Optimal Regret Bounds via Low-Rank Structured Variation in Non-Stationary Reinforcement Learning

Tuan Dam
Hanoi University of Science and Technology, Hanoi, Vietnam
tuandq@soict.hust.edu.vn

Abstract

We study reinforcement learning in non-stationary communicating MDPs whose transition drift admits a low-rank plus sparse structure. We propose **SVUCRL** (Structured Variation UCRL) and prove the dynamic-regret bound

$$\widetilde{\mathcal{O}}\big(D_{\max} S\sqrt{AT} + D_{\max}\sqrt{(B_r + B_p)KST} + D_{\max}\,\delta_B\,B_p\big).$$

where S is the number of states, A the number of actions, T the horizon, $D_{\max}$ the MDP diameter, B_r/B_p the total reward/transition variation budgets, and $K \ll SA$ the rank of the structured drift. The first term is the statistical price of learning in stationary problems; the second is the *non-stationarity price*, which scales with $\sqrt{K}$ rather than $\sqrt{SA}$ when drift is low-rank. This matches the $\sqrt{T}$ rate (up to logs) and improves on prior $T^{3/4}$-type guarantees. SVUCRL combines: (i) online low-rank tracking with explicit Frobenius guarantees, (ii) incremental RPCA to separate structured drift from sparse shocks, (iii) adaptive confidence widening via a bias-corrected local-variation estimator, and (iv) factor forecasting with an optimal shrinkage center.

1 Introduction

Reinforcement learning (RL) algorithms have achieved remarkable success in stationary environments with fixed reward distributions and state transition dynamics. However, many real-world applications involve non-stationary environments where dynamics evolve due to changing user preferences, environmental conditions, or system parameters. This non-stationarity poses significant challenges for traditional RL approaches that assume fixed environment dynamics.

Non-stationary RL faces environments whose reward and transition laws evolve in complex ways. Standard approaches use sliding-window techniques that focus on recent observations while discarding older data. Algorithms such as SWUCRL2–CW [6] widen confidence sets to cover temporal drift, providing theoretical guarantees at the cost of higher regret.

These existing approaches have a significant limitation: they use uniform widening parameters that ignore the *structure* of environmental evolution. In many real systems, however, drift exhibits exploitable patterns: it often lives in low-dimensional subspaces ($K \ll SA$) where only a few underlying factors drive changes; the changes follow smooth trajectories enabling short-term forecasting; and many environments exhibit structured evolution with occasional sparse shocks affecting only small subsets of state-action pairs.

This paper leverages these observations to develop SVUCRL (Structured Variation UCRL), which combines matrix factorization, robust statistics, and time-series analysis to achieve improved regret bounds and computational efficiency.

Our contributions:

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1. **Structured variation model and δ_B.** We model drift as low-rank plus a sparse component whose ℓ_1-mass consumes at most a δ_B-fraction of the transition-variation budget B_p. This separates the learnable, shared dynamics from idiosyncratic shocks (Assumption 1).
2. **Provable low-rank tracking.** A power–Frobenius inequality and randomized SVD with power iterations control the streaming approximation error with explicit constants (Lemmas 1, 2).
3. **Shock isolation.** An incremental RPCA update (Algorithm 2) separates the structured low–rank drift from sparse shocks and provides a per-step reconstruction bound, together with its per-update cost.(Proposition 1).
4. **Forecast–shrinkage center.** Forecasted factors define a low-variance center that is combined with empirical transitions via a James–Stein weight; the data-driven weight is asymptotically optimal (Theorem 1).
5. **Main regret bound.** Summing per-step bounds under the above controls yields the three-term dynamic-regret guarantee, matching $\sqrt{T}$ rates (Theorem 2).

The resulting algorithm, SVUCRL, enjoys the regret bound of Theorem 2 and is computationally $\mathcal{O}\big(TSA(SK+S)\log T\big)$.

Notation $S = |\mathcal{S}|$ denotes the size of the state space, A is the average number of actions, $D_{\max}$ the diameter of the MDP, B_r, B_p the reward/transition variation budgets, and $\|\cdot\|_{1,\infty}$ the maximum row ℓ_1 norm. Throughout $\widetilde{\mathcal{O}}$ hides $\mathrm{polylog}(T, S, A)$ factors.

