PROVABLY EFFICIENT ENSEMBLE SAMPLING FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Ensemble sampling offers a simple and practical approach to exploration in reinforcement learning (RL), yet establishing regret guarantees is challenging when persistent perturbations influence subsequent data collection. We propose an ensemble sampling framework with two variants that maintain persistently perturbed datasets, uniformly select one ensemble member per episode, and execute its greedy policy. By retaining historical perturbations and fitting only the selected member, the proposed variants avoid repeated noise resampling and iterative posterior sampling, offering a simple design amenable to deep RL. We establish regret guarantees for both variants in linear Markov decision processes (MDPs). The signed variant achieves a leading regret term of $\widetilde{\mathcal{O}}(d^2H^2\sqrt{T})$ with $\widetilde{\mathcal{O}}(dH)$ ensemble members, matching the leading regret rate of a Thompson sampling-type method for linear MDPs while retaining perturbations across episodes. Here, d is the feature dimension, H the episode horizon, and T the total number of interaction steps. The absolute variant improves this leading regret term to $\widetilde{\mathcal{O}}(d^2H^{3/2}\sqrt{T})$ with only $\widetilde{\mathcal{O}}(d)$ members, at the cost of an additional reference regression at each step. Our analysis handles the dependence between persistent perturbations and adaptively collected data by establishing stochastic optimism through joint perturbation control across steps for the signed variant and recursive ensemble-average optimism for the absolute variant. Experiments demonstrate that the proposed algorithms achieve performance competitive with existing baselines despite the simplicity of their ensemble sampling mechanism.

1 INTRODUCTION

Efficient exploration is a central challenge in reinforcement learning (RL), where an agent seeks to achieve high cumulative reward with limited interactions with an unknown environment. Two major approaches are optimistic and randomized exploration. Optimistic methods assign high values to uncertain states and actions, often through upper confidence bounds (UCBs) (Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018; 2020), whereas randomized methods, including Thompson sampling (TS) and posterior sampling, incorporate uncertainty in model or value estimates through randomized policy selection (Osband et al., 2013; 2016b; Osband and Van Roy, 2017; Osband et al., 2019; Russo, 2019; Zanette et al., 2020; Dann et al., 2021; Xiong et al., 2022; Agrawal et al., 2026).

Both approaches admit theoretical guarantees in tabular and function approximation settings. In deep RL, however, constructing and computing theoretically valid confidence bounds with neural network function approximators is generally difficult. This has motivated alternative approaches that induce exploration by introducing randomness into value function learning. One line of work randomizes value estimates using newly sampled perturbations in each episode (Zanette et al., 2020; Ishfaq et al., 2021), while another generates randomized value functions through iterative approximate posterior sampling (Ishfaq et al., 2024a;b). A separate line of work uses multiple value estimates collectively to construct optimistic values or policies, for example by taking maxima or quantiles across estimates (Liu et al., 2021; Tiapkin et al., 2023; Cassel et al., 2024; Wang et al., 2026; Cassel and Rosenberg, 2026). These methods provide theoretical foundations for a variety of exploration procedures, but their guarantees rely on regenerating perturbations, performing iterative approximate sampling, or comparing and aggregating multiple estimates.

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.

From a practical perspective, ensemble sampling (ES) provides a simple exploration mechanism. Bootstrapped DQN and its extensions with randomized prior functions (Osband et al., 2016a; 2018) maintain multiple value function models, uniformly select one member per episode, and follow its greedy policy throughout the episode. This mechanism enables temporally consistent exploration with neural network function approximation and has shown strong empirical performance in deep RL. Motivated by these practical advantages, recent work on linear bandits has also established regret guarantees for ES with finite ensembles, including rates comparable to TS (Lee and Oh, 2024; Janz et al., 2026; Hwang and Oh, 2026). However, it remains unclear whether the same persistent ensemble and uniform selection mechanism can guarantee efficient exploration in episodic RL with function approximation.

We address this question by proposing *Ensemble Sampling with Least-Squares Value Iteration* (**ES-LSVI**), which combines least-squares value iteration with persistently perturbed datasets. The main analytical challenge is that reused perturbations affect future data collection, so conditional arguments based on freshly sampled noise no longer apply directly. This dependence must also be controlled together with error propagation through Bellman recursion. We consider two variants of **ES-LSVI** that share the same uniform member selection mechanism. The *signed mode* uses the perturbations directly and achieves the same leading regret order as a TS-type method for linear MDPs (Zanette et al., 2020). The *absolute mode* uses the magnitude of the perturbation and achieves a tighter regret bound with a smaller ensemble, at the cost of an additional reference regression. Our main contributions are summarized as follows:

- We propose **ES-LSVI**, a provably efficient ensemble sampling framework for RL that maintains persistently perturbed datasets. In each episode, **ES-LSVI** uniformly selects one ensemble member, fits its value function, and executes its greedy policy. By reusing stored perturbations, **ES-LSVI** avoids regenerating historical noise or running an iterative posterior sampler, yielding a simple implementation amenable to scalable deep RL.
- We show that, in linear MDPs, the signed mode of **ES-LSVI** achieves a regret bound of $\widetilde{\mathcal{O}}(d^2H^2\sqrt{T})$ with $\widetilde{\mathcal{O}}(dH)$ ensemble members, where d is the feature dimension, H is the episode horizon, and $T = KH$ is the total number of interaction steps over K episodes. This matches the leading regret order of a TS-type method for linear MDPs (Zanette et al., 2020). To our knowledge, this is the first regret guarantee showing that uniform selection from a persistent ensemble can attain this rate in linear MDPs. Our analysis establishes stochastic optimism by jointly controlling the accumulated persistent perturbations across steps, despite their dependence on adaptively collected data.
- We further show that the absolute mode of **ES-LSVI** achieves a sharper regret bound with a smaller ensemble, at the cost of fitting an additional reference regression at each step. Specifically, it achieves $\widetilde{\mathcal{O}}(d^2H^{3/2}\sqrt{T})$ regret with only $\widetilde{\mathcal{O}}(d)$ ensemble members. By establishing optimism of the ensemble-averaged value functions at every step, our analysis reduces the sufficient ensemble size by a factor of H and improves the leading regret term by a factor of $\sqrt{H}$ relative to the signed mode, up to logarithmic factors.
- We empirically demonstrate that **ES-LSVI** achieves competitive performance against existing exploration baselines across diverse environments while retaining a simple implementation based on ensemble sampling.

2 PRELIMINARIES

Notation. For a positive integer n, let $[n] := \{1, \dots, n\}$. For vectors $x, y \in \mathbb{R}^d$, write $\langle x, y\rangle := x^\top y$ and $\|x\|_A := \sqrt{x^\top Ax}$ for $A \succeq 0_{d\times d}$. We use $\|\cdot\|_2$ and $\|\cdot\|_\infty$ for the Euclidean and supremum norms, respectively, and $\mathbb{1}\{\cdot\}$ for the indicator function. The notation $\mathcal{O}(\cdot)$ hides universal constants, while $\widetilde{\mathcal{O}}(\cdot)$ additionally suppresses logarithmic factors.

Markov decision process. We consider a finite-horizon Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and H is the episode horizon. The transition kernels $P = \{P_h\}_{h=1}^H$ and reward functions $r = \{r_h\}_{h=1}^H$ may vary across steps within an episode but remain fixed across episodes. For each $h \in [H]$, $P_h(\cdot \mid s, a)$ specifies the distribution of the next state, and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic reward function. A Markov policy

$\pi = \{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}_{h=1}^{H}$ specifies a distribution over actions at each state and step, where $\Delta(\mathcal{A})$ denotes the set of probability distributions over $\mathcal{A}$.

We evaluate the performance of a policy π using its value function, $V_h^\pi(s) := \mathbb{E}_{\pi,P}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right]$. For any bounded function $V : \mathcal{S} \to \mathbb{R}$, write $(P_h V)(s,a) := \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V(s')]$. The corresponding action-value function is $Q_h^\pi(s,a) := r_h(s,a) + (P_h V_{h+1}^\pi)(s,a)$, with $V_{H+1}^\pi \equiv 0$, and $V_h^\pi(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}[Q_h^\pi(s,a)]$. Let π^* be an optimal policy, with $V_h^* := V_h^{\pi^*}$ and $Q_h^* := Q_h^{\pi^*}$, so that $V_h^*(s) = \max_{a\in\mathcal{A}} Q_h^*(s,a)$.

The agent interacts with the MDP for K episodes, for a total of $T := KH$ interaction steps. At the beginning of episode $k \in [K]$, the agent observes an initial state $s_{k,1}$ and selects a policy $\pi_k = \{\pi_{k,h}\}_{h=1}^{H}$ based on this state, past observations, and its internal randomization. At each step $h \in [H]$, it samples an action $a_{k,h} \sim \pi_{k,h}(\cdot \mid s_{k,h})$, receives reward $r_h(s_{k,h}, a_{k,h})$, and observes the next state $s_{k,h+1} \sim P_h(\cdot \mid s_{k,h}, a_{k,h})$. The agent aims to minimize the cumulative regret, $\mathrm{Regret}(K) := \sum_{k=1}^{K} \left[V_1^*(s_{k,1}) - V_1^{\pi_k}(s_{k,1})\right]$.

3 ALGORITHMS

We introduce *ensemble sampling with least-squares value iteration* (**`ES-LSVI`**), summarized in Algorithm 1. The algorithm maintains an ensemble of m action-value models, each associated with a persistently perturbed dataset and an initial perturbation. At the beginning of each episode, it uniformly samples one member, fits that member through backward value iteration, and follows its greedy policy throughout the episode. We consider two modes, signed and $\mathrm{absolute}$, which differ in how the perturbed predictions are used to construct action-value estimates.

Persistent perturbations. For each step $h \in [H]$, we maintain a reference dataset $\mathcal{D}_h^0$ containing the observed transitions and a perturbed dataset $\mathcal{D}_h^j$ for each ensemble member $j \in [m]$. All datasets are initially empty. When a transition $(s_{k,h}, a_{k,h}, r_{k,h}, s_{k,h+1})$ is collected, where $r_{k,h} := r_h(s_{k,h}, a_{k,h})$, it is added to $\mathcal{D}_h^0$. For each member j, we independently sample $\xi_{k,h}^j \sim \mathcal{P}_R$ and add a copy of the transition with reward $r_{k,h} + \xi_{k,h}^j$ to $\mathcal{D}_h^j$. Each reward perturbation is sampled once upon collection and retained with its transition. At initialization, we also independently draw $g_h^j \sim \mathcal{P}_I$ for every $(h,j) \in [H] \times [m]$ and keep these vectors fixed across episodes.

Least-squares value iteration. At the beginning of episode k, we sample $J_k \sim \mathrm{Unif}([m])$ and perform backward value iteration only for the selected member. We initialize $V_{k,H+1}^{J_k} \equiv 0$. At each step $h = H, \ldots, 1$, the perturbed estimate $\widetilde{w}_{k,h}^{J_k}$ is obtained by fitting $Q(w; \phi(s,a))$ to the targets $\widetilde{r} + V_{k,h+1}^{J_k}(s')$ in $\mathcal{D}_h^{J_k}$, with regularization $\lambda\|w - g_h^{J_k}\|_2^2$. In the absolute mode, we additionally compute a reference estimate $\hat{w}_{k,h}^{J_k}$ using the targets $r + V_{k,h+1}^{J_k}(s')$ in $\mathcal{D}_h^0$ and regularization $\lambda\|w\|_2^2$. Both regressions use the same value estimate $V_{k,h+1}^{J_k}$; their objectives differ only through the stored reward perturbations and the initial perturbation $g_h^{J_k}$.

Signed and absolute estimates. At each backward step, the raw action-value estimate for the selected member is constructed as

$$\widetilde{Q}_{k,h}^{J_k}(s,a) := \begin{cases} Q(\widetilde{w}_{k,h}^{J_k}; \phi(s,a)), & \mathsf{M} = \mathrm{signed}, \\ Q(\hat{w}_{k,h}^{J_k}; \phi(s,a)) + \left|Q(\widetilde{w}_{k,h}^{J_k}; \phi(s,a)) - Q(\hat{w}_{k,h}^{J_k}; \phi(s,a))\right|, & \mathsf{M} = \mathrm{absolute}. \end{cases}$$

The signed mode directly uses the perturbed prediction. The absolute mode adds the magnitude of its deviation from the reference prediction, producing an upward adjustment to that reference. We then apply the value-bounding operator bdd and compute the corresponding value function:

$$Q_{k,h}^{J_k} := \mathrm{bdd}[\widetilde{Q}_{k,h}^{J_k}], \qquad V_{k,h}^{J_k}(s) := \max_{a\in\mathcal{A}} Q_{k,h}^{J_k}(s,a).$$

The operator bdd can take different forms to control the range of the value estimates, with clipping being one possible choice.

Algorithm 1 Ensemble Sampling with Least-Squares Value Iteration (**ES−LSVI**)

Require: Number of episodes K, mode $\mathsf{M} \in \{\text{signed}, \text{absolute}\}$, ensemble size m, regularization $\lambda > 0$, initial-perturbation distribution $\mathcal{P}_I$ on $\mathbb{R}^d$, reward-perturbation distribution $\mathcal{P}_R$ on $\mathbb{R}$, value-bounding operator bdd

Initialize $\mathcal{D}_h^j \leftarrow \emptyset$ for all $h \in [H]$ and $j \in \{0, \dots, m\}$
Independently sample $g_h^j \sim \mathcal{P}_I$ for all $(h, j) \in [H] \times [m]$
for each episode $k = 1, \dots, K$ **do**
  Observe $s_{k,1}$
  Sample $J_k \sim \mathrm{Unif}([m])$ and set $V_{k,H+1}^{J_k}(\cdot) \leftarrow 0$
  for each step $h = H, H-1, \dots, 1$ **do**

$$\widetilde{w}_{k,h}^{J_k} \in \operatorname*{argmin}_{w \in \mathbb{R}^d} \left\{ \sum_{(s,a,\widetilde{r},s') \in \mathcal{D}_h^{J_k}} \left(\widetilde{r} + V_{k,h+1}^{J_k}(s') - Q(w; \phi(s,a)) \right)^2 + \lambda \left\| w - g_h^{J_k} \right\|_2^2 \right\}$$

    if $\mathsf{M} = \text{signed}$ **then** ▷ Signed mode
      $\widetilde{Q}_{k,h}^{J_k}(\cdot,\cdot) \leftarrow Q(\widetilde{w}_{k,h}^{J_k}; \phi(\cdot,\cdot))$
    else if $\mathsf{M} = \text{absolute}$ **then** ▷ Absolute mode

$$\hat{w}_{k,h}^{J_k} \in \operatorname*{argmin}_{w \in \mathbb{R}^d} \left\{ \sum_{(s,a,r,s') \in \mathcal{D}_h^0} \left(r + V_{k,h+1}^{J_k}(s') - Q(w; \phi(s,a)) \right)^2 + \lambda \|w\|_2^2 \right\}$$

$$\widetilde{Q}_{k,h}^{J_k}(\cdot,\cdot) \leftarrow Q(\hat{w}_{k,h}^{J_k}; \phi(\cdot,\cdot)) + \left| Q(\widetilde{w}_{k,h}^{J_k}; \phi(\cdot,\cdot)) - Q(\hat{w}_{k,h}^{J_k}; \phi(\cdot,\cdot)) \right|$$

    end if
    $Q_{k,h}^{J_k}(\cdot,\cdot) \leftarrow \mathrm{bdd}[\widetilde{Q}_{k,h}^{J_k}(\cdot,\cdot)], \qquad V_{k,h}^{J_k}(\cdot) \leftarrow \max_{a \in \mathcal{A}} Q_{k,h}^{J_k}(\cdot, a)$
  end for
  for each step $h = 1, \dots, H$ **do**
    Take $a_{k,h} \in \operatorname{argmax}_{a \in \mathcal{A}} Q_{k,h}^{J_k}(s_{k,h}, a)$, and observe $r_{k,h} := r_h(s_{k,h}, a_{k,h})$ and $s_{k,h+1}$
    $\mathcal{D}_h^0 \leftarrow \mathcal{D}_h^0 \cup \{(s_{k,h}, a_{k,h}, r_{k,h}, s_{k,h+1})\}$
    for each ensemble member $j = 1, \dots, m$ **do**
      Sample $\xi_{k,h}^j \sim \mathcal{P}_R$ and $\mathcal{D}_h^j \leftarrow \mathcal{D}_h^j \cup \{(s_{k,h}, a_{k,h}, r_{k,h} + \xi_{k,h}^j, s_{k,h+1})\}$
    end for
  end for
end for

Once backward value iteration is complete, the agent interacts with the environment using the selected member's greedy policy. At each step h, it selects $a_{k,h} \in \operatorname{argmax}_{a \in \mathcal{A}} Q_{k,h}^{J_k}(s_{k,h}, a)$ and observes the reward and next state. Each collected transition is added to the reference dataset and, with independently perturbed rewards, to the datasets of all ensemble members. Thus, all members' datasets accumulate the observations collected during the episode, while only the selected member's value estimates are computed for interaction.

Remark 1 (Computational efficiency)**.** ***ES−LSVI*** *maintains the ensemble through persistent perturbed datasets and computes value estimates only for the member sampled at the beginning of each episode. Since each member's backward recursion uses only its own value estimates at the next step, the selected member can be fitted without computing those of the other members. Consequently, the number of regression problems solved per episode is independent of the ensemble size: one per step in the signed mode and two per step in the absolute mode. All members' datasets are still extended with each newly collected transition and independently perturbed rewards. This distinguishes* ***ES−LSVI*** *from ensemble methods that select the member with the highest estimated initial value (Cassel et al., 2024), aggregate predictions through quantiles (Liu et al., 2021; Cassel and Rosenberg, 2026), or construct uncertainty bonuses from their standard deviation (Bai et al., 2021). Such methods require predictions from multiple members for policy selection or value updates, whereas uniform member selection allows* ***ES−LSVI*** *to perform backward value iteration for only one member per episode.*

Remark 2 (Signed versus absolute modes)**.** *The two modes offer a trade-off between computational cost and regret guarantees. The signed mode uses only the perturbed regression estimate, whereas*

the absolute mode requires an additional reference regression at each step. As shown in Theorem 1, the absolute mode achieves a tighter regret bound at the cost of this additional computation.

4 THEORETICAL RESULTS

The proposed ensemble sampling framework accommodates different parameterizations of the action-value function. We study its theoretical guarantees in linear MDPs, a standard setting for analyzing exploration that enables comparison with the methods summarized in Table 1.

4.1 **ES-LSVI** FOR LINEAR MDPS

Linear MDPs. We consider linear MDPs (Jin et al., 2020), in which the reward functions and transition kernels share a linear representation.

Definition 1 (Linear MDP)**.** *An MDP is linear with respect to a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if, for each $h \in [H]$, there exist an unknown vector $\theta_h \in \mathbb{R}^d$ and a vector of unknown finite signed measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})^\top$ over $\mathcal{S}$ such that, for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$r_h(s, a) = \langle \phi(s, a), \theta_h \rangle, \qquad P_h(\cdot \mid s, a) = \langle \phi(s, a), \mu_h(\cdot) \rangle.$$

Without loss of generality, we assume that $\|\phi(s, a)\|_2 \le 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Moreover, for every $h \in [H]$ and bounded measurable function $V : \mathcal{S} \to \mathbb{R}$, we assume that $\|\theta_h\|_2 \le \sqrt{d}$ and $\left\| \int_{\mathcal{S}} V(s')\, \mu_h(ds') \right\|_2 \le \sqrt{d}\, \|V\|_\infty$.

