t(s,a) = \phi(s,a) - \phi(s, \hat{\pi}_t(s))$. Then

$$\theta \in \mathrm{Alt}_0(\hat{\pi}_t) \quad \Longleftrightarrow \quad \max_{s,a} v_t(s,a)^\top \theta > 0.$$

Membership of a sampled parameter requires only evaluating linear inequalities. Searching within a nonempty compact polyhedral set $\mathcal{P}_t$ can likewise be reduced to the linear programs $\max_{\theta \in \mathcal{P}_t} v_t(s,a)^\top \theta$, one per pair.

Disagreement on sampled states. For a large state space, construct a probe set $\mathcal{S}_t^{\mathrm{probe}}$ and evaluate

$$\mathrm{Gap}_t(Q) = \max_{s \in \mathcal{S}_t^{\mathrm{probe}}} \Big(\max_a Q(s,a) - Q(s, \hat{\pi}_t(s)) \Big).$$

A candidate passes if $\mathrm{Gap}_t(Q) > \kappa$, where $\kappa \geq 0$ is a chosen margin. The comparison is invariant to state-dependent additive shifts of the candidate's values. Probes can mix recent trajectory states, replay states, and draws from a specified valid-state distribution when available. A positive gap supplies a witness of disagreement; agreement on the probes does not imply agreement throughout the state space. In continuous actions, any action improving on $\hat{\pi}_t(s)$ supplies such a witness even if maximization is approximate. Probe randomness is independent of the current action draw conditional on the available history.

Selection from a finite pool. Given K candidates, define

$$\mathcal{K}_t = \Big\{ k \in \{1,\dots,K\} : \mathrm{Gap}_t(Q^{(k)}) > \kappa, \quad \widehat{S}_t(Q^{(k)}) \leq c_t^{\mathrm{fit}} \Big\}.$$

If independent proposals already follow the unrestricted Bellman Gibbs law, uniform selection from a nonempty eligible pool gives that law conditioned on the screening region. Gaussian head and NTK

proposals generally require importance correction. For proposals with density p_t^{prop} positive wherever the target density p_0 is positive, use normalized weights

$$w_k \propto \frac{p_0(\theta^{(k)})}{p_t^{\text{prop}}(\theta^{(k)})} \exp\Big(-\eta_t \widehat{S}_t(Q^{(k)})\Big) \mathbf{1}\{k \in \mathcal{K}_t\}.$$

The prior is zero outside Θ. Weights are evaluated on the log scale. Finite-pool importance resampling is not an exact conditional Gibbs draw, even with exact scores; finite probes and a hard filter introduce further discrepancies. For ensembles without a tractable proposal density, score-only reweighting is a heuristic. It does not supply the density ratio in (D.1.3). Report pool size, effective sample size $1/\sum_k w_k^2$, and the frequency of numerical or empty-pool failures.

If the pool is empty, additional batches may be drawn up to a prescribed budget. A concrete fallback is to select among candidates passing the fit filter, or use the leader if that set is also empty. This fallback need not be an alternative and must be recorded separately. Its fitted Bellman residual need not vanish, even when the leader itself is selected. An explicit exploration mixture maintains data collection independently of the fallback. Failure to find a challenger in a heuristic pool is not the calibrated sampling test of Appendix C.4.

D.1.4 Exploration updates

The selected challenger supplies the fitted-residual reward

$$\hat{r}_t^{\text{ex}}(s,a) = \frac{1}{2} \operatorname{clip}_{[-E,E]}\big(\hat{g}_{t,\widetilde{Q}_t}(s,a)\big)^2.$$

The leader is trained on environment rewards, while the explorer is trained on this estimated Bellman-residual reward. The two learners maintain separate targets and objectives. For bounded neural candidates $|Q| \le Q_{\max}$, a valid residual bound is $E = R_{\max} + (1+\gamma)Q_{\max}$; the implementation must specify how candidate boundedness is enforced. A smaller clipping threshold changes the oracle objective when the true residual exceeds it. The explorer can be instantiated using mirror descent or online policy-gradient updates, holding the fitted residual reward fixed within each update block.

Metrics. The learner's recommendation is evaluated every peridically transitions, separately from the behaviour policy. Empirical first identification is not a certified stopping decision; stopping statistics and error rates are reported separately when a calibrated test is implemented.

- **Identification rate** (DeepSea): the fraction of seeds with $\hat{\pi}_t \in \text{Greedy}(Q^\star)$ on *every* reachable state. $Q^\star$ must be computed from the tabular model for the same discount and terminal convention used by the learner, with comparison tolerance 10^{-9}. Uniform optimality over reachable states is stronger than an initial-state value-gap criterion for that same objective.
- **Return of the recommended policy**, computed exactly on DeepSea and by a greedy rollout of one full episode on Cartpole swingup.
- **Swingup rate** (Cartpole swingup): the fraction of seeds whose recommendation has reached a return of at least 100 (300, 700) by time t. Since $Q^\star$ is unknown there, this thresholded return replaces exact identification, and the curves are monotone by construction.
- **Transitions to identification**: the first evaluation at which a run meets the criterion. Runs that never do are censored at the budget, counted in the denominator, and marked in the figures.

Uncertainty reporting. We use 50 seeds for the linear agents and 20 for the neural agents. Rates are reported with 95% Wilson intervals. For transitions to identification, report the censoring fraction and estimate quantiles with censoring retained; a median not reached within the budget is reported as unreached.