Related Work Non-stationary reinforcement learning has been studied under various modeling assumptions. The sliding window approach has been explored extensively [6, 8], with algorithms that discard data outside a recent window. Change-point detection methods [21] attempt to identify significant shifts in environment dynamics. Bandit-based approaches [4, 22] use various weighting schemes to prioritize recent observations. Our work is most closely related to the confidence widening approach in SWUCRL2-CW [6], but we significantly improve upon it by exploiting structure in the variation. Our structured variation model bears some similarity to factored MDPs [16, 20], but we focus on the structure of *changes* rather than the structure of the MDP itself. The matrix decomposition techniques we employ relate to robust PCA [5] and online matrix factorization [17], but we adapt these methods to the specific challenges of sequential decision-making under non-stationarity. Our adaptive confidence widening connects to adaptive concentration inequalities in statistics [12, 13]. Beyond classical factored MDPs [16, 20], recent work explores low-rank structure for sample-efficient control and representation learning, e.g., low-rank MDPs with continuous actions [3] and model-based methods that exploit low-rank structure [1]. Our setting differs by allowing *time-varying* dynamics with a low-rank drift plus sparse shocks, and our analysis quantifies how exploiting this structure improves dynamic-regret rates.

Our $\tilde{O}(\sqrt{T})$ dependence does not contradict known lower bounds in the *unstructured* non-stationary setting: for communicating MDPs, Mao et al. [18] show any algorithm suffers at least $\Omega\big((B_r+B_p)^{1/3}T^{2/3}\big)$ when no structure is assumed. By contrast, our improvement *leverages* low-rank drift and sparse shocks. Unless otherwise stated, all comparisons in this paper are made under Assumption 1 (low-rank drift plus sparse shocks); in the same regime, SWUCRL2-CW [6] is the most relevant baseline.

2 Problem set-up

We study reinforcement learning in non-stationary environments, formalized as a sequence $\big(\mathcal{S}, \mathcal{A}, p_t, r_t\big)_{t=1}^{T}$ of communicating Markov Decision Processes (MDPs) with diameter at most $D_{\max}$.

MDP Sequence Each MDP M_t in the sequence shares the same state space $\mathcal{S}$ and action space $\mathcal{A}$ but has potentially different transition dynamics p_t and reward functions r_t at each time step t. The transition function $p_t(s'|s,a)$ specifies the probability of transitioning to state s' when taking action a in state s at time t. Similarly, the reward function $r_t(s,a)$ represents the expected reward for taking action a in state s at time t.

Communicating MDPs We assume that each MDP in the sequence is communicating, meaning that for any pair of states $s, s' \in \mathcal{S}$, there exists a policy that reaches s' from s with non-zero probability. The diameter $D_{\max}$ quantifies the worst-case expected time to navigate between any two states, providing a measure of the connectivity of the MDP.

Variation Budgets To quantify the degree of non-stationarity, we define variation budgets for both rewards and transitions: $B_r = \sum_t \max_{s,a} |r_{t+1}(s,a) - r_t(s,a)|$, $B_p = \sum_t \max_{s,a} \|p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a)\|_1$. The reward variation budget B_r measures the cumulative maximum change in rewards across all state-action pairs, while the transition variation budget B_p captures the cumulative maximum change in transition probabilities measured in ℓ_1 norm. These budgets provide a formal way to bound the total amount of non-stationarity in the environment.

Learning Protocol The learning process proceeds as follows: at each time step t, the agent observes the current state s_t, selects an action a_t based on its policy, and the environment generates a reward $r_t(s_t, a_t)$ and transitions to the next state s_{t+1} according to $p_t(\cdot|s_t, a_t)$. The agent then updates its policy based on the observation (s_t, a_t, r_t, s_{t+1}).

Dynamic regret The performance of a learning algorithm is measured by its dynamic regret, defined as:

$$\mathrm{DynReg}_T = \sum_{t=1}^{T} \big(\rho_t^* - r_t(s_t, a_t)\big)$$

where ρ_t^* is the optimal average reward for MDP t (achievable by an oracle that knows the dynamics of MDP t in advance) and define as

$$\rho_t^* \;:=\; \sup_{\pi} \lim_{T\to\infty} \frac{1}{T}\, \mathbb{E}_{\pi}^{M_t} \left[\sum_{i=0}^{T-1} r_t(s_i, a_i) \right],$$

where $a_i \sim \pi(\cdot \mid s_i)$ and $s_{i+1} \sim p_t(\cdot \mid s_i, a_i)$. The dynamic regret measures the cumulative difference between the reward obtained by the algorithm and the reward that could have been obtained by an optimal policy for each MDP in the sequence. This is a more challenging metric than the static regret often used in stationary environments, as it requires the algorithm to track the changing optimal policy over time.