Least-squares updates. The linear MDP structure implies that the action-value function of any Markov policy π admits the following linear representation (Jin et al., 2020, Proposition 2.3):

$$Q_h^\pi(s, a) = \langle \phi(s, a), w_h^\pi \rangle, \qquad w_h^\pi := \theta_h + \int_{\mathcal{S}} V_{h+1}^\pi(s')\, \mu_h(ds').$$

Accordingly, we use $Q(w; \phi(s, a)) = \langle \phi(s, a), w \rangle$ in Algorithm 1. For analysis, we define the regression and value estimates for every member $j \in [m]$ by applying the same backward recursion as in Algorithm 1, although the algorithm computes them only for the selected member $j = J_k$. Let $\phi_{i,h} := \phi(s_{i,h}, a_{i,h})$. At the beginning of episode k, define the regularized Gram matrix and the perturbation vector by

$$\Lambda_{k,h} := \lambda I_d + \sum_{i=1}^{k-1} \phi_{i,h} \phi_{i,h}^\top, \qquad s_{k,h}^j := \lambda g_h^j + \sum_{i=1}^{k-1} \phi_{i,h} \xi_{i,h}^j.$$

The matrix $\Lambda_{k,h}$ is shared across ensemble members, while $s_{k,h}^j$ records the persistent perturbations of member j. Because the reference and perturbed regressions use the same value estimate $V_{k,h+1}^j$ at the next step, their normal equations give

$$\hat{w}_{k,h}^j = \Lambda_{k,h}^{-1} \sum_{i=1}^{k-1} \phi_{i,h} \left[r_{i,h} + V_{k,h+1}^j(s_{i,h+1}) \right], \qquad \widetilde{w}_{k,h}^j = \hat{w}_{k,h}^j + \Lambda_{k,h}^{-1} s_{k,h}^j.$$

Here, $\hat{w}_{k,h}^j$ denotes the reference estimate in both modes; the signed mode can compute $\widetilde{w}_{k,h}^j$ directly without separately fitting $\hat{w}_{k,h}^j$. The raw action-value estimates therefore take the form

$$\widetilde{Q}_{k,h}^j(s, a) = \langle \phi(s, a), \hat{w}_{k,h}^j \rangle + \begin{cases} \phi(s, a)^\top \Lambda_{k,h}^{-1} s_{k,h}^j, & \mathsf{M} = \text{signed}, \\ \left| \phi(s, a)^\top \Lambda_{k,h}^{-1} s_{k,h}^j \right|, & \mathsf{M} = \text{absolute}. \end{cases}$$

Thus, the signed mode preserves the sign of the projected perturbation, while the absolute mode adds its magnitude to the reference prediction.

Value-bounding operator. To control the propagation of regression and perturbation errors through backward value iteration, we use the truncation scheme following prior randomized value-iteration methods (Zanette et al., 2020; Wu and Sun, 2024). For thresholds $0 < \alpha_L < \alpha_U$, define

$$\nu(u) := \mathbb{1}\{u \le \alpha_L\} + \frac{\alpha_U - u}{\alpha_U - \alpha_L}\mathbb{1}\{\alpha_L < u < \alpha_U\}.$$

For each ensemble member $j \in [m]$, we define

$$\mathrm{bdd}[\widetilde{Q}^j_{k,h}](s,a) := \nu\Big(\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}\Big)\,\widetilde{Q}^j_{k,h}(s,a) + \Big[1 - \nu\Big(\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}\Big)\Big]\,(H-h+1). \quad (1)$$

The operator retains the raw estimate when uncertainty is low, assigns the optimistic default $H-h+1$ when uncertainty is high, and interpolates continuously between these regimes. The interpolation weight depends only on the Gram matrix, which is shared across all ensemble members. Consequently, the operator applies the same affine transformation to every member's raw estimate at each state-action pair, a property used in our analysis.

4.2 Regret Bounds

We establish regret bounds for the signed and absolute modes of **`ES-LSVI`** in linear MDPs.

Theorem 1 (Regret bounds for **`ES-LSVI`**). *Consider a linear MDP as defined in Definition 1, and let $T := KH$ and $\delta \in (0, 1/2)$. Run Algorithm 1 with the linear parameterization, the value-bounding operator defined in Eq. (1), $\lambda = 1$, and Gaussian perturbation distributions $\mathcal{P}_I = \mathcal{N}(0, \eta^2 I_d)$ and $\mathcal{P}_R = \mathcal{N}(0, \eta^2)$. For each mode, there exist choices of m, η, and thresholds $0 < \alpha_L < \alpha_U$ such that the following guarantee holds with probability at least $1 - \delta$:*

1. ***Signed mode.*** *With $m = \widetilde{\mathcal{O}}(dH)$ and $\eta = \widetilde{\mathcal{O}}(dH^{3/2})$, $\mathrm{Regret}(K) \le \widetilde{\mathcal{O}}\Big(d^2H^2\sqrt{T} + d^4H^5\Big)$.*
2. ***Absolute mode.*** *With $m = \widetilde{\mathcal{O}}(d)$ and $\eta = \widetilde{\mathcal{O}}(dH)$, $\mathrm{Regret}(K) \le \widetilde{\mathcal{O}}\Big(d^2H^{3/2}\sqrt{T} + d^4H^4\Big)$.*

Discussion of Theorem 1. Theorem 1 shows that a finite ensemble with persistent perturbations suffices for provably efficient exploration in linear MDPs. For fixed d and H, both modes achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret with ensemble sizes that depend only logarithmically on the number of episodes. In particular, the signed mode matches the regret bound of the TS-type method `OPT-RLSVI` for linear MDPs (Zanette et al., 2020), while retaining perturbations across episodes. In contrast, `OPT-RLSVI` draws fresh Gaussian perturbations in every episode. To our knowledge, this is the first ensemble sampling guarantee matching this regret rate in linear MDPs with persistent perturbations.

The absolute mode reduces the sufficient ensemble size by a factor of H and improves the leading regret term by a factor of $\sqrt{H}$, up to logarithmic factors. The two analyses rely on different optimism guarantees: the signed analysis establishes that a constant fraction of ensemble members are optimistic at the initial state, whereas the absolute analysis establishes optimism of the ensemble-averaged value functions at every step. This per-step guarantee permits a smaller exploration scale. The absolute mode requires an additional reference regression at each step, yielding a trade-off between computation per member and the strength of the regret guarantee.

Compared with existing randomized exploration approaches, **`ES-LSVI`** retains both initial and reward perturbations across episodes. The perturbed-history exploration method of Ishfaq et al. (2021) redraws noise for the accumulated history, so the number of reward perturbations per fit at each step grows linearly with the episode index k. The bootstrapped value iteration method of Liu et al. (2021) has a sufficient ensemble size of $\mathcal{O}(d^6T^4k)$ in its analysis, which grows polynomially with T. Posterior sampling via approximate samplers (Ishfaq et al., 2024b) requires iterative updates, with sufficient cumulative iteration budgets in the analysis scaling superlinearly in both K and H.

4.3 Regret Analysis

In this section, we outline the key technical steps underlying Theorem 1. For convenience, let $w_h(V) := \theta_h + \int_{\mathcal{S}} V(s')\,\mu_h(ds')$ denote the population regression parameter for a value function V. Let $\mathcal{F}_k$ denote the information available immediately before sampling the ensemble index J_k, including the initial state, past observations, and all persistent perturbations generated so far.

4.3.1 STOCHASTIC OPTIMISM AND REGRET FOR THE SIGNED MODE

Step 1. Concentration of Bellman regression estimates. We first establish concentration for the Bellman regression estimates $\hat{w}^j_{k,h}$. The regression targets contain value estimates $V^j_{k,h+1}$ constructed from the same data used in the regression. We follow the standard uniform-concentration and backward-induction arguments developed for linear MDPs (Jin et al., 2020; Zanette et al., 2020). Applying these arguments requires controlling the ensemble perturbations, since those at step $h+1$ affect $V^j_{k,h+1}$ and hence the regression targets at step h. By uniformly controlling the perturbations over episodes, steps, and ensemble members, we obtain the following concentration bounds.

Lemma 1 (Bellman regression concentration; informal)**.** *There are confidence radii $\beta = \widetilde{\mathcal{O}}(dH)$ and $\Gamma_\delta = \widetilde{\mathcal{O}}(\sqrt{d})$ such that, with high probability, the following holds*

$$\forall k \in [K],\, h \in [H],\, j \in [m]: \quad \left\|\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\right\|_{\Lambda_{k,h}} \le \beta, \qquad \left\|\Lambda^{-1}_{k,h} s^j_{k,h}\right\|_{\Lambda_{k,h}} \le \eta\Gamma_\delta.$$

Combining Lemma 1 with the value-bounding operator and choosing the thresholds such that $\alpha_U(\beta + \eta\Gamma_\delta) \le 1$ yields bounded value estimates and the following Bellman residual bound:

$$\begin{aligned}\left|\Delta^j_{k,h}(s,a)\right| &:= \left|Q^j_{k,h}(s,a) - r_h(s,a) - (P_h V^j_{k,h+1})(s,a)\right| \\ &\le (\beta + \eta\Gamma_\delta)\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} + 2H\mathbb{1}\{\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} > \alpha_L\}\,.\end{aligned}$$

Step 2. Stochastic optimism. A key ingredient in randomized exploration is stochastic optimism: the estimated value upper-bounds the optimal value with sufficient frequency. A common approach in linear MDPs is to construct optimistic Bellman backups uniformly over state–action pairs and propagate optimism by backward induction. For example, the method of Ishfaq et al. (2021) takes the maximum of multiple perturbed estimates to achieve optimism with high probability, while the method of Liu et al. (2021) uses quantiles of posterior predictions together with extrapolation to establish optimism of its mean bootstrapped value functions. These approaches obtain optimism by aggregating multiple randomized predictions at each step.

In **`ES-LSVI`**, each ensemble member maintains a persistent accumulated perturbation vector at each step, and one member is sampled uniformly and used throughout the episode. Thus, analyses that establish per-step optimism by selecting or aggregating multiple perturbed estimates do not directly apply. Instead, following Zanette et al. (2020), we track how the perturbations accumulate along an optimal policy and affect the value estimate at the initial state. Combining the Bellman equations with Lemma 1, we obtain

$$V^j_{k,1}(s_{k,1}) - V^*_1(s_{k,1}) \ge \sum_{h=1}^{H} \langle \phi^*_{k,h}, \Lambda^{-1}_{k,h} s^j_{k,h}\rangle - \beta \sum_{h=1}^{H} \|\phi^*_{k,h}\|_{\Lambda^{-1}_{k,h}}. \tag{2}$$

Here, $\phi^*_{k,h} \in \mathbb{R}^d$ is a weighted average of features along trajectories under an optimal policy, with weights induced by the value-bounding operator. Given the initial state, these directions are determined by the history and are therefore fixed conditional on $\mathcal{F}_k$ and common to all ensemble members. The first term in Eq. (2) represents the cumulative perturbation, while the second bounds the accumulated regression error. Thus, optimism at the initial state holds whenever the cumulative perturbation offsets this error, without requiring every Bellman backup to be optimistic.

In TS-type methods such as `OPT-RLSVI` (Zanette et al., 2020), Gaussian pseudo-noise vectors are freshly sampled in each episode, allowing Gaussian anti-concentration to be applied conditional on the collected history. In contrast, **`ES-LSVI`** reuses persistent perturbations that have already influenced the collected data and hence the feature directions. Conditional on $\mathcal{F}_k$, the remaining randomness therefore comes from uniformly selecting J_k among fixed ensemble members. We accordingly show that a constant fraction of these members satisfy the cumulative optimism condition.

Lemma 2 (Stochastic optimism; informal)**.** *There are choices $m = \widetilde{\mathcal{O}}(dH)$ and $\eta = \Theta(\beta\sqrt{H})$ such that, with high probability, every episode satisfies*

$$\mathbb{P}\Big(V^{J_k}_{k,1}(s_{k,1}) \ge V^*_1(s_{k,1}) \,\Big|\, \mathcal{F}_k\Big) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\Big\{V^j_{k,1}(s_{k,1}) \ge V^*_1(s_{k,1})\Big\} \ge p_0,$$

where $p_0 > 0$ is a universal constant.

To establish this result, we stack each member's perturbation vectors across the H steps into a single vector in $\mathbb{R}^{dH}$. For any fixed direction $u = (u_1, \ldots, u_H) \in \mathbb{R}^{dH}$, the projected perturbations $(\sum_{h=1}^{H} \langle u_h, s_{k,h}^j \rangle)_{j=1}^m$ form a predictable diagonal martingale transform of Gaussian noise, since the coefficient at each update is determined before the corresponding reward perturbations are generated. We then apply the uniform self-normalized exceedance result of Janz et al. (2026), which guarantees that a constant fraction of the persistent ensemble exceeds a normalized threshold uniformly over episodes and directions. With a suitable ensemble size, this yields a constant lower bound on the self-normalized exceedance frequency uniformly over episodes and directions:

$$\min_{k\in[K]} \inf_{u=(u_1,\ldots,u_H)\in\mathbb{R}^{dH}\setminus\{\mathbf{0}_{dH}\}} \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left\{ \sum_{h=1}^{H} \langle u_h, s_{k,h}^j \rangle \geq c\eta \left(\sum_{h=1}^{H} \|u_h\|_{\Lambda_{k,h}}^2 \right)^{1/2} \right\} \geq p_0,$$

where $c, p_0 > 0$ are universal constants. Since this bound holds uniformly over directions, we may choose $u_h = \Lambda_{k,h}^{-1} \phi_{k,h}^*$ despite its dependence on the collected history. Under this choice, the projected perturbation coincides with the cumulative perturbation in Eq. (2). Cauchy–Schwarz introduces a factor of $\sqrt{H}$ when relating the accumulated regression error to the self-normalized perturbation. Thus, $\eta = \Theta(\beta\sqrt{H})$ suffices to make a constant fraction of members optimistic.

Step 3. Conditional expected regret. A constant optimistic fraction does not imply that the selected member is optimistic in every episode. Nevertheless, it suffices to control the conditional expected regret under uniform ensemble sampling. Let π_k^j denote the greedy policy of $Q_{k,h}^j$.

Lemma 3 (From stochastic optimism to regret)**.** *On the events of Lemmas 1 and 2,*

$$\mathbb{E}\left[V_1^*(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1}) \,\middle|\, \mathcal{F}_k \right] \leq \left(1 + \frac{2}{p_0}\right) \mathbb{E}\left[\sum_{h=1}^{H} \max_{j\in[m]} \left| \Delta_{k,h}^j(s_{k,h}, a_{k,h}) \right| \,\middle|\, \mathcal{F}_k \right].$$

Unlike the analysis of `OPT-RLSVI` (Zanette et al., 2020), our argument requires neither independent copies of the pseudo-noise nor an auxiliary minimization over admissible perturbations. Instead, Lemma 3 follows from Bellman telescoping that directly controls pairwise differences between ensemble value estimates. Combining this lemma with the Bellman residual bound above, martingale concentration, and the elliptical potential lemma along the agent's realized trajectories yields the desired high-probability regret bound. Detailed proofs are provided in the appendix.

4.3.2 Ensemble-average optimism for the absolute mode

The signed analysis establishes optimism by controlling each member's cumulative perturbation across steps. The absolute mode instead establishes optimism of the ensemble average: the absolute perturbation terms offset regression errors after averaging at each step, without requiring individual members to be optimistic.

Lemma 4 (Ensemble-average optimism; informal)**.** *There exist choices $m = \widetilde{\mathcal{O}}(d)$ and $\eta = \Theta(\beta)$ such that, with high probability, for all $k \in [K]$, $h \in [H]$, and $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$\frac{1}{m}\sum_{j=1}^{m} Q_{k,h}^j(s,a) \geq Q_h^*(s,a), \qquad \frac{1}{m}\sum_{j=1}^{m} V_{k,h}^j(s) \geq V_h^*(s).$$

The key is to establish a Bellman inequality for ensemble averages despite the different regression targets across members. At each step, the uniform exceedance guarantee for projections of $s_{k,h}^j$, together with the absolute perturbation terms, yields

$$\frac{1}{m}\sum_{j=1}^{m} \left| \phi(s,a)^\top \Lambda_{k,h}^{-1} s_{k,h}^j \right| \geq c\eta \|\phi(s,a)\|_{\Lambda_{k,h}^{-1}},$$

where $c > 0$ is a universal constant. For $c\eta \geq \beta$, this offsets the regression error after averaging. Assuming ensemble-average optimism at step $h+1$, the regression bound and the linearity and monotonicity of the transition operator P_h then give

$$\frac{1}{m}\sum_{j=1}^{m} \widetilde{Q}_{k,h}^j(s,a) \geq r_h(s,a) + P_h\left(\frac{1}{m}\sum_{j=1}^{m} V_{k,h+1}^j\right)(s,a) \geq r_h(s,a) + P_h V_{h+1}^\star(s,a) = Q_h^\star(s,a)\,.$$

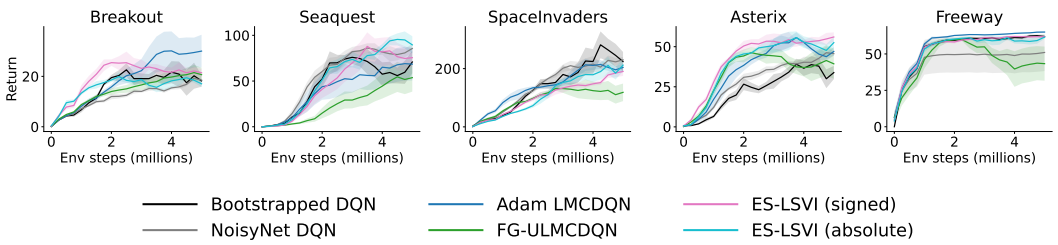


Figure 1: Learning curves on the five MinAtar games over five million environment steps. Solid lines denote the mean return over five random seeds, and shaded regions indicate the standard error.

Starting from the zero terminal values, the result follows by backward induction. Unlike the signed analysis, this argument closes after averaging at each step and therefore requires only separate d-dimensional perturbation guarantees, rather than joint control across all H steps. This yields a smaller ensemble size and exploration scale, leading to the improved regret bound.

5 Experiments

We evaluate neural implementations of **ES-LSVI** on all five MinAtar games: Breakout, Seaquest, Space Invaders, Asterix, and Freeway (Young and Tian, 2019), using five million environment steps and five random seeds. We compare the signed and absolute variants of **ES-LSVI** with Bootstrapped DQN with randomized prior functions (Osband et al., 2016a; 2018), NoisyNet DQN (Fortunato et al., 2018), Adam LMCDQN (Ishfaq et al., 2024a), and FG-ULMCDQN (Ishfaq et al., 2024b). For each method, we first perform a grid search over hyperparameters and use the selected configurations for the final comparison.

Figure 1 shows that **ES-LSVI** achieves competitive learning performance across multiple environments, while the relative performance of the signed and absolute variants varies across tasks. In particular, the signed variant performs strongly on Asterix, whereas the absolute variant performs strongly on Seaquest. This suggests that, despite the stronger regret guarantee of the absolute variant in theory, the relative effectiveness of the two perturbation schemes can depend on the environment in neural implementations. Figure 2 compares the wall-clock cost of the methods on Asterix. **ES-LSVI** achieves competitive performance with comparable or lower wall-clock cost than the evaluated posterior sampling methods. Overall, these results indicate that **ES-LSVI** can be effectively applied with nonlinear value function approximation while retaining practical computational efficiency. Additional experimental results and details are provided in Appendix D.

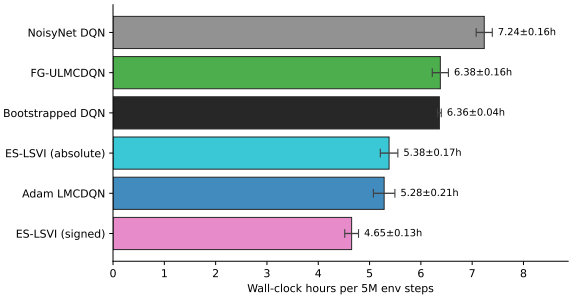


Figure 2: Wall-clock training time on Asterix. Results are measured on the same hardware and reported as the mean and standard deviation over five random seeds.

6 Conclusion

We proposed **ES-LSVI**, a practical ensemble sampling framework with a simple implementation amenable to scalable RL. It retains perturbations across episodes and performs planning only for a uniformly selected member. In linear MDPs, the signed mode matches the regret rate of a provably efficient TS-based method, while the absolute mode achieves an improved rate with a smaller ensemble through an ensemble-average optimism analysis. To our knowledge, these are the first regret guarantees for ensemble sampling with persistent perturbations and uniform member selection in linear MDPs. Experiments demonstrate competitive performance across diverse environments, supporting the practical effectiveness of the framework. Future work will extend these guarantees to more general and nonlinear function classes.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with the writing, refinement, and verification of mathematical proofs, and providing feedback on experimental design and the interpretation of experimental results. The remaining required-disclosure tasks were not applicable to this work. Additionally, we used generative AI tools for improving the readability of the manuscript and providing feedback on the organization and presentation of the paper. We reviewed all AI-assisted mathematical arguments, experimental interpretations, and literature references. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Priyank Agrawal, Jinglin Chen, and Nan Jiang. Improved worst-case regret bounds for randomized least-squares value iteration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6566–6573, 2021.

Priyank Agrawal, Shipra Agrawal, and Azmat Azati. Q-learning with posterior sampling. In *International Conference on Learning Representations*, volume 2026, pages 97730–97760, 2026.

Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Chenjia Bai, Lingxiao Wang, Lei Han, Jianye Hao, Animesh Garg, Peng Liu, and Zhaoran Wang. Principled exploration via optimistic bootstrapping and backward induction. In *International Conference on Machine Learning*, pages 577–587. PMLR, 2021.