Challenges Non-stationary RL presents several key challenges. The exploration-exploitation tradeoff requires the agent to balance exploring to learn the changing dynamics with exploiting current knowledge to maximize reward. The agent must also demonstrate adaptivity by adapting quickly to changes in the environment without discarding too much relevant historical data. Additionally, computational efficiency is crucial as processing the continuous stream of observations requires efficient algorithms, especially for large state and action spaces.

Our approach addresses these challenges by exploiting structure in the environmental changes, allowing for more efficient learning and better adaptation to the evolving dynamics.

3 Structured variation model

The core insight of our approach is that changes in the environment dynamics often exhibit structure that can be exploited for more efficient learning. We formalize this intuition in our structured variation model. Most of the *change* in the dynamics from step t to $t+1$ can be explained by a few latent drivers, plus occasional localized shocks.

Assumption 1 (Low-rank drift)**.** For each t, the transition change $\Delta P_t := P_{t+1} - P_t \in \mathbb{R}^{SA\times S}$ admits the decomposition

$$\Delta P_t \;=\; \sum_{k=1}^{K} u_k(t)\, \underbrace{v_k}_{\in\mathbb{R}^{SA}}\, \underbrace{w_k^{\top}}_{\in\mathbb{R}^{S}} + \epsilon_t, \qquad \sum_t \max_{s,a} \|\epsilon_t(s,a,\cdot)\|_1 \;\le\; \delta_B\, B_p,$$

with per-factor bounds $\|w_k\|_1 \le 1$ and $|v_k(s,a)| \le 1$ for all (s,a).

In this assumption:

- $u_k(t)$ is the *time weight* of factor k at step t.
- v_k is a pattern over state–action rows (s, a) saying *which* parts of the MDP are affected by factor k.
- w_k is a pattern over next states s' saying *how* probability mass is reallocated when factor k acts.
- ϵ_t is a *sparse shock* capturing rare, localized, hard-to-predict changes.

Why the constraints matter. The simple bounds $|v_k(s, a)| \leq 1$ and $\|w_k\|_1 \leq 1$ are a convenient *scaling convention*: they push the overall magnitude of each factor into $u_k(t)$. This makes the per-step change interpretable and ensures that $\max_{s,a} \| \sum_k u_k(t) v_k(s, a) w_k \|_1$ is directly controlled by $\sum_k |u_k(t)|$. The budget $\sum_t \max_{s,a} \|\epsilon_t(s, a, \cdot)\|_1 \leq \delta_B B_p$ says that shocks consume at most a δ_B–fraction of the total transition variation B_p, so most drift is structured.

What the model buys you. Low rank means *shared structure across rows*: many (s, a)-rows move in a correlated way (via the same w_k pattern), at amplitudes set by $v_k(s, a)$, and with time profiles $u_k(t)$. When $K \ll SA$, SVUCRL can track these few drivers instead of estimating all SA rows independently, enabling tighter confidence sets and $\sqrt{(B_r+B_p)KST}$ dependence in the non-stationary term of the regret bound.

This model is motivated by several observations about real-world systems:

- *Traffic/control:* a weather factor shifts many transitions in the same direction (wet roads); v_k highlights the affected links, w_k encodes where mass moves (slower lanes), $u_k(t)$ follows the storm's intensity.
- *Recommendation:* a global popularity wave reweights next-state preferences for many user contexts in tandem, with occasional item-specific shocks handled by ϵ_t.

How it relates to the variation budgets. Recall $B_p = \sum_t \max_{s,a} \|P_{t+1}(\cdot|s, a) - P_t(\cdot|s, a)\|_1$. Under Assumption 1, the *structured* part of each row change is $\sum_k u_k(t) v_k(s, a) w_k$, whose ℓ_1-size per row is at most $\sum_k |u_k(t)|$ by the bounds on v_k, w_k. Hence the same few coefficients $u_k(t)$ concurrently govern the row-wise maxima that enter B_p, this is the leverage SVUCRL exploits.