Asaf Cassel and Aviv Rosenberg. Quantile of means: A bonus-free ensemble method for minimax optimal reinforcement learning. *arXiv preprint arXiv:2606.20107*, 2026.

Asaf Cassel, Haipeng Luo, Aviv Rosenberg, and Dmitry Sotnikov. Near-optimal regret in linear MDPs with aggregate bandit feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in neural information processing systems*, 34:12040–12051, 2021.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *International Conference on Learning Representations*, 2018.

Taehyun Hwang and Min-hwan Oh. Linear ensemble sampling with smaller ensembles. *arXiv preprint arXiv:2609.13954*, 2026.

Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, volume 139, pages 4607–4616. PMLR, 2021.

Haque Ishfaq, Qingfeng Lan, Pan Xu, A. Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. In *The Twelfth International Conference on Learning Representations*, 2024a.

Haque Ishfaq, Yixin Tan, Yu Yang, Qingfeng Lan, Jianfeng Lu, A. Rupam Mahmood, Doina Precup, and Pan Xu. More efficient randomized exploration for reinforcement learning via approximate sampling. In *Reinforcement Learning Conference*, 2024b.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

David Janz, Alexander Litvak, and Csaba Szepesvari. Ensemble sampling for linear bandits: small ensembles suffice. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

David Janz, Arya Akhavan, and Csaba Szepesvári. Sharp analysis of linear ensemble sampling. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 3716–3750. PMLR, 2026.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, pages 4868–4878, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Harin Lee and Min-hwan Oh. Improved regret of linear ensemble sampling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Boovi: Provably efficient bootstrapped value iteration. *Advances in Neural Information Processing Systems*, 34:7041–7053, 2021.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.

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

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

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016b.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Ian Osband, Benjamin Van Roy, Daniel J Russo, Zheng Wen, et al. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.

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

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Pierre Perrault, Michal Valko, and Pierre Ménard. Model-free posterior sampling via learning rate randomization. *Advances in Neural Information Processing Systems*, 36:73719–73774, 2023.

He Wang, Xingyu Xu, and Yuejie Chi. Provably efficient and agile randomized q-learning. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Runzhe Wu and Wen Sun. Making RL with preference-based feedback efficient via randomization. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhihan Xiong, Ruoqi Shen, Qiwen Cui, Maryam Fazel, and Simon S Du. Near-optimal randomized exploration for tabular markov decision processes. *Advances in neural information processing systems*, 35:6358–6371, 2022.

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

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

CONTENTS OF APPENDIX

A Related Work **14**

B Theoretical Guarantees for Linear MDPs **15**

B.1 Regret Bound for the Signed Mode 15

B.1.1 Bellman residual bounds for the signed mode 16

B.1.2 Stochastic optimism of signed ensemble sampling 18

B.1.3 Cumulative regret 19

B.2 Regret Bound for the Absolute Mode 21

B.2.1 Bellman residual bounds for the absolute mode 21

B.2.2 Optimism of ensemble-averaged value functions 21

B.2.3 Cumulative regret 22

B.3 Proof of Technical Lemmas 24

B.3.1 Proof of Lemma 5 24

B.3.2 Proof of Lemma 6 24

B.3.3 Proof of Lemma 7 29

B.3.4 Proof of Lemma 8 30

B.3.5 Proof of Lemma 3 33

B.3.6 Proof of Lemma 9 34

B.3.7 Proof of Lemma 10 36

C Auxiliary Lemmas **38**

D Experiment Details **39**

A Related Work

Table 1: Comparison of regret guarantees for provably efficient exploration algorithms in linear MDPs and ensemble-based algorithms in tabular MDPs. Logarithmic factors and selected lower-order terms are omitted. Here, d is the feature dimension, H is the episode horizon, $T = KH$ is the total number of interaction steps over K episodes, and S and A are the numbers of states and actions, respectively.

Algorithm	Exploration	Regret	Note
`LSVI-UCB` (Jin et al., 2020)	UCB	$\widetilde{\mathcal{O}}\left(d^{3/2}H^{3/2}\sqrt{T}\right)$	Explicit exploration bonus
`OPT-RLSVI` (Zanette et al., 2020)	TS	$\widetilde{\mathcal{O}}\left(d^{2}H^{2}\sqrt{T}\right)$	Noise resampling per episode
`LSVI-PHE` (Ishfaq et al., 2021)	PHE	$\widetilde{\mathcal{O}}\left(d^{3/2}H^{3/2}\sqrt{T}\right)$	Resampling noise for past rewards
`LMC-LSVI` (Ishfaq et al., 2024a)	LMC	$\widetilde{\mathcal{O}}\left(d^{3/2}H^{3/2}\sqrt{T}\right)$	Iterative Langevin sampling
`LSVI-ASE` (Ishfaq et al., 2024b)	FG-TS	$\widetilde{\mathcal{O}}\left(dH^{3/2}\sqrt{T}\right)$	Approximate posterior sampling
`BooVI` (Liu et al., 2021)	Bootstrap & quantile	$\widetilde{\mathcal{O}}\left(d^{3/2}H^{3/2}\sqrt{T}\right)$	Repeated posterior sampling
`RE-LSVI` (Cassel et al., 2024)	Max-over-ensemble	$\widetilde{\mathcal{O}}\left(d^{5/2}H^{3}\sqrt{T}\right)$	Aggregate feedback; noise resampling
`ES-LSVI` (Theorem 1)	ES (signed)	$\widetilde{\mathcal{O}}\left(d^{2}H^{2}\sqrt{T}\right)$	Persistent perturbations
`ES-LSVI` (Theorem 1)	ES (absolute)	$\widetilde{\mathcal{O}}\left(d^{2}H^{3/2}\sqrt{T}\right)$	Persistent perturbations
`Staged-RandQL` (Tiapkin et al., 2023)	Max-over-ensemble	$\widetilde{\mathcal{O}}\left(H^{2}\sqrt{SAT}\right)$	Tabular MDP
`RandomizedQ` (Wang et al., 2026)	Max-over-ensemble & mixing	$\widetilde{\mathcal{O}}\left(H^{2}\sqrt{SAT}\right)$	Tabular MDP
`VIBE` (Cassel and Rosenberg, 2026)	Data partitioning & quantile	$\widetilde{\mathcal{O}}\left(H\sqrt{SAT}\right)$	Tabular MDP

Randomized exploration in RL. Randomized exploration has been studied through both posterior sampling over MDP models and randomization of value functions. In tabular RL, model-based posterior sampling selects a policy by planning in an MDP sampled from the posterior (Osband et al., 2013; Osband and Van Roy, 2017; Agrawal and Jia, 2017). Model-free randomized methods instead induce exploration through sampled or perturbed value functions, including randomized least-squares value iteration (Osband et al., 2016b; 2019; Russo, 2019; Agrawal et al., 2021) and posterior-sampling methods over value functions (Dann et al., 2021; Agrawal et al., 2026).

In linear MDPs, Zanette et al. (2020) proposed `OPT-RLSVI`, a TS-based algorithm that draws fresh Gaussian perturbations in each episode and achieves a leading regret term of $\widetilde{\mathcal{O}}(d^2H^2\sqrt{T})$. Ishfaq et al. (2021) proposed `LSVI-PHE`, a perturbed-history exploration (PHE) algorithm that takes the maximum of multiple perturbed regression predictions and achieves $\widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T})$ regret. It resamples noise for all previously observed rewards in each episode, so the number of reward perturbations per fit grows linearly with the episode index k. For approximate posterior sampling, Ishfaq et al. (2024a) proposed `LMC-LSVI`, which uses Langevin Monte Carlo (LMC) updates and achieves $\widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T})$ regret. Ishfaq et al. (2024b) proposed `LSVI-ASE`, which combines Feel-Good Thompson sampling (FG-TS) with approximate samplers and establishes $\widetilde{\mathcal{O}}(dH^{3/2}\sqrt{T})$ expected regret under suitable sampling accuracy. In contrast, our method retains perturbations across episodes and does not require iterative posterior sampling.

Ensemble and bootstrap methods in RL. Ensemble methods use multiple value estimates in different ways. `Bootstrapped DQN` and its extensions with randomized prior functions (Osband et al., 2016a; 2018) select one member per episode and follow its greedy policy, demonstrating the practical promise of temporally consistent exploration. Building on this structure, **`ES-LSVI`** combines uniform member selection with persistent perturbations and least-squares value iteration, for which we establish regret guarantees in linear MDPs. Other methods instead use multiple estimates collectively to construct or select optimistic predictions.

In linear MDPs, Liu et al. (2021) proposed `BooVI`, which combines empirical quantiles of posterior predictions with extrapolation and achieves $\widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T})$ regret. Its analysis specifies a sufficient budget of $N_k = \mathcal{O}(d^6T^4k/\delta^4)$ posterior samples per step in episode k, which grows polynomially with T and linearly with k. Under aggregate reward feedback, Cassel et al. (2024) proposed `RE-LSVI`, which combines randomized reward estimates with explicit transition bonuses

and selects the ensemble member with the highest initial policy value. It achieves $\widetilde{\mathcal{O}}(d^{5/2}H^3\sqrt{T})$ regret using an ensemble of logarithmic size that is regenerated in each episode.

In tabular MDPs, Tiapkin et al. (2023) proposed `Staged-RandQL`, which updates an ensemble using randomized learning rates and takes its maximum at stage boundaries. Wang et al. (2026) proposed `RandomizedQ`, which combines a frequently updated ensemble maximum with a staged optimistic estimate. Both methods achieve leading regret bounds of $\widetilde{\mathcal{O}}(H^2\sqrt{SAT})$. Cassel and Rosenberg (2026) proposed `VIBE`, which partitions observations into disjoint batches and uses a quantile of their value estimates, achieving a leading worst-case regret bound of $\widetilde{\mathcal{O}}(H\sqrt{SAT})$. These tabular guarantees require only logarithmic ensemble sizes. The distinction from our method lies in *how the ensemble is used*: these methods compare or aggregate predictions across members, whereas **ES-LSVI** retains persistent perturbations and follows one uniformly selected member without comparing or aggregating ensemble predictions.

Ensemble sampling in bandits. Recent analyses of linear ensemble sampling have progressively improved both regret guarantees and sufficient ensemble sizes. Janz et al. (2024) establish $\widetilde{\mathcal{O}}(d^{5/2}\sqrt{T})$ frequentist regret for a symmetrized ES algorithm with $m = \Theta(d \log T)$ members, allowing infinite action sets, where d is the feature dimension and T is the number of interaction rounds. Lee and Oh (2024) improve the regret to $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ without symmetrization, matching the standard linear TS benchmark, but their sufficient ensemble size scales linearly with the number of arms and logarithmically with T. Janz et al. (2026) obtain the same regret order with $m = \Theta(d \log T)$ for potentially infinite action sets. Their analysis embeds the persistent Gaussian perturbation processes into Brownian motions and controls the fraction of members exceeding a threshold uniformly over time and directions. More recently, Hwang and Oh (2026) introduce occasional ensemble refreshes based on changes in the Gram matrix, reducing the sufficient ensemble size to $\mathcal{O}(d \log d + d \log\log T)$ while retaining $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret.

Our work builds on the analysis of persistent ensembles but addresses additional challenges in RL: ensemble-dependent Bellman targets and the propagation of estimation errors across steps. We establish regret guarantees through stochastic optimism based on cumulative perturbations for the signed mode and ensemble-average optimism for the absolute mode.

B Theoretical Guarantees for Linear MDPs

In this section, we analyze the regret of **ES-LSVI** in linear MDPs. Algorithm 2 gives the corresponding linear specialization.

B.1 Regret Bound for the Signed Mode

We first state the regret guarantee for the signed mode of **ES-LSVI**.

Theorem 2 (Regret bound for the signed mode)**.** *Let $\delta \in (0, 1/2)$ and $T := KH$, and set*

$$\lambda = 1, \qquad n := \max\{KH, dH, 2\}, \qquad m := \left\lceil 2000dH \log \frac{8n}{\delta} \right\rceil.$$

There exist choices of $\eta > 0$ and $\alpha_U > 0$ satisfying

$$\eta = \widetilde{\mathcal{O}}(dH^{3/2}), \qquad \alpha_U^{-1} = \widetilde{\mathcal{O}}(d^{3/2}H^{3/2}),$$

such that, with $\alpha_L = \alpha_U/2$, running Algorithm 2 in the signed *mode yields, with probability at least $1 - \delta$,*

$$\mathrm{Regret}(K) \le \widetilde{\mathcal{O}}\left(d^2H^2\sqrt{T} + d^4H^5\right).$$

The proof has three main steps. First, we control the regression errors and ensemble perturbations uniformly, establish boundedness of the recursively constructed values, and derive a common Bellman residual bound. Second, we combine a persistent ensemble exceedance bound with the Bellman recursion along an optimal policy to show that a constant fraction of ensemble members are optimistic at the initial state of each episode. Finally, Bellman telescoping relates this optimism to the

Algorithm 2 `ES-LSVI` for Linear MDPs

Require: Number of episodes K, ensemble size m, mode $\mathsf{M} \in \{\text{signed}, \text{absolute}\}$, regularization $\lambda > 0$, exploration scale $\eta > 0$, value cutoff parameters $0 < \alpha_L < \alpha_U$

Initialize $\Lambda_{1,h} \leftarrow \lambda I_d$ for all $h \in [H]$
Independently sample $g_h^j \sim \mathcal{N}(0, \eta^2 I_d)$ and set $s_{1,h}^j \leftarrow \lambda g_h^j$ for all $(h, j) \in [H] \times [m]$
for each episode $k = 1, \ldots, K$ **do**
  Observe $s_{k,1}$
  Sample $J_k \sim \mathrm{Unif}([m])$
  Set $V_{k,H+1}^{J_k}(\cdot) \leftarrow 0$
  for each step $h = H, H-1, \ldots, 1$ **do**
    $\widetilde{w}_{k,h}^{J_k} \leftarrow \Lambda_{k,h}^{-1} \left(\sum_{i=1}^{k-1} \phi_{i,h} \left[r_{i,h} + V_{k,h+1}^{J_k}(s_{i,h+1}) \right] + s_{k,h}^{J_k} \right)$
    if $\mathsf{M} = \text{signed}$ **then**
      $\widetilde{Q}_{k,h}^{J_k}(s,a) \leftarrow \langle \phi(s,a), \widetilde{w}_{k,h}^{J_k} \rangle$
    else if $\mathsf{M} = \text{absolute}$ **then**
      $\hat{w}_{k,h}^{J_k} \leftarrow \Lambda_{k,h}^{-1} \sum_{i=1}^{k-1} \phi_{i,h} \left[r_{i,h} + V_{k,h+1}^{J_k}(s_{i,h+1}) \right]$
      $\widetilde{Q}_{k,h}^{J_k}(s,a) \leftarrow \langle \phi(s,a), \hat{w}_{k,h}^{J_k} \rangle + \left| \langle \phi(s,a), \widetilde{w}_{k,h}^{J_k} - \hat{w}_{k,h}^{J_k} \rangle \right|$
    end if
    $Q_{k,h}^{J_k}(s,a) \leftarrow \nu\Big(\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}} \Big) \widetilde{Q}_{k,h}^{J_k}(s,a) + \Big[1 - \nu\Big(\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}} \Big) \Big] (H - h + 1)$
    $V_{k,h}^{J_k}(\cdot) \leftarrow \max_{a \in \mathcal{A}} Q_{k,h}^{J_k}(\cdot, a)$
  end for
  for each step $h = 1, \ldots, H$ **do**
    Take $a_{k,h} \in \mathrm{argmax}_{a \in \mathcal{A}} Q_{k,h}^{J_k}(s_{k,h}, a)$, and observe $r_{k,h} := r_h(s_{k,h}, a_{k,h})$ and $s_{k,h+1}$
    Set $\phi_{k,h} \leftarrow \phi(s_{k,h}, a_{k,h})$ and $\Lambda_{k+1,h} \leftarrow \Lambda_{k,h} + \phi_{k,h}\phi_{k,h}^\top$
    for each ensemble member $j = 1, \ldots, m$ **do**
      Independently sample $\xi_{k,h}^j \sim \mathcal{N}(0, \eta^2)$ and set $s_{k+1,h}^j \leftarrow s_{k,h}^j + \phi_{k,h}\xi_{k,h}^j$
    end for
  end for
end for

regret of the uniformly selected member. Elliptical potential and martingale bounds then yield the cumulative regret guarantee.

For the analysis, let π_k^j denote the greedy policy induced by $Q_{k,h}^j$, i.e.,

$$\pi_{k,h}^j(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} Q_{k,h}^j(s,a), \qquad h \in [H].$$

B.1.1 Bellman residual bounds for the signed mode

We establish a uniform bound on the Bellman residual of Algorithm 2, defined by

$$\Delta_{k,h}^j(s,a) := Q_{k,h}^j(s,a) - r_h(s,a) - (P_h V_{k,h+1}^j)(s,a). \tag{3}$$

The analysis proceeds in three steps. We first control the persistent ensemble perturbations. We then bound the regression error uniformly over a deterministic class of bounded continuation values. Finally, a backward induction verifies that the actual continuation values belong to this class and establishes the Bellman residual bound.

Uniform bound on ensemble perturbations. The perturbation in each raw action-value estimate is $\phi(s,a)^\top \Lambda_{k,h}^{-1} s_{k,h}^j$. We control its magnitude through a uniform bound on $\|s_{k,h}^j\|_{\Lambda_{k,h}^{-1}}$ across all episodes, steps, and ensemble members.

Lemma 5. *Suppose that $\|\phi(s,a)\|_2 \le 1$ and let $\lambda \ge 1$. For any $\delta \in (0,1)$, with probability at least $1-\delta/8$, for all $k \in [K]$, $h \in [H]$, and $j \in [m]$, we have*

$$\|s^j_{k,h}\|_{\Lambda^{-1}_{k,h}} \le \eta \underbrace{\left[\sqrt{\lambda}\left(\sqrt{d} + \sqrt{2\log\frac{16mH}{\delta}}\right) + \sqrt{d\log\left(1+\frac{K}{\lambda d}\right) + 2\log\frac{16mH}{\delta}} \right]}_{:=\Gamma_\delta}.$$

By the Cauchy–Schwarz inequality, this lemma implies

$$\left|\phi(s,a)^\top \Lambda^{-1}_{k,h} s^j_{k,h}\right| \le \eta\Gamma_\delta \|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}$$

uniformly over all k, h, j and (s,a). Thus, the perturbation is controlled by the corresponding feature uncertainty. We next establish an analogous bound for the error of the reference regression estimate.

Concentration of Bellman regression estimates. The value estimates $V^j_{k,h+1}$ at the next step, which appear in the regression targets, depend on the same data used to fit the regression. To handle this dependence, we follow the framework of Zanette et al. (2020): cover a deterministic class of bounded value functions and apply a self-normalized martingale inequality to the resulting finite net.

This framework applies to the signed mode because the reference estimate and ensemble perturbation combine into a single coefficient vector:

$$\vartheta^j_{k,h} := \hat{w}^j_{k,h} + \Lambda^{-1}_{k,h} s^j_{k,h}, \qquad \widetilde{Q}^j_{k,h}(s,a) = \phi(s,a)^\top \vartheta^j_{k,h}. \tag{4}$$

For $\vartheta \in \mathbb{R}^d$, $A \succ 0$ and $B_h := H - h + 1$, define

$$q_{\vartheta,A,h}(s,a) := B_h + \nu(\|\phi(s,a)\|_A)\left(\phi(s,a)^\top \vartheta - B_h\right).$$

The value-bounding operator therefore gives

$$Q^j_{k,h} = q_{\vartheta^j_{k,h},\Lambda^{-1}_{k,h},h},$$

which has the same functional form as the truncated estimates considered in Zanette et al. (2020).

The perturbation bound controls the contribution of the ensemble noise to the coefficient radius. When the value estimates at subsequent stages are bounded, the reference ridge estimate also has a controlled radius. These observations allow us to place $V^j_{k,h+1}$ in a deterministic class with bounded parameters and bounded function values. We now state the formal version of Lemma 1. The result is phrased as an inductive implication; the required bounds at subsequent stages are established by backward induction in the proof below.

Lemma 6 (Concentration of Bellman regression estimates)**.** *Consider the* signed *mode of Algorithm 2. Fix $\lambda \ge 1, \eta > 0, 0 < \alpha_L < \alpha_U$ and $\delta \in (0,1)$. Let*

$$\widetilde{\mathcal{E}} := \left\{ \|s^j_{k,h}\|_{\Lambda^{-1}_{k,h}} \le \eta\Gamma_\delta \,, \forall k \in [K],\, h \in [H],\, j \in [m] \right\},$$

where Γ_δ is defined in Lemma 5. There exists an event $\hat{\mathcal{E}}$ with $\Pr(\hat{\mathcal{E}}) \ge 1-\delta/8$ such that, on $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}$, the following implication holds simultaneously for every $k \in [K]$, $h \in [H]$, and $j \in [m]$: if

$$\sup_{s,a} |Q^j_{k,\ell}(s,a) - Q^\star_\ell(s,a)| \le B_\ell \qquad \textit{for every } \ell \in \{h+1,\dots,H\}, \tag{5}$$

then

$$\left\| \hat{w}^j_{k,h} - w_h(V^j_{k,h+1}) \right\|_{\Lambda_{k,h}} \le C_\beta \sqrt{\lambda} H d\iota = \beta\,,$$

where $C_\beta > 0$ is an absolute constant and ι is the logarithmic term depending on $K, H, d, m, \eta, \alpha_U, \alpha_L, \lambda$, and $1/\delta$.

Value boundedness and Bellman residual bounds. On the event $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}$ and under the induction condition, the preceding two lemmas jointly control the raw Bellman residual:

$$\left|\widetilde{Q}^j_{k,h}(s,a) - r_h(s,a) - (P_h V^j_{k,h+1})(s,a)\right| \le (\beta + \eta\Gamma_\delta)\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}.$$

The two terms in the coefficient correspond to the regression error and the ensemble perturbation, respectively.

To complete the induction, we choose the thresholds so that this error is at most one whenever the raw estimate contributes to $Q^j_{k,h}$. Outside this region, the value-bounding operator sets the estimate to B_h. Starting from the terminal value zero, these properties preserve the required error bound relative to $Q^\star_h$ at every stage. The following lemma makes this argument precise and accounts for the additional residual introduced by the value-bounding operator.

Lemma 7 (Bellman residual bound for signed value iteration)**.** *Consider the signed mode of Algorithm 2. Let Γ_δ, β, $\widetilde{\mathcal{E}}$, and $\hat{\mathcal{E}}$ be as in Lemmas 5 and 6. Suppose that the thresholds satisfy*

$$0 < \alpha_L < \alpha_U, \qquad \alpha_U(\beta + \eta\Gamma_\delta) \le 1.$$

On the event $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}$, the following bounds hold simultaneously for all $k \in [K]$, $h \in [H]$, $j \in [m]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$|Q^j_{k,h}(s,a) - Q^\star_h(s,a)| \le B_h, \tag{6}$$

and

$$|\Delta^j_{k,h}(s,a)| \le (\beta + \eta\Gamma_\delta)\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} + 2H\mathbb{1}\left\{\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} > \alpha_L\right\}. \tag{7}$$

The residual bound separates the contribution of regression and perturbation errors from the correction due to the value-bounding operator. The correction vanishes whenever $\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} \le \alpha_L$. Moreover, the right-hand side is common to all ensemble members, providing a uniform residual bound for the subsequent regret analysis.

B.1.2 Stochastic optimism of signed ensemble sampling

We show that a uniformly selected ensemble member is optimistic at the initial state with sufficient probability. Specifically, we prove that a constant fraction of the ensemble members are optimistic in every episode. The following is the formal statement of Lemma 2.

Lemma 8 (Stochastic optimism)**.** *Consider the signed mode of Algorithm 2. Fix $\lambda \ge 1$ and $\delta \in (0, 1/2)$, and let Γ_δ, β, $\widetilde{\mathcal{E}}$, and $\hat{\mathcal{E}}$ be as in Lemmas 5 and 6. Set $n := \max\{KH, dH, 2\}$ and suppose that*

$$m \ge 2000dH\log\frac{8n}{\delta}, \qquad \eta \ge 40\beta\sqrt{H},$$

and

$$0 < \alpha_L < \alpha_U, \qquad \alpha_U(\beta + \eta\Gamma_\delta) \le 1.$$

Let $\mathcal{F}_k$ denote the information available immediately before drawing J_k, including the initial state $s_{k,1}$, past observations, and all ensemble perturbations generated so far. Then, there exists an event $\mathcal{E}_{\mathrm{exc}}$ with $\Pr(\mathcal{E}_{\mathrm{exc}}) \ge 1 - \delta/8$ such that, on

$$\mathcal{E} := \widetilde{\mathcal{E}} \cap \hat{\mathcal{E}} \cap \mathcal{E}_{\mathrm{exc}},$$

the following holds simultaneously for every $k \in [K]$:

$$\Pr\left(V^{J_k}_{k,1}(s_{k,1}) \ge V^\star_1(s_{k,1}) \,\middle|\, \mathcal{F}_k\right) = \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\left\{V^j_{k,1}(s_{k,1}) \ge V^\star_1(s_{k,1})\right\} \ge \frac{1}{10}.$$

Note that the exploration in `OPT-RLSVI` (Zanette et al., 2020) exploits fresh, conditionally Gaussian perturbations, whereas the proposed persistent ensemble perturbations are coupled with the data they helped collect. Beyond the dependence induced by adaptive data collection in bandit ensemble sampling (Janz et al., 2024; 2026; Hwang and Oh, 2026), Bellman recursion involves value estimates at the next stage that differ across ensemble members and requires the same member's perturbations to offset errors accumulated across stages. We address these issues by combining uniform Bellman

regression concentration with the gated recursion of Zanette et al. (2020), which yields feature directions shared by all ensemble members. Stacking the stages into a single dH-dimensional process then allows the uniform-exceedance result of Janz et al. (2026) to control the aggregate perturbation along these data-dependent directions. The $\sqrt{H}$ inflation in η compensates for the sum of stagewise regression errors, yielding a constant optimistic fraction in the persistent ensemble.

B.1.3 CUMULATIVE REGRET

Now we are ready to prove Theorem 2.

Proof of Theorem 2. We begin by decomposing the cumulative regret as follows:

$$\begin{aligned}\mathrm{Regret}(K) &= \sum_{k=1}^{K}(V_1^\star(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1})) \\ &= \sum_{k=1}^{K}\mathbb{E}\left[V_1^\star(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1})\,\middle|\,\mathcal{F}_k\right] + \sum_{k=1}^{K}\underbrace{\left(\mathbb{E}\left[V_1^{\pi_k^{J_k}}(s_{k,1})\,\middle|\,\mathcal{F}_k\right] - V_1^{\pi_k^{J_k}}(s_{k,1})\right)}_{=:X_k},\end{aligned}$$

where we used that $V_1^\star(s_{k,1})$ is $\mathcal{F}_k$-measurable. By construction,

$$\mathbb{E}[X_k \mid \mathcal{F}_k] = 0, \qquad |X_k| \le H,$$

where the second property follows from $0 \le V_1^{\pi_k^{J_k}}(s_{k,1}) \le H$. Thus, $(X_k)_{k=1}^K$ is a martingale difference sequence with increments bounded in absolute value by H. Azuma–Hoeffding therefore gives, with probability at least $1-\delta/4$,

$$\sum_{k=1}^{K} X_k \le H\sqrt{2K\log\frac{4}{\delta}}.$$

It remains to control the conditional expected regret. The following lemma uses stochastic optimism to bound this quantity by the expected sum of the maximum absolute Bellman residuals along the sampled policy.

Lemma 3 (From stochastic optimism to regret). *Under the conditions of Lemma 8, on $\mathcal{E}$, simultaneously for all $k \in [K]$,*

$$\mathbb{E}\left[V_1^\star(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1})\,\middle|\,\mathcal{F}_k\right] \le 21\mathbb{E}\left[\sum_{h=1}^{H}\max_{j\in[m]}|\Delta_{k,h}^{j}(s_{k,h},a_{k,h})|\,\middle|\,\mathcal{F}_k\right].$$

This lemma shows that having a constant fraction of optimistic ensemble members suffices to control the expected regret under uniform ensemble sampling. The proof of Lemma 3 is provided in Appendix B.3.

On $\mathcal{E}$, Lemmas 3 and 7 imply

$$\mathbb{E}\left[V_1^\star(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1})\,\middle|\,\mathcal{F}_k\right] \le 21\mathbb{E}\left[\sum_{h=1}^{H}\left((\beta+\eta\Gamma_\delta)\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} + 2H\mathbb{1}\left\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} > \alpha_L\right\}\right)\,\middle|\,\mathcal{F}_k\right].$$

Summing over episodes, we have

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}\left[V_1^\star(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1})\,\middle|\,\mathcal{F}_k\right] &\le 21(\beta+\eta\Gamma_\delta)\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} \\ &\quad + 42H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{1}\left\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} > \alpha_L\right\} + \sum_{k=1}^{K}Y_k\,,\end{aligned}$$

where we denote

$$Y_k := 21\mathbb{E}\left[\sum_{h=1}^{H}\left((\beta+\eta\Gamma_\delta)\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}+2H\mathbb{1}\left\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}>\alpha_L\right\}\right)\,\middle|\,\mathcal{F}_k\right]$$
$$-21(\beta+\eta\Gamma_\delta)\sum_{h=1}^{H}\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}-42H\sum_{h=1}^{H}\mathbb{1}\left\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}>\alpha_L\right\}.$$

By construction, Y_k is $\mathcal{F}_{k+1}$-measurable and satisfies $\mathbb{E}[Y_k \mid \mathcal{F}_k]=0$. Moreover, since $\lambda \geq 1$ and $\|\phi_{k,h}\|_2 \leq 1$, we have

$$0 \leq 21(\beta+\eta\Gamma_\delta)\sum_{h=1}^{H}\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}+42H\sum_{h=1}^{H}\mathbb{1}\left\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}>\alpha_L\right\}$$
$$\leq 21H(\beta+\eta\Gamma_\delta+2H).$$

Its conditional expectation lies in the same interval, so

$$|Y_k| \leq 21H(\beta+\eta\Gamma_\delta+2H).$$

Thus, $(Y_k)_{k=1}^K$ is a bounded martingale difference sequence on the full probability space. Azuma–Hoeffding therefore gives, with probability at least $1-\delta/4$,

$$\sum_{k=1}^{K}Y_k \leq 21H(\beta+\eta\Gamma_\delta+2H)\sqrt{2K\log\frac{4}{\delta}}.$$

Combining these bounds with $\Pr(\mathcal{E}) \geq 1-3\delta/8$ and applying a union bound, we obtain, with probability at least $1-7\delta/8 \geq 1-\delta$,

$$\mathrm{Regret}(K) \leq 21(\beta+\eta\Gamma_\delta)\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}+42H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{1}\left\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}>\alpha_L\right\}$$
$$+H\sqrt{2K\log\frac{4}{\delta}}+21H(\beta+\eta\Gamma_\delta+2H)\sqrt{2K\log\frac{4}{\delta}}.$$

Note that Cauchy–Schwarz and the elliptical potential lemma (Lemma 15) give

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} \leq \sqrt{KH\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi_{k,h}\|^2_{\Lambda_{k,h}^{-1}}} \leq \sqrt{2dH^2K\log\left(1+\frac{K}{\lambda d}\right)}.$$

Also, since $\mathbb{1}\{x>\alpha_L\} \leq x^2/\alpha_L^2$, we obtain

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{1}\left\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}>\alpha_L\right\} \leq \frac{1}{\alpha_L^2}\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi_{k,h}\|^2_{\Lambda_{k,h}^{-1}} \leq \frac{2dH}{\alpha_L^2}\log\left(1+\frac{K}{\lambda d}\right).$$

Choosing

$$\eta = 40\beta\sqrt{H}, \qquad \alpha_U = \frac{1}{\beta+\eta\Gamma_\delta}, \qquad \alpha_L = \frac{\alpha_U}{2},$$

with the logarithmic factor in β chosen sufficiently large to satisfy Lemma 6, gives

$$\beta+\eta\Gamma_\delta = \widetilde{O}\left(d^{3/2}H^{3/2}\right).$$

Consequently, with $T=KH$,

$$\mathrm{Regret}(K) = \widetilde{\mathcal{O}}(d^2H^2\sqrt{T}+d^4H^5+H\sqrt{K}+d^{3/2}H^2\sqrt{T}).$$

□

B.2 Regret Bound for the Absolute Mode

Theorem 3 (Regret bound for the absolute mode)**.** *Let $\delta \in (0, 1/2)$ and $T := KH$, and set*

$$\lambda = 1, \qquad n := \max\{K, d, 2\}, \qquad m := \left\lceil 2000 d \log \frac{8Hn}{\delta} \right\rceil .$$

There exist choices of $\eta > 0$ and $\alpha_U > 0$ satisfying

$$\eta = \widetilde{\mathcal{O}}(dH), \qquad \alpha_U^{-1} = \widetilde{\mathcal{O}}(d^{3/2}H),$$

such that, with $\alpha_L = \alpha_U/2$, Algorithm 2 in the absolute *mode satisfies, with probability at least $1 - \delta$,*

$$\mathrm{Regret}(K) \le \widetilde{\mathcal{O}} \left(d^2 H^{3/2} \sqrt{T} + d^4 H^4 \right).$$

B.2.1 Bellman residual bounds for the absolute mode

The regression concentration argument for the absolute mode requires only a minor modification of the proof of Lemma 6. In this mode, the raw action-value estimate takes the form

$$\widetilde{Q}^j_{k,h}(s,a) = \phi(s,a)^\top \hat{w}^j_{k,h} + \left| \phi(s,a)^\top \Lambda^{-1}_{k,h} s^j_{k,h} \right| .$$

We therefore cover the regression and perturbation vectors separately. Since the absolute-value map is 1-Lipschitz, this changes only the constants in the covering-number bound and yields the same order of concentration.

Lemma 9 (Concentration of Bellman regression estimates for the absolute mode)**.** *Consider the absolute mode of Algorithm 2. Fix $\lambda \ge 1$, $\eta > 0$, $0 < \alpha_L < \alpha_U$, and $\delta \in (0, 1)$. Let $\widetilde{\mathcal{E}}$ be the perturbation event defined in Lemma 6. There exists an event $\hat{\mathcal{E}}_{\mathrm{abs}}$ with probability at least $1 - \delta/8$ such that, on $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}_{\mathrm{abs}}$, the following implication holds simultaneously for every $k \in [K]$, $h \in [H]$, and $j \in [m]$: if*

$$\sup_{s,a} \left| Q^j_{k,\ell}(s,a) - Q^\star_\ell(s,a) \right| \le B_\ell, \qquad \forall \ell \in \{h+1, \dots, H\}, \tag{8}$$

then

$$\left\| \hat{w}^j_{k,h} - w_h(V^j_{k,h+1}) \right\|_{\Lambda_{k,h}} \le C_\beta \sqrt{\lambda}\, H d\, \iota = \beta.$$

Here, ι is the logarithmic factor used in Lemma 6, and $C_\beta > 0$ is a sufficiently large universal constant.

Under $\alpha_U(\beta + \eta\Gamma_\delta) \le 1$, the backward induction in the proof of Lemma 7 applies with Lemma 9. Thus, the same value boundedness and Bellman residual bounds hold in the absolute mode on $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}_{\mathrm{abs}}$.

B.2.2 Optimism of ensemble-averaged value functions

We show that the ensemble-averaged value functions are optimistic at every stage and state. The key is to make the average absolute perturbation dominate the regression error at each state-action pair. This local inequality then propagates through the Bellman recursion. The following is the formal statement of Lemma 4.

Lemma 10 (Optimism of ensemble-averaged value functions)**.** *Consider the absolute mode of Algorithm 2. Fix $\delta \in (0, 1/2)$ and $\lambda \ge 1$, and set $n := \max\{K, d, 2\}$. Let Γ_δ, β, $\widetilde{\mathcal{E}}$, and $\hat{\mathcal{E}}_{\mathrm{abs}}$ be as in Lemmas 5 and 9. Suppose that*

$$m \ge 2000 d \log \frac{8Hn}{\delta}, \quad \eta \ge 400\beta, \quad 0 < \alpha_L < \alpha_U, \quad \alpha_U(\beta + \eta\Gamma_\delta) \le 1 .$$

Then there exists an event $\mathcal{E}^{\mathrm{abs}}_{\mathrm{exc}}$ with probability at least $1 - \delta/8$ such that, on

$$\mathcal{E}_{\mathrm{abs}} := \widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}_{\mathrm{abs}} \cap \mathcal{E}^{\mathrm{abs}}_{\mathrm{exc}},$$

the following inequalities hold simultaneously for all $k \in [K]$, $h \in [H]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\frac{1}{m}\sum_{j=1}^{m} Q_{k,h}^{j}(s,a) \geq Q_h^\star(s,a), \qquad \frac{1}{m}\sum_{j=1}^{m} V_{k,h}^{j}(s) \geq V_h^\star(s). \tag{9}$$

In particular, on the same event,

$$\mathbb{E}\Big[V_{k,1}^{J_k}(s_{k,1}) \,\Big|\, \mathcal{F}_k\Big] \geq V_1^\star(s_{k,1}) \qquad \forall k \in [K].$$

B.2.3 Cumulative regret

We now combine ensemble-average optimism with the Bellman residual bound to prove Theorem 3.

Proof of Theorem 3. We first derive a regret bound for deterministic parameters satisfying the conditions of Lemma 10, and then specify suitable parameter choices.

We begin by decomposing the cumulative regret:

$$\begin{aligned}\mathrm{Regret}(K) &= \sum_{k=1}^{K}\left(V_1^\star(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1})\right) \\ &= \sum_{k=1}^{K}\mathbb{E}\left[V_1^\star(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1}) \,\middle|\, \mathcal{F}_k\right] + \sum_{k=1}^{K} X_k,\end{aligned}$$

where

$$X_k := \mathbb{E}\left[V_1^{\pi_k^{J_k}}(s_{k,1}) \,\middle|\, \mathcal{F}_k\right] - V_1^{\pi_k^{J_k}}(s_{k,1}).$$

Here, we used that $V_1^\star(s_{k,1})$ is $\mathcal{F}_k$-measurable. By construction, X_k is $\mathcal{F}_{k+1}$-measurable and satisfies

$$\mathbb{E}[X_k \mid \mathcal{F}_k] = 0, \qquad |X_k| \leq H,$$

where the second property follows from $0 \leq V_1^{\pi_k^{J_k}}(s_{k,1}) \leq H$. Thus, Azuma–Hoeffding gives, with probability at least $1-\delta/4$,

$$\sum_{k=1}^{K} X_k \leq H\sqrt{2K\log\frac{4}{\delta}}.$$

It remains to control the conditional expected regret. Let $\mathcal{E}_{\mathrm{abs}}$ be the event in Lemma 10, with $\Pr(\mathcal{E}_{\mathrm{abs}}) \geq 1 - 3\delta/8$. On this event, by Lemma 10, the average value estimate is optimistic:

$$\mathbb{E}\Big[V_{k,1}^{J_k}(s_{k,1}) \,\Big|\, \mathcal{F}_k\Big] \geq V_1^\star(s_{k,1}) \qquad \forall k \in [K].$$

On $\mathcal{E}_{\mathrm{abs}}$, using Lemma 10 and Eq. (13) established in the proof of Lemma 3, we obtain

$$\begin{aligned}&\mathbb{E}\left[V_1^\star(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1}) \,\middle|\, \mathcal{F}_k\right] \\ &= \underbrace{\mathbb{E}\Big[V_1^\star(s_{k,1}) - V_{k,1}^{J_k}(s_{k,1}) \,\Big|\, \mathcal{F}_k\Big]}_{\leq 0} + \mathbb{E}\left[V_{k,1}^{J_k}(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1}) \,\middle|\, \mathcal{F}_k\right] \\ &\leq \mathbb{E}\left[\sum_{h=1}^{H}\Delta_{k,h}^{J_k}(s_{k,h},a_{k,h}) \,\middle|\, \mathcal{F}_k\right] \\ &\leq \mathbb{E}\left[\sum_{h=1}^{H}\Big((\beta+\eta\Gamma_\delta)\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} + 2H\mathbb{1}\Big\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} > \alpha_L\Big\}\Big) \,\middle|\, \mathcal{F}_k\right],\end{aligned}$$

where the last inequality follows from the uniform Bellman residual bound for the absolute mode (Lemma 9).

Define

$$Z_k := \sum_{h=1}^{H} \left((\beta + \eta\Gamma_\delta)\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} + 2H\mathbb{1}\Big\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} > \alpha_L\Big\}\right), \quad Y_k := \mathbb{E}[Z_k \mid \mathcal{F}_k] - Z_k .$$

Summing the conditional regret bounds over episodes therefore gives, on $\mathcal{E}_{\mathrm{abs}}$,

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}\left[V_1^\star(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1}) \,\middle|\, \mathcal{F}_k \right] \\
&\le \sum_{k=1}^{K} Z_k + \sum_{k=1}^{K} Y_k \\
&= (\beta + \eta\Gamma_\delta) \sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} + 2H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{1}\Big\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} > \alpha_L\Big\} + \sum_{k=1}^{K} Y_k.
\end{aligned}$$

By construction, Y_k is $\mathcal{F}_{k+1}$-measurable and satisfies $\mathbb{E}[Y_k \mid \mathcal{F}_k] = 0$. Moreover, since $\lambda = 1$ and $\|\phi_{k,h}\|_2 \le 1$, we have

$$0 \le Z_k \le H(\beta + \eta\Gamma_\delta + 2H) .$$

Its conditional expectation lies in the same interval, so

$$|Y_k| \le H(\beta + \eta\Gamma_\delta + 2H).$$

Thus, $(Y_k)_{k=1}^K$ is a bounded martingale difference sequence. Azuma–Hoeffding gives, with probability at least $1 - \delta/4$,

$$\sum_{k=1}^{K} Y_k \le H(\beta + \eta\Gamma_\delta + 2H)\sqrt{2K \log \frac{4}{\delta}}.$$

Combining these bounds with $\Pr(\mathcal{E}_{\mathrm{abs}}) \ge 1 - 3\delta/8$ and applying a union bound, we obtain, with probability at least $1 - 7\delta/8 \ge 1 - \delta$,

$$\begin{aligned}
\mathrm{Regret}(K) \le{}& (\beta + \eta\Gamma_\delta) \sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} + 2H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{1}\Big\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} > \alpha_L\Big\} \\
&+ H(\beta + \eta\Gamma_\delta + 2H + 1)\sqrt{2K \log \frac{4}{\delta}}.
\end{aligned}$$

The Cauchy–Schwarz inequality and the elliptical potential lemma (Lemma 15) with the determinant-trace inequality (Lemma 14) give

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} \le \sqrt{KH \sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi_{k,h}\|^2_{\Lambda_{k,h}^{-1}}} \le \sqrt{2dH^2K \log\left(1 + \frac{K}{d}\right)}.$$

Also, using $\mathbb{1}\{x > \alpha_L\} \le x^2/\alpha_L^2$, we obtain

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{1}\Big\{\|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} > \alpha_L\Big\} \le \frac{1}{\alpha_L^2} \sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi_{k,h}\|^2_{\Lambda_{k,h}^{-1}} \le \frac{2dH}{\alpha_L^2} \log\left(1 + \frac{K}{d}\right).$$

Choosing

$$\eta = 400\beta, \qquad \alpha_U = \frac{1}{\beta + \eta\Gamma_\delta}, \qquad \alpha_L = \frac{\alpha_U}{2},$$

with the logarithmic factors in $\beta = \widetilde{\mathcal{O}}(Hd)$ chosen sufficiently large to satisfy Lemma 9, gives

$$\eta = \widetilde{\mathcal{O}}(dH), \qquad \beta + \eta\Gamma_\delta = \widetilde{\mathcal{O}}(d^{3/2}H), \qquad \alpha_L^{-1} = 2\alpha_U^{-1} = \widetilde{\mathcal{O}}(d^{3/2}H).$$

These choices satisfy the conditions of Lemma 10. Consequently, with $T = KH$,

$$\begin{aligned}
\mathrm{Regret}(K) &\le \widetilde{\mathcal{O}}\Big(d^2H^{3/2}\sqrt{T} + d^4H^4 + d^{3/2}H^{3/2}\sqrt{T}\Big) \\
&= \widetilde{\mathcal{O}}\Big(d^2H^{3/2}\sqrt{T} + d^4H^4\Big) .
\end{aligned}$$

□

B.3 Proof of Technical Lemmas

B.3.1 Proof of Lemma 5

Proof of Lemma 5. The claim follows by adapting the uniform perturbation bound for linear ensemble sampling (Janz et al., 2026; Hwang and Oh, 2026) to each stage $h \in [H]$. We include the proof for completeness.

Fix $(h, j) \in [H] \times [m]$ and write

$$s^j_{k,h} = \lambda g^j_h + M^j_{k,h}, \qquad M^j_{k,h} := \sum_{i=1}^{k-1} \phi_{i,h} \xi^j_{i,h}.$$

Since $\Lambda_{k,h} \succeq \lambda I_d$,

$$\|s^j_{k,h}\|_{\Lambda^{-1}_{k,h}} \le \|\lambda g^j_h\|_{\Lambda^{-1}_{k,h}} + \|M^j_{k,h}\|_{\Lambda^{-1}_{k,h}} \le \sqrt{\lambda}\|g^j_h\|_2 + \|M^j_{k,h}\|_{\Lambda^{-1}_{k,h}}.$$

Because $g^j_h/\eta \sim \mathcal{N}(0, I_d)$, Gaussian concentration (Lemma 12) gives

$$\|g^j_h\|_2 \le \eta \left(\sqrt{d} + \sqrt{2 \log \frac{16mH}{\delta}} \right)$$

with failure probability at most $\delta/(16mH)$.

For the second term, $\phi_{i,h}$ is determined before $\xi^j_{i,h}$ is sampled, and $\xi^j_{i,h}$ has distribution $\mathcal{N}(0, \eta^2)$ conditional on the entire history preceding its generation. Thus, the noise is conditionally η-sub-Gaussian. The self-normalized martingale inequality (Lemma 13) therefore gives, with failure probability at most $\delta/(16mH)$,

$$\|M^j_{k,h}\|^2_{\Lambda^{-1}_{k,h}} \le \eta^2 \left[\log \frac{\det \Lambda_{k,h}}{\lambda^d} + 2 \log \frac{16mH}{\delta} \right]$$

simultaneously for all $k \in [K]$.

Using $\|\phi_{i,h}\|_2 \le 1$, the determinant–trace inequality (Lemma 14) gives

$$\log \frac{\det \Lambda_{k,h}}{\lambda^d} \le d \log \left(1 + \frac{k-1}{\lambda d} \right) \le d \log \left(1 + \frac{K}{\lambda d} \right).$$

Combining these bounds, we obtain

$$\|s^j_{k,h}\|_{\Lambda^{-1}_{k,h}} \le \eta \left[\sqrt{\lambda} \left(\sqrt{d} + \sqrt{2 \log \frac{16mH}{\delta}} \right) + \sqrt{d \log \left(1 + \frac{K}{\lambda d} \right) + 2 \log \frac{16mH}{\delta}} \right] = \eta \Gamma_\delta$$

simultaneously for all $k \in [K]$. Taking a union bound over the two failure events for each of the mH pairs (h, j) gives total failure probability at most

$$2mH \cdot \frac{\delta}{16mH} = \frac{\delta}{8},$$

which proves the claim. □

B.3.2 Proof of Lemma 6

Proof of Lemma 6. We adapt the covering and concentration argument of Zanette et al. (2020) to control the transition noise term uniformly over a deterministic class of bounded value functions. In the signed mode, the raw predictor has coefficient $\hat{w}^j_{k,\ell} + \Lambda^{-1}_{k,\ell} s^j_{k,\ell}$, so the parameter radius must account for both the reference ridge estimate and the persistent ensemble perturbation. We provide the argument for completeness.

Set

$$L := 2H \sqrt{\frac{K}{\lambda}} + \frac{\eta \Gamma_\delta}{\sqrt{\lambda}}, \qquad \epsilon := (K+1)^{-2},$$

and define

$$q_{\vartheta,A,\ell}(s,a) := B_\ell + \nu(\|\phi(s,a)\|_A)\left(\phi(s,a)^\top \vartheta - B_\ell\right).$$

For each $\ell \in [H]$, consider

$$O_\ell := \left\{(\vartheta, A) : \|\vartheta\|_2 \le L,\ 0 \prec A \preceq \lambda^{-1} I_d,\ \sup_{s,a} |q_{\vartheta,A,\ell}(s,a) - Q^\star_\ell(s,a)| \le B_\ell\right\},$$

and let

$$\mathcal{V}_\ell := \left\{s \mapsto \max_{a\in\mathcal{A}} q_{\vartheta,A,\ell}(s,a) : (\vartheta, A) \in O_\ell\right\}, \qquad \mathcal{V}_{H+1} := \{0\}.$$

These classes are deterministic because $Q^\star_\ell$ is fixed by the underlying MDP. Moreover, $0 \le Q^\star_\ell(s,a) \le B_\ell$ implies that every $V \in \mathcal{V}_\ell$ takes values in $[-B_\ell, 2B_\ell]$.

We first bound the sensitivity to the parameters. Fix $(\vartheta, A), (\vartheta', A') \in O_\ell$ and write $x := \phi(s,a)$. Adding and subtracting $\nu(\|x\|_A)(x^\top\vartheta' - B_\ell)$ gives

$$q_{\vartheta,A,\ell}(s,a) - q_{\vartheta',A',\ell}(s,a) = \nu(\|x\|_A)x^\top(\vartheta - \vartheta') + \left[\nu(\|x\|_A) - \nu(\|x\|_{A'})\right](x^\top\vartheta' - B_\ell).$$

Since $0 \le \nu \le 1$, ν is $(\alpha_U - \alpha_L)^{-1}$-Lipschitz, and $|x^\top\vartheta' - B_\ell| \le L + H$, it follows that

$$|q_{\vartheta,A,\ell}(s,a) - q_{\vartheta',A',\ell}(s,a)| \le \|\vartheta - \vartheta'\|_2 + \frac{L+H}{\alpha_U - \alpha_L}\left|\|x\|_A - \|x\|_{A'}\right|.$$

The dependence on A is controlled by

$$\left|\|x\|_A - \|x\|_{A'}\right| = \left|\sqrt{x^\top A x} - \sqrt{x^\top A' x}\right| \le \sqrt{|x^\top(A - A')x|} \le \sqrt{\|A - A'\|_{\mathrm{F}}},$$

where we used $|\sqrt{u} - \sqrt{v}| \le \sqrt{|u - v|}$ for $u, v \ge 0$ and $\|x\|_2 \le 1$. Taking the supremum over (s,a) therefore yields

$$\|q_{\vartheta,A,\ell} - q_{\vartheta',A',\ell}\|_\infty \le \|\vartheta - \vartheta'\|_2 + \frac{L+H}{\alpha_U - \alpha_L}\sqrt{\|A - A'\|_{\mathrm{F}}}.$$

Define

$$\Theta_L := \{\vartheta \in \mathbb{R}^d : \|\vartheta\|_2 \le L\}, \qquad \mathcal{B}_\lambda := \left\{A \in \mathbb{S}^d : \|A\|_{\mathrm{F}} \le \frac{\sqrt{d}}{\lambda}\right\},$$

where $\mathbb{S}^d := \{A \in \mathbb{R}^{d\times d} : A = A^\top\}$. Since $0 \prec A \preceq \lambda^{-1} I_d$ implies $\|A\|_{\mathrm{F}} \le \sqrt{d}/\lambda$, we have $O_\ell \subseteq \Theta_L \times \mathcal{B}_\lambda$. The sets Θ_L and $\mathcal{B}_\lambda$ are Euclidean balls in $(\mathbb{R}^d, \|\cdot\|_2)$ and $(\mathbb{S}^d, \|\cdot\|_{\mathrm{F}})$, respectively; these spaces have dimensions d and $d(d+1)/2$. Choose

$$r_\theta := \frac{\epsilon}{4}, \qquad r_A := \frac{(\alpha_U - \alpha_L)^2\epsilon^2}{16(L+H)^2}.$$

Let $\mathcal{C}_\theta$ be a minimum-cardinality r_θ-cover of Θ_L under $\|\cdot\|_2$, and let $\mathcal{C}_A$ be a minimum-cardinality r_A-cover of $\mathcal{B}_\lambda$ under $\|\cdot\|_{\mathrm{F}}$. By the standard covering-number bound for Euclidean balls (Lemma 16), we have

$$|\mathcal{C}_\theta| \le \left(1 + \frac{2L}{r_\theta}\right)^d, \qquad |\mathcal{C}_A| \le \left(1 + \frac{2\sqrt{d}}{\lambda r_A}\right)^{d(d+1)/2}.$$

For each product of covering balls intersecting O_ℓ, choose a representative from that intersection. Every $(\vartheta, A) \in O_\ell$ then shares a product of covering balls with some representative $(\vartheta', A') \in O_\ell$, so

$$\|\vartheta - \vartheta'\|_2 \le 2r_\theta, \qquad \|A - A'\|_{\mathrm{F}} \le 2r_A.$$

The preceding continuity bound gives

$$\|q_{\vartheta,A,\ell} - q_{\vartheta',A',\ell}\|_\infty \le 2r_\theta + \frac{L+H}{\alpha_U - \alpha_L}\sqrt{2r_A} = \frac{\epsilon}{2} + \frac{\epsilon}{2\sqrt{2}} \le \epsilon.$$

Since maximization over actions is nonexpansive, these representatives induce a deterministic ϵ-net $\mathcal{C}_\ell \subseteq \mathcal{V}_\ell$ with at most $|\mathcal{C}_\theta|\,|\mathcal{C}_A|$ elements. Then, we have

$$\log \mathcal{N}(\mathcal{V}_\ell, \epsilon, \|\cdot\|_\infty) \le \log|\mathcal{C}_\ell| \le \log|\mathcal{C}_\theta| + \log|\mathcal{C}_A| \le \underbrace{d\log\left(1+\frac{8L}{\epsilon}\right) + \frac{d(d+1)}{2}\log\left(1+\frac{32\sqrt{d}\,(L+H)^2}{\lambda(\alpha_U-\alpha_L)^2\epsilon^2}\right)}_{=:D_\epsilon}$$

Because the representatives lie in O_ℓ, every function in $\mathcal{C}_\ell$ takes values in $[-B_\ell, 2B_\ell]$. Set $\mathcal{C}_{H+1} := \{0\}$. We next use these deterministic nets to obtain uniform concentration over the value-function classes $\mathcal{V}_{h+1}$. For a deterministic function V, define

$$M_{k,h}(V) := \sum_{i=1}^{k-1} \phi_{i,h}\left[V(s_{i,h+1}) - (P_hV)(s_{i,h}, a_{i,h})\right].$$

For each fixed $\bar{V} \in \mathcal{C}_{h+1}$, the centered transition residuals are martingale differences bounded in absolute value by $3H$, and hence are conditionally $3H$-sub-Gaussian. Moreover, $\phi_{i,h}$ is measurable before $s_{i,h+1}$ is sampled. Applying the time-uniform self-normalized inequality (Lemma 13) with initial matrix λI_d and failure probability $\delta/(8H|\mathcal{C}_{h+1}|)$, and taking a union bound over the net functions and stages, yields an event $\hat{\mathcal{E}}$ with $\Pr(\hat{\mathcal{E}}) \ge 1-\delta/8$ on which

$$\|M_{k,h}(\bar{V})\|_{\Lambda_{k,h}^{-1}} \le 3H\left[\log\frac{\det\Lambda_{k,h}}{\lambda^d} + 2\log\frac{8H|\mathcal{C}_{h+1}|}{\delta}\right]^{1/2}$$

simultaneously for all $k \in [K]$, $h \in [H]$, and $\bar{V} \in \mathcal{C}_{h+1}$.

To extend these bounds to $\mathcal{V}_{h+1}$, take any $V \in \mathcal{V}_{h+1}$ and choose $\bar{V} \in \mathcal{C}_{h+1}$ with $\|V-\bar{V}\|_\infty \le \epsilon$. By the dual norm identity,

$$\|M_{k,h}(V) - M_{k,h}(\bar{V})\|_{\Lambda_{k,h}^{-1}} = \sup_{\|u\|_{\Lambda_{k,h}}\le 1} \left|u^\top\big(M_{k,h}(V) - M_{k,h}(\bar{V})\big)\right|.$$

Fix $u \in \mathbb{R}^d$ with $\|u\|_{\Lambda_{k,h}} \le 1$. The definition of $\Lambda_{k,h}$ gives

$$\sum_{i=1}^{k-1}(u^\top\phi_{i,h})^2 = u^\top(\Lambda_{k,h} - \lambda I_d)u \le \|u\|^2_{\Lambda_{k,h}} \le 1.$$

Also, since $P_h(\cdot \mid s,a)$ is a probability measure,

$$\left|(V-\bar{V})(s_{i,h+1}) - \big(P_h(V-\bar{V})\big)(s_{i,h},a_{i,h})\right| \le \|V-\bar{V}\|_\infty + \big(P_h|V-\bar{V}|\big)(s_{i,h},a_{i,h}) \le 2\epsilon\,.$$

Applying the Cauchy–Schwarz inequality and using these two bounds, we obtain

$$\begin{aligned}
&\left|u^\top\big(M_{k,h}(V) - M_{k,h}(\bar{V})\big)\right| \\
&= \left|\sum_{i=1}^{k-1}(u^\top\phi_{i,h})\left[(V-\bar{V})(s_{i,h+1}) - \big(P_h(V-\bar{V})\big)(s_{i,h},a_{i,h})\right]\right| \\
&\le \left(\sum_{i=1}^{k-1}(u^\top\phi_{i,h})^2\right)^{1/2}\left(\sum_{i=1}^{k-1}\left[(V-\bar{V})(s_{i,h+1}) - \big(P_h(V-\bar{V})\big)(s_{i,h},a_{i,h})\right]^2\right)^{1/2} \\
&\le 2\sqrt{k-1}\,\epsilon.
\end{aligned}$$

Taking the supremum over all u with $\|u\|_{\Lambda_{k,h}} \le 1$ therefore yields

$$\|M_{k,h}(V) - M_{k,h}(\bar{V})\|_{\Lambda_{k,h}^{-1}} \le 2\sqrt{k-1}\,\epsilon.$$

Therefore, on $\hat{\mathcal{E}}$, for any $V \in \mathcal{V}_{h+1}$, we have

$$\begin{aligned}\|M_{k,h}(V)\|_{\Lambda_{k,h}^{-1}} &\le \|M_{k,h}(V) - M_{k,h}(\bar{V})\|_{\Lambda_{k,h}^{-1}} + \|M_{k,h}(\bar{V})\|_{\Lambda_{k,h}^{-1}} \\ &\le 2\sqrt{k-1}\epsilon + 3H\left[\log\frac{\det\Lambda_{k,h}}{\lambda^d} + 2\log\frac{8H|\mathcal{C}_{h+1}|}{\delta}\right]^{1/2}. \end{aligned} \tag{10}$$

Combining this approximation bound with $\log|\mathcal{C}_{h+1}| \le D_\epsilon$ and the determinant–trace inequality (Lemma 14) yields

$$\|M_{k,h}(V)\|_{\Lambda_{k,h}^{-1}} \le 3H\left[d\log\left(1+\frac{K}{\lambda d}\right) + 2D_\epsilon + 2\log\frac{8H}{\delta}\right]^{1/2} + 2\sqrt{K}\,\epsilon,$$

simultaneously for all $k \in [K]$, $h \in [H]$, and $V \in \mathcal{V}_{h+1}$.

It remains to verify that the value estimates $V_{k,h+1}^j$ at the next stage belong to the deterministic classes under the induction condition. Fix $k \in [K]$, $h \in [H]$, and $j \in [m]$. On $\widetilde{\mathcal{E}}$, suppose that Eq.(5) holds. Since $0 \le Q_\ell^\star \le B_\ell$, this condition implies

$$\|V_{k,\ell}^j\|_\infty \le 2B_\ell \qquad \forall \ell \in \{h+1,\dots,H\}.$$

If $h = H$, then $V_{k,H+1}^j = 0 \in \mathcal{V}_{H+1}$. Otherwise, let $\ell := h+1$. The preceding boundedness condition, together with $r_{i,\ell} \in [0,1]$ and $V_{k,H+1}^j = 0$, gives

$$\left|r_{i,\ell} + V_{k,\ell+1}^j(s_{i,\ell+1})\right| \le 2H \qquad \forall i < k.$$

By the optimality of $\hat{w}_{k,\ell}^j$, its ridge objective value is no greater than that obtained by setting $w = 0$ while keeping the regression targets fixed. Since the squared-error term is nonnegative, this gives

$$\begin{aligned}\lambda\|\hat{w}_{k,\ell}^j\|_2^2 &\le \sum_{i=1}^{k-1}\left(\langle\phi_{i,\ell}, \hat{w}_{k,\ell}^j\rangle - r_{i,\ell} - V_{k,\ell+1}^j(s_{i,\ell+1})\right)^2 + \lambda\|\hat{w}_{k,\ell}^j\|_2^2 \\ &\le \sum_{i=1}^{k-1}\left(r_{i,\ell} + V_{k,\ell+1}^j(s_{i,\ell+1})\right)^2 \le 4H^2K.\end{aligned}$$

Consequently, we have $\|\hat{w}_{k,\ell}^j\|_2 \le 2H\sqrt{K/\lambda}$. Define the combined signed parameter by

$$\vartheta_{k,\ell}^j := \hat{w}_{k,\ell}^j + \Lambda_{k,\ell}^{-1}s_{k,\ell}^j.$$

Since $\Lambda_{k,\ell} \succeq \lambda I_d$, on $\widetilde{\mathcal{E}}$ we have

$$\|\vartheta_{k,\ell}^j\|_2 \le \|\hat{w}_{k,\ell}^j\|_2 + \frac{1}{\sqrt{\lambda}}\|s_{k,\ell}^j\|_{\Lambda_{k,\ell}^{-1}} \le 2H\sqrt{\frac{K}{\lambda}} + \frac{\eta\Gamma_\delta}{\sqrt{\lambda}} = L.$$

This parameterization gives

$$Q_{k,\ell}^j = q_{\vartheta_{k,\ell}^j, \Lambda_{k,\ell}^{-1}, \ell}.$$

Together with $0 \prec \Lambda_{k,\ell}^{-1} \preceq \lambda^{-1}I_d$ and the induction condition at stage ℓ, this verifies all defining conditions of O_ℓ. Thus,

$$(\vartheta_{k,\ell}^j, \Lambda_{k,\ell}^{-1}) \in O_\ell, \qquad V_{k,h+1}^j \in \mathcal{V}_{h+1}.$$

Therefore, under the stated conditions, the uniform concentration bound in (10) applies to $M_{k,h}(V_{k,h+1}^j)$ on $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}$ for every $h \in [H]$.

By the linear Bellman representation,

$$\phi_{i,h}^\top w_h(V_{k,h+1}^j) = r_{i,h} + (P_hV_{k,h+1}^j)(s_{i,h}, a_{i,h}).$$

Then, we obtain

$$\begin{aligned}
&\Lambda_{k,h}\big[\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\big] \\
&= \sum_{i=1}^{k-1} \phi_{i,h}\left[r_{i,h} + V^j_{k,h+1}(s_{i,h+1}) - \phi_{i,h}^\top w_h(V^j_{k,h+1})\right] - \lambda w_h(V^j_{k,h+1}) \\
&= \sum_{i=1}^{k-1} \phi_{i,h}\left[V^j_{k,h+1}(s_{i,h+1}) - (P_h V^j_{k,h+1})(s_{i,h},a_{i,h})\right] - \lambda w_h(V^j_{k,h+1}) \\
&= M_{k,h}(V^j_{k,h+1}) - \lambda w_h(V^j_{k,h+1}).
\end{aligned}$$

Recall that the linear Bellman parameter is

$$w_h(V) = \theta_h + \int V(s')\,\mu_h(ds'),$$

and the boundedness assumption gives

$$\|\theta_h\|_2 \le \sqrt{d}, \qquad \left\|\int V(s')\,\mu_h(ds')\right\|_2 \le \sqrt{d}\,\|V\|_\infty$$

for every bounded function V. By the triangle inequality and $\|V^j_{k,h+1}\|_\infty \le 2H$, we obtain

$$\begin{aligned}
\|w_h(V^j_{k,h+1})\|_2 &\le \|\theta_h\|_2 + \left\|\int V^j_{k,h+1}(s')\,\mu_h(ds')\right\|_2 \\
&\le \sqrt{d} + \sqrt{d}\,\|V^j_{k,h+1}\|_\infty \\
&\le (1+2H)\sqrt{d} \le 3H\sqrt{d}\,.
\end{aligned}$$

Hence, on $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}$ and under the induction condition, we obtain,

$$\begin{aligned}
\|\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\|_{\Lambda_{k,h}} &\le \|M_{k,h}(V^j_{k,h+1})\|_{\Lambda^{-1}_{k,h}} + \lambda\|w_h(V^j_{k,h+1})\|_{\Lambda^{-1}_{k,h}} \\
&\le \|M_{k,h}(V^j_{k,h+1})\|_{\Lambda^{-1}_{k,h}} + \sqrt{\lambda}\,\|w_h(V^j_{k,h+1})\|_2 \\
&\le \|M_{k,h}(V^j_{k,h+1})\|_{\Lambda^{-1}_{k,h}} + 3H\sqrt{\lambda d} \\
&\le 3H\left[d\log\left(1+\frac{K}{\lambda d}\right) + 2D_\epsilon + 2\log\frac{8H}{\delta}\right]^{1/2} + 2\sqrt{K}\,\epsilon + 3H\sqrt{\lambda d},
\end{aligned}$$

where the last inequality follows from the uniform concentration bound in Eq. (10). Substituting $\epsilon = (K+1)^{-2}$ and expanding D_ϵ, we conclude that

$$\begin{aligned}
&\|\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\|_{\Lambda_{k,h}} \\
&\le 3H\left[d\log\left(1+\frac{K}{\lambda d}\right) + 2d\log\left(1+8L(K+1)^2\right) + d(d+1)\log\left(1+\frac{32\sqrt{d}\,(L+H)^2(K+1)^4}{\lambda(\alpha_U-\alpha_L)^2}\right)\right. \\
&\quad \left. + 2\log\frac{8H}{\delta}\right]^{1/2} + \frac{2\sqrt{K}}{(K+1)^2} + 3H\sqrt{\lambda d} \\
&\le C_\beta\sqrt{\lambda}Hd\iota = \beta\,,
\end{aligned}$$

where ι is the logarithmic term depending on $K, H, d, m, \eta, \alpha_U, \alpha_L, \lambda$, and $1/\delta$. The event $\hat{\mathcal{E}}$ was constructed uniformly over the deterministic classes, without conditioning on $\widetilde{\mathcal{E}}$ or the induction condition. These conditions are used only to establish $V^j_{k,h+1} \in \mathcal{V}_{h+1}$. Therefore, on $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}$, the stated implication holds simultaneously for all k, h, j, with no additional union bound over the ensemble members. □

B.3.3 PROOF OF LEMMA 7

Proof of Lemma 7. Suppose that the event $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}$ holds. Fix k, j. We first prove Eq. (6) by backward induction. At the terminal stage, $V^j_{k,H+1} = V^\star_{H+1} = 0$. For convenience, set $B_{H+1} := 0$.

Suppose that (6) holds at every stage $\ell > h$. Lemma 6 then gives

$$\|\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\|_{\Lambda_{k,h}} \le \beta.$$

The signed parameterization and the linear Bellman representation imply

$$\widetilde{Q}^j_{k,h}(s,a) - r_h(s,a) - (P_h V^j_{k,h+1})(s,a) = \phi(s,a)^\top \left[\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\right] + \phi(s,a)^\top \Lambda^{-1}_{k,h} s^j_{k,h}.$$

By the Cauchy–Schwarz inequality and the definition of $\widetilde{\mathcal{E}}$,

$$\begin{aligned}
&\left|\widetilde{Q}^j_{k,h}(s,a) - r_h(s,a) - (P_h V^j_{k,h+1})(s,a)\right| \\
&\le \|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} \left[\|\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\|_{\Lambda_{k,h}} + \|s^j_{k,h}\|_{\Lambda^{-1}_{k,h}}\right] \\
&\le (\beta + \eta\Gamma_\delta)\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}.
\end{aligned}$$

The induction hypothesis also implies $\|V^j_{k,h+1} - V^\star_{h+1}\|_\infty \le B_{h+1}$. Using the Bellman equation for $Q^\star_h$, we obtain

$$\begin{aligned}
|\widetilde{Q}^j_{k,h}(s,a) - Q^\star_h(s,a)| &\le (\beta + \eta\Gamma_\delta)\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} + \left|P_h(V^j_{k,h+1} - V^\star_{h+1})(s,a)\right| \\
&\le (\beta + \eta\Gamma_\delta)\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} + B_{h+1}.
\end{aligned}$$

We now bound $Q^j_{k,h}(s,a)$ by considering the three regimes of the value-bounding operator.

Case 1: $\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} \le \alpha_L$. In this case, we obtain $Q^j_{k,h}(s,a) = \widetilde{Q}^j_{k,h}(s,a)$. The preceding bound and $\alpha_U(\beta + \eta\Gamma_\delta) \le 1$ give

$$|Q^j_{k,h}(s,a) - Q^\star_h(s,a)| \le (\beta + \eta\Gamma_\delta)\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} + B_{h+1} \le 1 + B_{h+1} = B_h.$$

Case 2: $\alpha_L < \|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} < \alpha_U$. Here, the operator interpolates between $\widetilde{Q}^j_{k,h}(s,a)$ and B_h. The raw estimate satisfies

$$|\widetilde{Q}^j_{k,h}(s,a) - Q^\star_h(s,a)| \le (\beta + \eta\Gamma_\delta)\alpha_U + B_{h+1} \le B_h.$$

Also, $0 \le Q^\star_h(s,a) \le B_h$ implies $|B_h - Q^\star_h(s,a)| \le B_h$. Thus, both values being interpolated lie in the interval

$$[Q^\star_h(s,a) - B_h,\ Q^\star_h(s,a) + B_h].$$

Their convex combination $Q^j_{k,h}(s,a)$ lies in the same interval, and hence

$$|Q^j_{k,h}(s,a) - Q^\star_h(s,a)| \le B_h.$$

Case 3: $\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} \ge \alpha_U$. In this case, we obtain $Q^j_{k,h}(s,a) = B_h$. Therefore,

$$|Q^j_{k,h}(s,a) - Q^\star_h(s,a)| = B_h - Q^\star_h(s,a) \le B_h.$$

The bound holds in all three cases and therefore uniformly over (s,a) at stage h. This completes the backward induction.

We next bound the Bellman residual. The definition of the value-bounding operator gives

$$\begin{aligned}
\Delta^j_{k,h}(s,a) = \nu\Big(\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}\Big) &\left[\widetilde{Q}^j_{k,h}(s,a) - r_h(s,a) - (P_h V^j_{k,h+1})(s,a)\right] \\
&+ \left[1 - \nu\Big(\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}\Big)\right] \left[B_h - r_h(s,a) - (P_h V^j_{k,h+1})(s,a)\right].
\end{aligned}$$

By Eq. (6),

$$-B_{h+1} \le V^j_{k,h+1}(s') \le 2B_{h+1}$$

for every s', including the terminal stage. Together with $r_h(s,a) \in [0,1]$, this implies

$$\left| B_h - r_h(s,a) - (P_h V^j_{k,h+1})(s,a) \right| \le B_h + B_{h+1} \le 2H.$$

Combining this bound with the raw residual bound yields

$$|\Delta^j_{k,h}(s,a)| \le \nu\Big(\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}\Big)\,(\beta + \eta\Gamma_\delta)\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} + 2H\left[1 - \nu\Big(\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}\Big)\right].$$

Finally, $0 \le \nu \le 1$ and $\nu(u) = 1$ for $u \le \alpha_L$ give Eq. (7).

Both underlying events are uniform over k, h, j, so the conclusions hold simultaneously for all episodes, stages, and ensemble members. □

B.3.4 Proof of Lemma 8

Proof of Lemma 8. Fix an episode k and an optimal deterministic policy $\pi^\star$. Following the optimistic recursion framework of Zanette et al. (2020, Lemma F.1), let $\rho_{k,1}$ be the probability measure concentrated at $s_{k,1}$, defined by

$$\rho_{k,1}(C) := \mathbb{1}\{s_{k,1} \in C\} \qquad \text{for every measurable } C \subseteq \mathcal{S}.$$

Define the gated occupancy measures recursively by

$$\bar\rho_{k,h}(ds) := \nu\Big(\|\phi(s,\pi^\star_h(s))\|_{\Lambda^{-1}_{k,h}}\Big)\,\rho_{k,h}(ds),$$

and

$$\rho_{k,h+1}(C) := \int_{\mathcal{S}} P_h(C \mid s, \pi^\star_h(s))\,\bar\rho_{k,h}(ds).$$

These are nonnegative sub-probability measures. Define the corresponding feature directions by

$$\phi^*_{k,h} := \int_{\mathcal{S}} \phi(s,\pi^\star_h(s))\,\bar\rho_{k,h}(ds).$$

Note that the measures and directions depend on the observed Gram matrices and the initial state, and are common to all ensemble members. By definition, the ensemble value function maximizes its own action-value estimate over all actions:

$$V^j_{k,h}(s) = \max_{a\in\mathcal{A}} Q^j_{k,h}(s,a) \ge Q^j_{k,h}(s,\pi^\star_h(s)).$$

Hence, we obtain

$$V^j_{k,h}(s) - V^\star_h(s) \ge Q^j_{k,h}(s,\pi^\star_h(s)) - Q^\star_h(s,\pi^\star_h(s)).$$

The definition of the value-bounding operator and $B_h \ge Q^\star_h(s,\pi^\star_h(s))$ then imply

$$Q^j_{k,h}(s,\pi^\star_h(s)) - Q^\star_h(s,\pi^\star_h(s)) \ge \nu\Big(\|\phi(s,\pi^\star_h(s))\|_{\Lambda^{-1}_{k,h}}\Big)\left[\widetilde{Q}^j_{k,h}(s,\pi^\star_h(s)) - Q^\star_h(s,\pi^\star_h(s))\right].$$

Using the linear Bellman representation, the term $\widetilde{Q}^j_{k,h}(s,\pi^\star_h(s)) - Q^\star_h(s,\pi^\star_h(s))$ can be written as

$$\widetilde{Q}^j_{k,h}(s,\pi^\star_h(s)) - Q^\star_h(s,\pi^\star_h(s)) = \phi(s,\pi^\star_h(s))^\top\left[\hat w^j_{k,h} - w_h(V^j_{k,h+1})\right] + \phi(s,\pi^\star_h(s))^\top\Lambda^{-1}_{k,h}s^j_{k,h} + \big(P_h(V^j_{k,h+1} - V^\star_{h+1})\big)(s,\pi^\star_h(s)).$$

Integrating with respect to $\rho_{k,h}$ yields,

$$\begin{aligned}
&\int_{\mathcal{S}} (V^j_{k,h} - V^\star_h)(s)\,\rho_{k,h}(ds) \\
&\ge \int_{\mathcal{S}} \nu\Big(\|\phi(s,\pi^\star_h(s))\|_{\Lambda^{-1}_{k,h}}\Big)\left[\phi(s,\pi^\star_h(s))^\top\left[\hat w^j_{k,h} - w_h(V^j_{k,h+1})\right] + \phi(s,\pi^\star_h(s))^\top\Lambda^{-1}_{k,h}s^j_{k,h}\right. \\
&\quad \left. + \big(P_h(V^j_{k,h+1} - V^\star_{h+1})\big)(s,\pi^\star_h(s))\right]\rho_{k,h}(ds)\,.
\end{aligned}$$

Note that the regression and perturbation terms, the vectors $\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})$ and $\Lambda^{-1}_{k,h} s^j_{k,h}$ do not depend on the integration variable s. Linearity of integration with the definition of $\bar{\rho}_{k,h}$ therefore gives

$$\begin{aligned}
&\int_{\mathcal{S}} \phi(s, \pi^\star_h(s))^\top \left[\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\right] \bar{\rho}_{k,h}(ds) + \int_{\mathcal{S}} \phi(s, \pi^\star_h(s))^\top \Lambda^{-1}_{k,h} s^j_{k,h}\, \bar{\rho}_{k,h}(ds) \\
&= \left(\int_{\mathcal{S}} \phi(s, \pi^\star_h(s))\, \bar{\rho}_{k,h}(ds)\right)^\top \left[\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\right] + \left(\int_{\mathcal{S}} \phi(s, \pi^\star_h(s))\, \bar{\rho}_{k,h}(ds)\right)^\top \Lambda^{-1}_{k,h} s^j_{k,h} \\
&= (\phi^\star_{k,h})^\top \left[\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\right] + (\phi^\star_{k,h})^\top \Lambda^{-1}_{k,h} s^j_{k,h}\,.
\end{aligned}$$

For the term involving the value error at the next stage, expanding the transition operator yields

$$\begin{aligned}
&\int_{\mathcal{S}} \big(P_h(V^j_{k,h+1} - V^\star_{h+1})\big)(s, \pi^\star_h(s))\, \bar{\rho}_{k,h}(ds) \\
&= \int_{\mathcal{S}} \left[\int_{\mathcal{S}} (V^j_{k,h+1} - V^\star_{h+1})(s')\, P_h(ds' \mid s, \pi^\star_h(s))\right] \bar{\rho}_{k,h}(ds) \\
&= \int_{\mathcal{S}} (V^j_{k,h+1} - V^\star_{h+1})(s')\, \rho_{k,h+1}(ds')\,,
\end{aligned}$$

where the last equality follows from the definition of $\rho_{k,h+1}$ and Fubini's theorem. Combining these inequalities with Cauchy-Schwarz inequality gives,

$$\begin{aligned}
\int_{\mathcal{S}} (V^j_{k,h} - V^\star_h)(s) \rho_{k,h}(ds) &\geq -\|\phi^\star_{k,h}\|_{\Lambda^{-1}_{k,h}} \|\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\|_{\Lambda_{k,h}} + (\phi^\star_{k,h})^\top \Lambda^{-1}_{k,h} s^j_{k,h} \\
&\quad + \int_{\mathcal{S}} (V^j_{k,h+1} - V^\star_{h+1})(s') \rho_{k,h+1}(ds') \\
&\geq -\beta \|\phi^\star_{k,h}\|_{\Lambda^{-1}_{k,h}} + (\phi^\star_{k,h})^\top \Lambda^{-1}_{k,h} s^j_{k,h} \\
&\quad + \int_{\mathcal{S}} (V^j_{k,h+1} - V^\star_{h+1})(s') \rho_{k,h+1}(ds')\,,
\end{aligned}$$

where the last inequality follows from $\|\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\|_{\Lambda_{k,h}} \leq \beta$ (Lemma 6). Summing over h telescopes the integrals of the value errors at successive stages. Since $\rho_{k,1}$ assigns unit mass to $s_{k,1}$ and both terminal values are zero, we obtain

$$V^j_{k,1}(s_{k,1}) - V^\star_1(s_{k,1}) \geq \sum_{h=1}^{H} (\phi^\star_{k,h})^\top \Lambda^{-1}_{k,h} s^j_{k,h} - \beta \sum_{h=1}^{H} \|\phi^\star_{k,h}\|_{\Lambda^{-1}_{k,h}}\,. \tag{11}$$

By Eq. (11), an ensemble member is optimistic whenever its cumulative perturbation offsets the regression error. To establish constant stochastic optimism, it therefore suffices to show that this occurs for a constant fraction of the members, simultaneously in every episode.

By the Cauchy–Schwarz inequality, the regression-error term satisfies

$$\beta \sum_{h=1}^{H} \|\phi^\star_{k,h}\|_{\Lambda^{-1}_{k,h}} \leq \beta\sqrt{H} \left(\sum_{h=1}^{H} \|\phi^\star_{k,h}\|^2_{\Lambda^{-1}_{k,h}}\right)^{1/2}.$$

Since $\eta \geq 40\beta\sqrt{H}$, it is enough to prove that at least $m/10$ members satisfy

$$\sum_{h=1}^{H} (\phi^\star_{k,h})^\top \Lambda^{-1}_{k,h} s^j_{k,h} \geq \frac{\eta}{40} \left(\sum_{h=1}^{H} \|\phi^\star_{k,h}\|^2_{\Lambda^{-1}_{k,h}}\right)^{1/2}. \tag{12}$$

For these members, the right-hand side of Eq. (11) is nonnegative.

To establish Eq. (12), we use the uniform-exceedance analysis of Janz et al. (2026) for ensemble sampling in linear bandits. They show that, with a sufficiently large ensemble, a constant fraction of the persistent perturbations has a normalized projection above a positive threshold, simultaneously

at every round and in every direction. Their result allows the features to depend on past ensemble perturbations, and therefore accommodates the dependence induced by adaptive data collection. This uniformity is useful here because the directions $\phi^\star_{k,h}$ depend on the observed Gram matrices and hence on the ensemble history. We state the required result below in a predictable-design formulation of Janz et al. (2026).

Lemma 11 (Corollary 5.6 in Janz et al. (2026))**.** *Let $D \geq 1$, $\lambda \geq 1$, $n \geq \max\{2, D\}$, and $\delta' \in (0, 1/2)$. Let $g^1, \dots, g^m$ be independent $\mathcal{N}(0, I_D)$ vectors. At each round t, a feature $x_t \in \mathbb{R}^D$ with $\|x_t\|_2 \leq 1$ is determined before generating $(\xi_t^1, \dots, \xi_t^m) \sim \mathcal{N}(0, I_m)$ independently of all preceding information. The feature may depend on the initial vectors and previous observations and perturbations.*

Define

$$G_t := \lambda I_D + \sum_{\tau=1}^{t-1} x_\tau x_\tau^\top, \qquad z_t^j := \sqrt{\lambda}\, g^j + \sum_{\tau=1}^{t-1} x_\tau \xi_\tau^j.$$

If

$$m \geq 2000 D \max\left\{1, \log \frac{n}{\delta'}\right\},$$

then, with probability at least $1 - \delta'$,

$$\frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\left\{ v^\top z_t^j \geq \frac{1}{40} \|v\|_{G_t} \right\} \geq \frac{1}{10}$$

simultaneously for every $t \in [n]$ and $v \in \mathbb{R}^D \setminus \{0\}$.

Lemma 11 provides the desired lower bound on the fraction of members with sufficiently large perturbations. To apply it to Eq. (12), we express the cumulative perturbation across the H stages as a single projection of a persistent Gaussian perturbation process. Since this sum involves the same member at every stage, we stack its stage-wise perturbations into a vector in $\mathbb{R}^{dH}$.

Define

$$\mathbf{\Lambda}_k := \mathrm{diag}(\Lambda_{k,1}, \dots, \Lambda_{k,H}), \qquad \boldsymbol{s}_k^j := \begin{pmatrix} s_{k,1}^j \\ \vdots \\ s_{k,H}^j \end{pmatrix}, \qquad \boldsymbol{\phi}_k^\star := \begin{pmatrix} \phi_{k,1}^\star \\ \vdots \\ \phi_{k,H}^\star \end{pmatrix}.$$

Then

$$\sum_{h=1}^{H} (\phi_{k,h}^\star)^\top \Lambda_{k,h}^{-1} s_{k,h}^j = (\boldsymbol{\phi}_k^\star)^\top \mathbf{\Lambda}_k^{-1} \boldsymbol{s}_k^j.$$

We now verify that the normalized block perturbations have the update structure required by Lemma 11. Order the observed transitions chronologically and, for $t = (i-1)H + h$, set

$$x_t := e_h \otimes \phi_{i,h}, \qquad \bar{\xi}_t^j := \frac{\xi_{i,h}^j}{\eta}, \qquad \bar{g}^j := \frac{1}{\eta} \begin{pmatrix} g_1^j \\ \vdots \\ g_H^j \end{pmatrix}.$$

Then, the vectors $\bar{g}^j$ are independent $\mathcal{N}(0, I_{dH})$ vectors. Each feature x_t has norm at most one and is determined before the corresponding noise vector $(\bar{\xi}_t^j)_{j=1}^m \sim \mathcal{N}(0, I_m)$ is generated independently of the preceding information.

Apply Lemma 11 with its regularization parameter set to λ^2 and

$$D = dH, \qquad n = \max\{KH, dH, 2\}, \qquad \delta' = \delta/8.$$

If $n > KH$, append zero features after the actual updates. At the beginning of episode k, let $t_k := (k-1)H + 1$. The processes in Lemma 11 then satisfy

$$G_{t_k} = \lambda^2 I_{dH} + \sum_{t < t_k} x_t x_t^\top = \mathbf{\Lambda}_k + (\lambda^2 - \lambda) I_{dH} \succeq \mathbf{\Lambda}_k,$$

where the matrix inequality uses $\lambda \geq 1$ and

$$z_{t_k}^j = \lambda \bar{g}^j + \sum_{t<t_k} x_t \bar{\xi}_t^j = \frac{\boldsymbol{s}_k^j}{\eta}.$$

The assumed ensemble size therefore yields an event $\mathcal{E}_{\mathrm{exc}}$ with $\Pr(\mathcal{E}_{\mathrm{exc}}) \geq 1 - \delta/8$ on which

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\Big\{v^\top \boldsymbol{s}_k^j \geq \frac{\eta}{40}\|v\|_{G_{t_k}}\Big\} \geq \frac{1}{10}$$

simultaneously for all $k \in [K]$ and nonzero $v \in \mathbb{R}^{dH}$. Since $G_{t_k} \succeq \boldsymbol{\Lambda}_k$, this implies

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\Big\{v^\top \boldsymbol{s}_k^j \geq \frac{\eta}{40}\|v\|_{\boldsymbol{\Lambda}_k}\Big\} \geq \frac{1}{10}.$$

If $\boldsymbol{\phi}_k^\star \neq 0$, choose

$$v := \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\phi}_k^\star.$$

For this direction,

$$v^\top \boldsymbol{s}_k^j = \sum_{h=1}^{H} (\phi_{k,h}^\star)^\top \Lambda_{k,h}^{-1} s_{k,h}^j, \qquad \|v\|_{\boldsymbol{\Lambda}_k} = \left(\sum_{h=1}^{H} \|\phi_{k,h}^\star\|_{\Lambda_{k,h}^{-1}}^2\right)^{1/2}.$$

Hence, by Lemma 11 at least $m/10$ members satisfy Eq. (12). The direction may depend on the ensemble history, since the exceedance guarantee holds simultaneously for every direction. If $\boldsymbol{\phi}_k^\star = 0$, the optimistic recursion directly gives a nonnegative value difference for every member.

Consequently, on $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}} \cap \mathcal{E}_{\mathrm{exc}}$,

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\Big\{V_{k,1}^j(s_{k,1}) \geq V_1^\star(s_{k,1})\Big\} \geq \frac{1}{10}$$

simultaneously for every episode k. For each j, the value $V_{k,1}^j(s_{k,1})$ is $\mathcal{F}_k$-measurable. Since J_k is uniform on $[m]$ conditional on $\mathcal{F}_k$, we obtain

$$\Pr\Big(V_{k,1}^{J_k}(s_{k,1}) \geq V_1^\star(s_{k,1}) \,\Big|\, \mathcal{F}_k\Big) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\Big\{V_{k,1}^j(s_{k,1}) \geq V_1^\star(s_{k,1})\Big\} \geq \frac{1}{10}.$$

□

B.3.5 Proof of Lemma 3

Proof of Lemma 3. Fix k on $\mathcal{E}$. Note that conditional on $\mathcal{F}_k$, the ensemble value functions and their greedy policies are fixed. For $a = \pi_{k,h}^j(s)$, the Bellman equation and the definition of $\Delta_{k,h}^j$ give

$$\begin{aligned} V_{k,h}^j(s) - V_h^{\pi_k^j}(s) &= Q_{k,h}^j(s,a) - r_h(s,a) - (P_h V_{h+1}^{\pi_k^j})(s,a) \\ &= \Delta_{k,h}^j(s,a) + \big[P_h(V_{k,h+1}^j - V_{h+1}^{\pi_k^j})\big](s,a). \end{aligned}$$

Taking iterated conditional expectations along π_k^j cancels the continuation terms, since $V_{k,H+1}^j = V_{H+1}^{\pi_k^j} = 0$. Averaging over the uniform choice of J_k therefore gives

$$\begin{aligned} \mathbb{E}\left[V_{k,1}^{J_k}(s_{k,1}) - V_1^{\pi_k^{J_k}}(s_{k,1}) \,\middle|\, \mathcal{F}_k\right] &= \mathbb{E}\left[\sum_{h=1}^{H} \Delta_{k,h}^{J_k}(s_{k,h}, a_{k,h}) \,\middle|\, \mathcal{F}_k\right] \\ &\leq \mathbb{E}\left[\sum_{h=1}^{H} \max_{j' \in [m]} |\Delta_{k,h}^{j'}(s_{k,h}, a_{k,h})| \,\middle|\, \mathcal{F}_k\right]. \end{aligned} \tag{13}$$

Next, for any ensemble pairs (i, j), let $a = \pi^i_{k,h}(s)$. Then we obtain

$$\begin{aligned} V^i_{k,h}(s) - V^j_{k,h}(s) &\le Q^i_{k,h}(s,a) - Q^j_{k,h}(s,a) \\ &= \Delta^i_{k,h}(s,a) - \Delta^j_{k,h}(s,a) + \left[P_h(V^i_{k,h+1} - V^j_{k,h+1})\right](s,a). \end{aligned}$$

Iterating along π^i_k and using the zero terminal values bounds the initial value difference by the expected sum of $\Delta^i_{k,h} - \Delta^j_{k,h}$ as follows:

$$V^i_{k,1}(s_{k,1}) - V^j_{k,1}(s_{k,1}) \le \mathbb{E}\left[\sum_{h=1}^{H}\left(\Delta^i_{k,h}(s_{k,h},a_{k,h}) - \Delta^j_{k,h}(s_{k,h},a_{k,h})\right)\middle|\, \mathcal{F}_k, J_k = i\right].$$

Let

$$\mathcal{O}_k := \left\{i \in [m] : V^i_{k,1}(s_{k,1}) \ge V^\star_1(s_{k,1})\right\}$$

be the set of optimistic ensemble members. By Lemma 8, $|\mathcal{O}_k| \ge m/10$. Fix $j \in [m]$. Since $V^i_{k,1}(s_{k,1}) \ge V^\star_1(s_{k,1})$ for every $i \in \mathcal{O}_k$, averaging over these members and applying the pairwise bound gives

$$\begin{aligned} &V^\star_1(s_{k,1}) - V^j_{k,1}(s_{k,1}) \\ &\le \frac{1}{|\mathcal{O}_k|}\sum_{i\in\mathcal{O}_k}\left[V^i_{k,1}(s_{k,1}) - V^j_{k,1}(s_{k,1})\right] \\ &\le \frac{2}{|\mathcal{O}_k|}\sum_{i\in\mathcal{O}_k}\mathbb{E}\left[\sum_{h=1}^{H}\max_{j'\in[m]}|\Delta^{j'}_{k,h}(s_{k,h},a_{k,h})|\,\middle|\,\mathcal{F}_k, J_k = i\right] \\ &\le \frac{20}{m}\sum_{i=1}^{m}\mathbb{E}\left[\sum_{h=1}^{H}\max_{j'\in[m]}|\Delta^{j'}_{k,h}(s_{k,h},a_{k,h})|\,\middle|\,\mathcal{F}_k, J_k = i\right] \\ &= 20\mathbb{E}\left[\sum_{h=1}^{H}\max_{j'\in[m]}|\Delta^{j'}_{k,h}(s_{k,h},a_{k,h})|\,\middle|\,\mathcal{F}_k\right]. \end{aligned}$$

Here the second inequality uses $\Delta^i_{k,h} - \Delta^j_{k,h} \le 2\max_{j'}|\Delta^{j'}_{k,h}|$. The third uses nonnegativity and $|\mathcal{O}_k| \ge m/10$, and the final equality follows from the uniform conditional distribution of J_k.

Adding and subtracting $V^{J_k}_{k,1}(s_{k,1})$ and using the self-evaluation bound established above, we conclude that

$$\begin{aligned} \mathbb{E}\left[V^\star_1(s_{k,1}) - V^{\pi^{J_k}_k}_1(s_{k,1})\,\middle|\,\mathcal{F}_k\right] &= \mathbb{E}\left[V^\star_1(s_{k,1}) - V^{J_k}_{k,1}(s_{k,1})\,\middle|\,\mathcal{F}_k\right] \\ &\quad + \mathbb{E}\left[V^{J_k}_{k,1}(s_{k,1}) - V^{\pi^{J_k}_k}_1(s_{k,1})\,\middle|\,\mathcal{F}_k\right] \\ &\le 21\mathbb{E}\left[\sum_{h=1}^{H}\max_{j\in[m]}|\Delta^{j}_{k,h}(s_{k,h},a_{k,h})|\,\middle|\,\mathcal{F}_k\right]. \end{aligned}$$

□

B.3.6 Proof of Lemma 9

Proof of Lemma 9. As in the proof of Lemma 6, set

$$L := 2H\sqrt{\frac{K}{\lambda}} + \frac{\eta\Gamma_\delta}{\sqrt{\lambda}}, \qquad \epsilon := (K+1)^{-2}.$$

For each $\ell \in [H]$, define

$$q_{w,z,A,\ell}(s,a) := B_\ell + \nu(\|\phi(s,a)\|_A)\left[\phi(s,a)^\top w + |\phi(s,a)^\top z| - B_\ell\right],$$

and let

$$O_\ell^{\mathrm{abs}} := \Big\{(w,z,A) : \|w\|_2 \le 2H\sqrt{K/\lambda}, \quad \|z\|_2 \le \eta\Gamma_\delta/\sqrt{\lambda}, \quad 0 \prec A \preceq \lambda^{-1} I_d, \\ \sup_{s,a} |q_{w,z,A,\ell}(s,a) - Q_\ell^\star(s,a)| \le B_\ell\Big\}.$$

Define the deterministic value-function class

$$\mathcal{V}_\ell^{\mathrm{abs}} := \left\{ s \mapsto \max_{a\in\mathcal{A}} q_{w,z,A,\ell}(s,a) : (w,z,A) \in O_\ell^{\mathrm{abs}} \right\}, \qquad \mathcal{V}_{H+1}^{\mathrm{abs}} := \{0\}.$$

Every function in $\mathcal{V}_\ell^{\mathrm{abs}}$ takes values in $[-B_\ell, 2B_\ell]$.

For any $(w,z,A),(w',z',A') \in O_\ell^{\mathrm{abs}}$, the inequality $\big||x|-|y|\big| \le |x-y|$ and the continuity argument in the signed proof give

$$\|q_{w,z,A,\ell} - q_{w',z',A',\ell}\|_\infty \le \|w-w'\|_2 + \|z-z'\|_2 \\ + \frac{L+H}{\alpha_U-\alpha_L}\sqrt{\|A-A'\|_F}.$$

Maximization over actions preserves this bound.

Following the covering construction in the signed proof, cover the two vector balls, both contained in the Euclidean ball of radius L, at radius $\epsilon/8$. The admissible matrices lie in the Frobenius ball of radius $\sqrt{d}/\lambda$ in the space of symmetric matrices. Cover this ball at radius

$$r_A := \frac{(\alpha_U-\alpha_L)^2\epsilon^2}{16(L+H)^2}.$$

For each product of covering balls that intersects O_ℓ^{abs}, choose a deterministic representative $(\bar w,\bar z,\bar A) \in O_\ell^{\mathrm{abs}}$ from that intersection. Any admissible triple (w,z,A) in the same product is within twice the corresponding covering radius of this representative. Hence

$$\|q_{w,z,A,\ell} - q_{\bar w,\bar z,\bar A,\ell}\|_\infty \le \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{L+H}{\alpha_U-\alpha_L}\sqrt{2r_A} \\ = \left(\frac{1}{2} + \frac{1}{2\sqrt{2}}\right)\epsilon < \epsilon.$$

Therefore, the value functions corresponding to these representatives form a deterministic ϵ-net $\mathcal{C}_\ell^{\mathrm{abs}} \subset \mathcal{V}_\ell^{\mathrm{abs}}$ satisfying

$$\log|\mathcal{C}_\ell^{\mathrm{abs}}| \le 2d\log\left(1+\frac{16L}{\epsilon}\right) + \frac{d(d+1)}{2}\log\left(1+\frac{32\sqrt{d}(L+H)^2}{\lambda(\alpha_U-\alpha_L)^2\epsilon^2}\right).$$

Only the contribution of the vector covers changes, so the logarithmic covering number has the same order as in the signed case. Set $\mathcal{C}_{H+1}^{\mathrm{abs}} := \{0\}$.

For $M_{k,h}(V)$ defined in the signed proof, apply the same self-normalized inequality to each $V \in \mathcal{C}_{h+1}^{\mathrm{abs}}$, uniformly over k, with failure probability $\delta/(8H|\mathcal{C}_{h+1}^{\mathrm{abs}}|)$. A union bound over h and the net points, followed by the same ϵ-approximation argument, gives an event $\hat{\mathcal{E}}_{\mathrm{abs}}$ of probability at least $1-\delta/8$ on which

$$\|M_{k,h}(V)\|_{\Lambda_{k,h}^{-1}} \le 3H\left[d\log\left(1+\frac{K}{\lambda d}\right) + 2\log\left(\frac{8H|\mathcal{C}_{h+1}^{\mathrm{abs}}|}{\delta}\right)\right]^{1/2} + 2\sqrt{K}\,\epsilon$$

simultaneously for all $k\in[K]$, $h\in[H]$, and $V \in \mathcal{V}_{h+1}^{\mathrm{abs}}$. Here, $2\sqrt{K}\,\epsilon$ is the discretization error established in the signed proof. Since the classes and their nets are deterministic, this probability guarantee holds before imposing $\widetilde{\mathcal{E}}$ or (8).

Finally, fix k,h,j and suppose (8) holds. If $h=H$, then $V_{k,H+1}^j = 0 \in \mathcal{V}_{H+1}^{\mathrm{abs}}$. Otherwise, set $\ell = h+1$. The induction condition and the zero terminal values bound the regression targets at

stage ℓ by $2H$ in absolute value. Thus, on $\widetilde{\mathcal{E}}$, the same ridge-objective comparison as in the signed proof and the perturbation bound give

$$\|\hat{w}^j_{k,\ell}\|_2 \le 2H\sqrt{K/\lambda}, \qquad \|\Lambda^{-1}_{k,\ell} s^j_{k,\ell}\|_2 \le \eta\Gamma_\delta/\sqrt{\lambda}.$$

The absolute parameterization gives

$$Q^j_{k,\ell} = q_{\hat{w}^j_{k,\ell},\, \Lambda^{-1}_{k,\ell} s^j_{k,\ell},\, \Lambda^{-1}_{k,\ell},\, \ell}.$$

Together with the preceding norm bounds, $0 \prec \Lambda^{-1}_{k,\ell} \preceq \lambda^{-1} I_d$, and the induction condition at stage ℓ, this verifies all defining conditions of O^{abs}_ℓ. Hence $V^j_{k,h+1} \in \mathcal{V}^{\mathrm{abs}}_{h+1}$.

Therefore, on $\widetilde{\mathcal{E}} \cap \hat{\mathcal{E}}_{\mathrm{abs}}$ and under (8), the uniform concentration bound applies to $M_{k,h}(V^j_{k,h+1})$. Using the same normal-equation decomposition and the bound $\|w_h(V^j_{k,h+1})\|_2 \le 3H\sqrt{d}$ as in the signed proof, we obtain

$$\begin{aligned}
\left\|\hat{w}^j_{k,h} - w_h(V^j_{k,h+1})\right\|_{\Lambda_{k,h}} &\le \|M_{k,h}(V^j_{k,h+1})\|_{\Lambda^{-1}_{k,h}} + 3H\sqrt{\lambda d} \\
&\le 3H\left[d\log\left(1+\frac{K}{\lambda d}\right) + 2\log\left(\frac{8H|\mathcal{C}^{\mathrm{abs}}_{h+1}|}{\delta}\right)\right]^{1/2} \\
&\quad + 2\sqrt{K}\,\epsilon + 3H\sqrt{\lambda d} \\
&\le \beta,
\end{aligned}$$

where the last inequality follows from the covering-number bound, $\epsilon = (K+1)^{-2}$, and a sufficiently large universal constant in the definition of β. □

B.3.7 Proof of Lemma 10

Proof of Lemma 10. We first show that, uniformly over all episodes, stages, and state-action pairs, the average absolute perturbation across ensemble members is lower bounded as follows:

$$\frac{1}{m}\sum_{j=1}^{m}\left|\phi(s,a)^\top \Lambda^{-1}_{k,h} s^j_{k,h}\right| \ge \frac{\eta}{400}\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}}. \tag{14}$$

This lower bound allows the average perturbation to offset the regression error when $\eta \ge 400\beta$, which we then use to propagate optimism through backward induction.

Recall that

$$s^j_{k,h} = \lambda g^j_h + \sum_{i<k}\phi_{i,h}\xi^j_{i,h}, \qquad g^j_h \sim \mathcal{N}(0,\eta^2 I_d), \qquad \xi^j_{i,h} \sim \mathcal{N}(0,\eta^2).$$

For each fixed stage h, define

$$\bar{g}^j_h := g^j_h/\eta, \qquad \bar{\xi}^j_{i,h} := \xi^j_{i,h}/\eta,$$

and

$$G_{k,h} := \lambda^2 I_d + \sum_{i<k}\phi_{i,h}\phi^\top_{i,h}, \qquad \bar{s}^j_{k,h} := s^j_{k,h}/\eta = \lambda\bar{g}^j_h + \sum_{i<k}\phi_{i,h}\bar{\xi}^j_{i,h}.$$

The initial vectors $\bar{g}^j_h$ are independent $\mathcal{N}(0, I_d)$ vectors. Each feature $\phi_{i,h}$ has norm at most one and is determined before $(\bar{\xi}^j_{i,h})_{j\in[m]}$ is generated. Conditional on all information preceding its generation, this noise vector has distribution $\mathcal{N}(0, I_m)$. Thus, the updates are

$$G_{k+1,h} = G_{k,h} + \phi_{k,h}\phi^\top_{k,h}, \qquad \bar{s}^j_{k+1,h} = \bar{s}^j_{k,h} + \phi_{k,h}\bar{\xi}^j_{k,h},$$

with initial values $G_{1,h} = \lambda^2 I_d$ and $\bar{s}^j_{1,h} = \lambda\bar{g}^j_h$.

Apply Lemma 11 separately to each stage with dimension $D = d$, regularization parameter $\lambda^2 \ge 1$, horizon n, and failure probability $\delta/(8H)$, appending zero features if needed. The assumed

ensemble size and a union bound over $h \in [H]$ give an event $\mathcal{E}_{\text{exc}}^{\text{abs}}$ of probability at least $1 - \delta/8$ on which

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left\{ v^\top \bar{s}_{k,h}^j \geq \frac{1}{40}\|v\|_{G_{k,h}} \right\} \geq \frac{1}{10}$$

simultaneously for all $k \in [K]$, $h \in [H]$, and nonzero $v \in \mathbb{R}^d$.

Since $\lambda \geq 1$,

$$G_{k,h} = \Lambda_{k,h} + (\lambda^2 - \lambda) I_d \succeq \Lambda_{k,h}.$$

Together with $s_{k,h}^j = \eta \bar{s}_{k,h}^j$, this implies

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left\{ v^\top s_{k,h}^j \geq \frac{\eta}{40}\|v\|_{\Lambda_{k,h}} \right\} \geq \frac{1}{10}.$$

Since absolute projections are nonnegative,

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m} |v^\top s_{k,h}^j| &\geq \frac{\eta\|v\|_{\Lambda_{k,h}}}{40m}\sum_{j=1}^{m} \mathbb{1}\left\{ v^\top s_{k,h}^j \geq \frac{\eta}{40}\|v\|_{\Lambda_{k,h}} \right\} \\ &\geq \frac{\eta}{400}\|v\|_{\Lambda_{k,h}}.\end{aligned}$$

The same lower bound holds trivially for $v = 0$. Choosing $v = \Lambda_{k,h}^{-1}\phi(s,a)$ yields

$$\frac{1}{m}\sum_{j=1}^{m} \left| \phi(s,a)^\top \Lambda_{k,h}^{-1} s_{k,h}^j \right| \geq \frac{\eta}{400}\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}},$$

simultaneously for every k, h, s, a. This data-dependent choice of direction is valid because the exceedance guarantee is uniform over v.

Fix k and work on $\mathcal{E}_{\text{abs}}$. Under the threshold condition, the regression concentration and value-boundedness arguments in Lemma 9 give

$$\left\| \hat{w}_{k,h}^j - w_h(V_{k,h+1}^j) \right\|_{\Lambda_{k,h}} \leq \beta \qquad \forall h \in [H],\ j \in [m].$$

We now prove Eq. (9) by backward induction. The terminal values satisfy $V_{k,H+1}^j = V_{H+1}^\star = 0$. Suppose that

$$\frac{1}{m}\sum_{j=1}^{m} V_{k,h+1}^j(s') \geq V_{h+1}^\star(s') \qquad \forall s' \in \mathcal{S}\,.$$

For every (s,a) and j, the regression bound and the linear Bellman representation imply

$$\begin{aligned}\phi(s,a)^\top \hat{w}_{k,h}^j &= \phi(s,a)^\top \left(\hat{w}_{k,h}^j - w_h(V_{k,h+1}^j)\right) + \phi(s,a)^\top w_h(V_{k,h+1}^j) \\ &\geq -\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}} \|\hat{w}_{k,h}^j - w_h(V_{k,h+1}^j)\|_{\Lambda_{k,h}} + r_h(s,a) + (P_h V_{k,h+1}^j)(s,a) \\ &\geq -\beta\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}} + r_h(s,a) + (P_h V_{k,h+1}^j)(s,a)\,. \end{aligned} \tag{15}$$

Recall that the raw action-value estimate is

$$\widetilde{Q}_{k,h}^j(s,a) = \phi(s,a)^\top \hat{w}_{k,h}^j + \left| \phi(s,a)^\top \Lambda_{k,h}^{-1} s_{k,h}^j \right|.$$

Averaging the raw action-value estimates over j and applying Eqs. (14) and (15), we obtain

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m} \widetilde{Q}_{k,h}^j(s,a) &= \frac{1}{m}\sum_{j=1}^{m} \phi(s,a)^\top \hat{w}_{k,h}^j + \frac{1}{m}\sum_{j=1}^{m} \left| \phi(s,a)^\top \Lambda_{k,h}^{-1} s_{k,h}^j \right| \\ &\geq r_h(s,a) + \frac{1}{m}\sum_{j=1}^{m} (P_h V_{k,h+1}^j)(s,a) - \beta\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}} + \frac{\eta}{400}\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}} \\ &= r_h(s,a) + \left[P_h \left(\frac{1}{m}\sum_{j=1}^{m} V_{k,h+1}^j \right) \right](s,a) + \left(\frac{\eta}{400} - \beta \right) \|\phi(s,a)\|_{\Lambda_{k,h}^{-1}},\end{aligned}$$

where the equality follows from the linearity of P_h. The condition $\eta \geq 400\beta$ ensures that the final term is nonnegative.

Moreover, the induction hypothesis holds pointwise at the next stage, so the monotonicity of P_h implies

$$\left[P_h\left(\frac{1}{m}\sum_{j=1}^{m} V_{k,h+1}^{j}\right)\right](s,a) \geq (P_h V_{h+1}^{\star})(s,a).$$

Combining these observations with the optimal Bellman equation yields

$$\frac{1}{m}\sum_{j=1}^{m} \widetilde{Q}_{k,h}^{j}(s,a) \geq r_h(s,a) + (P_h V_{h+1}^{\star})(s,a) = Q_h^{\star}(s,a).$$

Note that the interpolation weight $\nu(\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}})$ depends only on the feature vector and the shared Gram matrix, and is therefore identical across ensemble members. Averaging the value-bounding operator thus gives

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m} Q_{k,h}^{j}(s,a) &= \nu\Big(\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}}\Big)\,\frac{1}{m}\sum_{j=1}^{m}\widetilde{Q}_{k,h}^{j}(s,a) + \Big[1-\nu\Big(\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}}\Big)\Big]\,B_h \\ &\geq Q_h^{\star}(s,a).\end{aligned}$$

The inequality follows because both the average raw estimate and B_h are at least $Q_h^{\star}(s,a)$, and the interpolation weight lies in $[0,1]$. Hence, their convex combination preserves optimism.

To obtain the corresponding bound for the value functions, note that, for any fixed action a,

$$\frac{1}{m}\sum_{j=1}^{m}\max_{a'\in\mathcal{A}} Q_{k,h}^{j}(s,a') \geq \frac{1}{m}\sum_{j=1}^{m} Q_{k,h}^{j}(s,a).$$

Since this holds for every a, we may maximize the right-hand side over actions. Combining this observation with the bound on the average action-value estimate yields

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m} V_{k,h}^{j}(s) &= \frac{1}{m}\sum_{j=1}^{m}\max_{a\in\mathcal{A}} Q_{k,h}^{j}(s,a) \\ &\geq \max_{a\in\mathcal{A}}\frac{1}{m}\sum_{j=1}^{m} Q_{k,h}^{j}(s,a) \\ &\geq \max_{a\in\mathcal{A}} Q_h^{\star}(s,a) = V_h^{\star}(s).\end{aligned}$$

Both bounds hold at every state and action at stage h, which completes the backward induction.

Finally, the ensemble value functions and $s_{k,1}$ are $\mathcal{F}_k$-measurable, and J_k is uniform conditional on $\mathcal{F}_k$. Thus

$$\mathbb{E}\Big[V_{k,1}^{J_k}(s_{k,1})\,\Big|\,\mathcal{F}_k\Big] = \frac{1}{m}\sum_{j=1}^{m} V_{k,1}^{j}(s_{k,1}) \geq V_1^{\star}(s_{k,1}).$$

□

C Auxiliary Lemmas

Lemma 12 (Lemma 7 in Lee and Oh (2024)). *If $Z \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ is a d-dimensional multivariate Gaussian vector, then for any $\delta \in (0,1]$,*

$$\mathbb{P}\left(\|Z\|_2 \geq \sqrt{d} + \sqrt{2\log\frac{1}{\delta}}\right) \leq \delta\,.$$

Lemma 13 (Theorem 1 in Abbasi-Yadkori et al. (2011)). *Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is F_t-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$. Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is F_{t-1}-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$, define*

$$\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top, \qquad S_t = \sum_{s=1}^{t} \eta_s X_s.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$,

$$\|S_t\|^2_{\overline{V}_t^{-1}} \leq 2R^2 \log\left(\frac{\det(\overline{V}_t)^{1/2} \det(V)^{-1/2}}{\delta}\right).$$

Lemma 14 (Lemma 10 in Abbasi-Yadkori et al. (2011)). *Suppose $X_1, \ldots, X_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$, $\|X_s\|_2 \leq L$. Let $\bar{V}_t = \lambda I_d + \sum_{s=1}^{t} X_s X_s^\top$ for some $\lambda > 0$. Then,*

$$\det \bar{V}_t \leq (\lambda + tL^2/d)^d.$$

Lemma 15 (Lemma 11 in Abbasi-Yadkori et al. (2011)). *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$, V a $d \times d$ positive definite matrix and define $\bar{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top$. If $\|X_t\|_2 \leq L$ for all t and $\lambda_{\min}(V) \geq \max\{1, L^2\}$, then*

$$\sum_{t=1}^{n} \|X_t\|^2_{\bar{V}_{t-1}^{-1}} \leq 2 \log \frac{\det \bar{V}_n}{\det V}.$$

Lemma 16 (Lemma D.5 in (Jin et al., 2020)). *For any $\varepsilon > 0$, the ε-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\varepsilon)^d$*

D Experiment Details

RiverSwim Environment. We evaluate the proposed algorithms on a finite-horizon RiverSwim task with $n = 12$ states and horizon $H = 40$ (Figure 3). Each episode starts at the left end of a chain, where the agent chooses between moving left and attempting to move right. Leftward actions move deterministically toward the starting state, whereas rightward actions have stochastic outcomes that can delay or reverse progress. Taking the left action at the leftmost state yields a small reward of 0.005, while taking the right action at the rightmost state yields a reward of 1; all other state–action pairs give zero reward. This creates an exploration challenge: the agent must forgo an easily accessible reward to discover a more rewarding region farther along the chain. We use one-hot state–action features, giving a feature dimension of $d = |\mathcal{S}| \times |\mathcal{A}| = 24$.

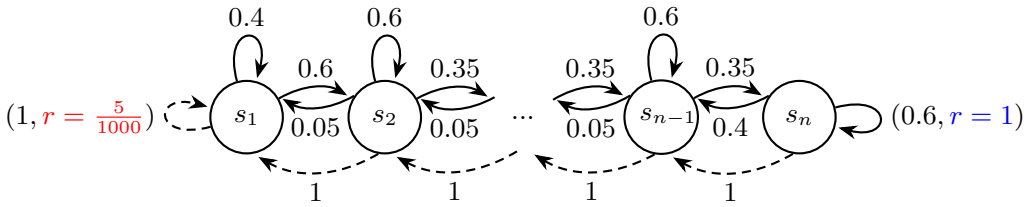


Figure 3: The "RiverSwim" environment with n states (Osband et al., 2013)

We compare the signed and absolute variants of **`ES-LSVI`**, each with ensemble size $m = 4$, against **`LSVI-UCB`** (Jin et al., 2020), **`LSVI-PHE`** (Ishfaq et al., 2021), **`OPT-RLSVI`** (Zanette et al., 2020), **`LMC-LSVI`** (Ishfaq et al., 2024a), and **`BooVI`** (Liu et al., 2021). Hyperparameters are selected through grid search on three validation seeds, using average return over the training run as the selection criterion. We then evaluate the selected configurations for 20,000 episodes across five separate seeds and report mean episodic returns with 100-episode moving-average smoothing. An oracle policy computed by backward induction provides a reference for optimal performance.

Figure 4 shows that both variants of **`ES-LSVI`** rapidly learn high-return policies and approach the oracle performance during the early stage of training. As highlighted by the first 1,000 episodes, the

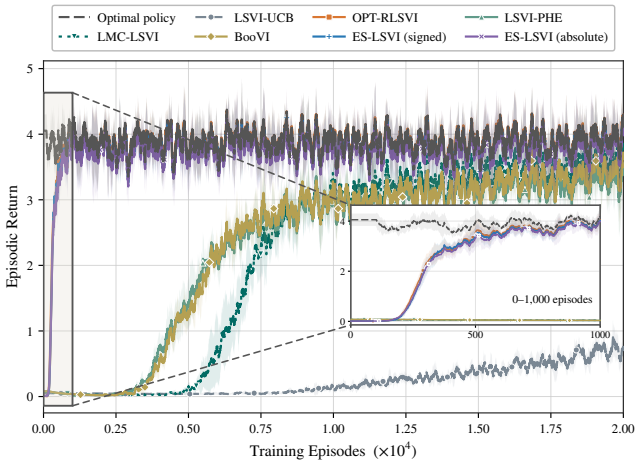
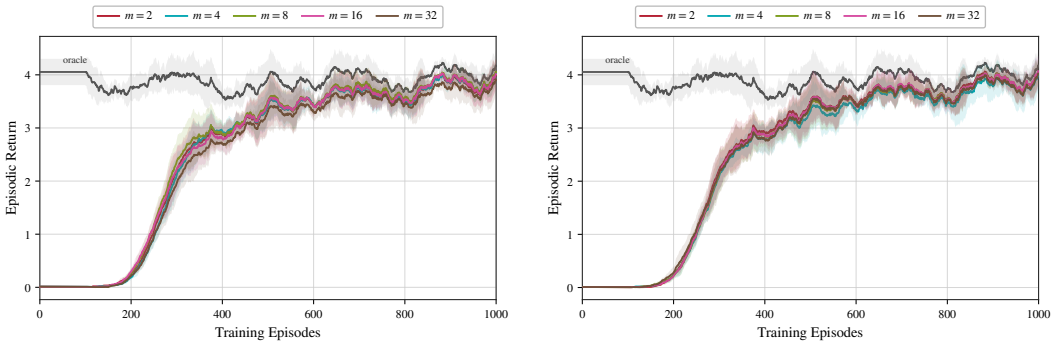


Figure 4: Performance on the 12-state RiverSwim environment over 20,000 training episodes. Curves show mean episodic returns over five seeds with 100-episode moving-average smoothing, and shaded regions indicate variability across seeds.

signed and absolute variants improve substantially earlier than several of the competing exploration methods and maintain high returns throughout training. The two variants exhibit similar learning behavior on this task, despite relying on different perturbation constructions and having different theoretical guarantees. These results indicate that persistent perturbations combined with uniform ensemble-member selection can provide effective exploration in RiverSwim without repeatedly resampling the perturbations.



(a) Episodic returns of the signed mode with different ensemble sizes.

(b) Episodic returns of the absolute mode with different ensemble sizes.

Figure 5: **Effect of ensemble size on `ES-LSVI`.** Episodic returns of (a) the signed mode and (b) the absolute mode for different ensemble sizes $m \in \{2, 4, 8, 16, 32\}$. Both modes achieve similar learning performance across the tested ensemble sizes, indicating limited sensitivity to the ensemble size in this environment. The black curve denotes the oracle performance.

We further examine the effect of the ensemble size in Figure 5. For both the signed and absolute modes, the learning curves remain similar across $m \in \{2, 4, 8, 16, 32\}$, with all tested ensemble sizes eventually approaching the oracle performance. In particular, even very small ensembles achieve comparable learning behavior in this environment.

MinAtar Environments. We provide additional implementation details and numerical results for the MinAtar experiments. All methods use a common convolutional feature-extractor architecture and the same replay-buffer and target-network settings. Our neural implementations retain persistent reward perturbations and uniform member selection at the beginning of each episode, while replacing exact least-squares regression with minibatch gradient updates. Unlike the selected-member fitting in Algorithm 1, the neural implementation updates all ensemble heads jointly.

Table 2: Hyperparameter values considered for **ES-LSVI** in the RiverSwim experiments.

Hyperparameter	Values
noise scale η	$\{0.001, 0.002, 0.005, 0.01, 0.05, 0.1\}$
regularization λ	$\{0.1, 1, 10\}$
ensemble size m	$\{2, 4, 8, 16, 32\}$
cutoff parameters (α_L, α_U)	$\{(0.15, 0.45), (0.3, 0.9), (0.45, 0.9)\}$

Table 3 lists the hyperparameter values considered for **ES-LSVI** across the MinAtar experiments. We tune the standard deviations of the initial and reward perturbations, σ_I and σ_R, the regularization coefficient λ, and the Adam learning rate. The ensemble size is fixed at $m = 10$ for the main comparison, and its effect is examined separately in the ablation below.

Table 3: Hyperparameter values considered for **ES-LSVI** in the MinAtar experiments.

Hyperparameter	Values
Initial perturbation std. σ_I	$\{0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 1\}$
Reward perturbation std. σ_R	$\{0.001, 0.003, 0.01, 0.03, 0.1, 0.3\}$
Regularization λ	$\{100, 1000, 10000\}$
Adam learning rate	$\{10^{-4}, 3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$
Ensemble size m	$\{10\}$

Table 4 reports the raw returns corresponding to the final comparison in Figure 1. For each seed, we average the returns from the five evaluations between four and five million environment steps. We then report the mean and standard error of these averages across the five seeds.

Table 4: Final-window returns on the five MinAtar games. Results are mean $\pm$ standard error over five random seeds.

Method	Breakout	Seaquest	Space Invaders	Asterix	Freeway
Bootstrapped DQN	20.00 ± 3.40	64.08 ± 13.46	225.92 ± 33.60	34.44 ± 4.42	62.20 ± 0.78
NoisyNet DQN	18.72 ± 0.55	83.62 ± 2.98	$\mathbf{230.83 \pm 13.34}$	44.98 ± 2.54	50.35 ± 12.60
Adam LMCDQN	$\mathbf{28.75 \pm 6.12}$	71.30 ± 18.86	199.48 ± 9.32	47.68 ± 2.69	$\mathbf{64.93 \pm 0.31}$
FG-ULMCDQN	20.63 ± 2.68	52.84 ± 15.74	115.02 ± 29.40	39.20 ± 9.33	43.06 ± 11.88
ES-LSVI (signed)	21.77 ± 1.18	74.40 ± 10.47	185.66 ± 13.89	$\mathbf{56.55 \pm 3.34}$	61.94 ± 0.56
ES-LSVI (absolute)	18.60 ± 0.85	$\mathbf{92.88 \pm 8.98}$	197.41 ± 21.88	49.90 ± 3.68	59.98 ± 1.93

The signed variant of **ES-LSVI** achieves the highest mean return on Asterix, while the absolute variant achieves the highest mean on Seaquest. Across the remaining games, the signed variant has higher mean returns than the absolute variant on Breakout and Freeway, whereas the absolute variant has a higher mean return on Space Invaders. These results are consistent with the learning curves in the main text and show that the relative performance of the two perturbation schemes varies across games.

We further examine the effect of the ensemble size on Seaquest. Figure 6 compares the learning curves of **ES-LSVI** with different numbers of ensemble members. This experiment complements the main comparison with $m = 10$ by examining the sensitivity of learning performance to the ensemble size under neural value approximation.

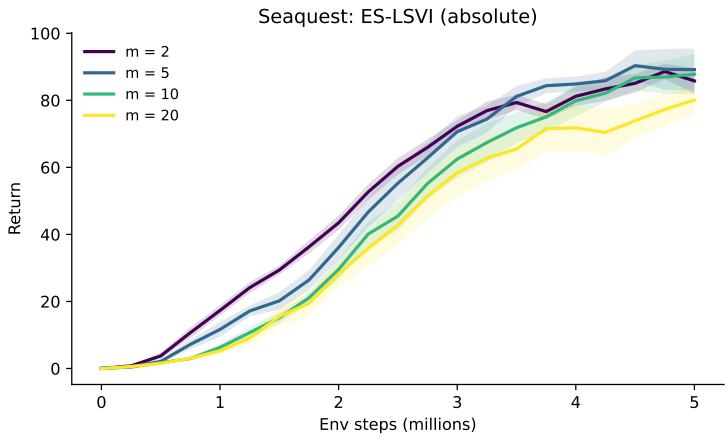


Figure 6: Effect of ensemble size on **`ES-LSVI`** in Seaquest. Solid lines show mean returns over five random seeds, and shaded regions indicate the standard error.