Sanity checks.

- *Stationary case:* if the environment is stationary, then $\Delta P_t \equiv 0$ and one can take $K = 0$, $\epsilon_t \equiv 0$.
- *Many weak drivers:* as K grows or δ_B increases, the advantage diminishes smoothly; the model reduces to general variation when K approaches SA or shocks dominate.

No need to know K a priori. SVUCRL *estimates* an effective rank online (via randomized SVD with oversampling and power iterations), and updates it as the spectrum evolves; the algorithm and guarantees do not require the true K as input.

What this model is *not*. We do *not* assume the MDP itself is factored; only the *drift* ΔP_t is approximately low rank plus sparse. This distinction lets us capture global but compact changes even when the underlying P_t has no simple structure.

4 Online low-rank approximation

A key component of our approach is efficiently tracking the low-rank structure of environmental changes as they evolve over time. This section develops the theoretical and algorithmic foundations for this tracking process.

To represent the changes in transition dynamics, we flatten the state-action pairs (s, a) into rows and the next-state indices into columns, forming matrices. We denote by $\mathbf{X}_t = [\Delta P_{t-W+1}, \ldots, \Delta P_t] \in \mathbb{R}^{SA \times WS}$ the matrix containing the last W changes in transition probabilities.

Algorithm 1 Randomised SVD with power iterations

Require: matrix $\mathbf{X}$, target rank $\widehat{K}$, oversampling s, iters q

$\Omega \leftarrow \mathcal{N}(0,1)^{WS\times(\widehat{K}+s)}$
$\mathbf{Y}_0 \leftarrow \mathbf{X}\Omega$
for $j = 1$ to q **do**
  $\mathbf{Y}_j \leftarrow \mathbf{X}\big(\mathbf{X}^\top \mathbf{Y}_{j-1}\big)$
end for
$\mathbf{Q} \leftarrow \mathrm{qr}(\mathbf{Y}_q)$
$\mathbf{B} \leftarrow \mathbf{Q}^\top \mathbf{X}$
$\mathbf{U}_B, \Sigma, \mathbf{V} \leftarrow \mathrm{svd}(\mathbf{B})$
$\mathbf{U} \leftarrow \mathbf{Q}\mathbf{U}_B$
return $(\mathbf{U}_{[:,1:\widehat{K}]}, \Sigma_{1:\widehat{K}}, \mathbf{V}_{[:,1:\widehat{K}]})$

4.1 A Frobenius power-iteration bound

We begin by establishing a theoretical result that relates the Frobenius norm error of a low-rank approximation to that of a power-iterated version of the matrix. This result is crucial for the theoretical guarantees of our randomized SVD algorithm.

Lemma 1 (Power–Frobenius)**.** Let $X = U\Sigma V^\top$ be the singular value decomposition of a (real) matrix X, and let

$$B \;:=\; (XX^\top)^q X \qquad \text{for an integer } q \geq 0.$$

For any rank-$\widehat{K}$ orthogonal projector P, define $m := \mathrm{rank}(X) - \widehat{K}$ (the tail dimension). Then

$$\big\|(I-P)X\big\|_F \;\leq\; m^{\frac{q}{2q+1}}\, \big\|(I-P)B\big\|_F^{\frac{1}{2q+1}}.$$

In particular, if $\mathrm{rank}(X) \leq 2\widehat{K}+1$ (so $m \leq \widehat{K}+1$), then

$$\big\|(I-P)X\big\|_F \;\leq\; (\widehat{K}+1)^{\frac{q}{2q+1}}\, \big\|(I-P)B\big\|_F^{\frac{1}{2q+1}}.$$

This lemma establishes that the error of a rank-$\widehat{K}$ approximation of $\mathbf{X}$ is related to the error of approximating the power-iterated matrix $\mathbf{B}$. Intuitively, power iteration amplifies the gap between the top $\widehat{K}$ singular values and the remaining ones, making it easier to identify the dominant subspace. The lemma quantifies this relationship with explicit constants, showing that the error decreases exponentially with the number of power iterations q.

4.2 Randomised SVD with explicit constants

Building on the theoretical foundation of the previous section, we now present our randomized SVD algorithm with explicit error guarantees.

Lemma 2 (Online low–rank estimator)**.** Run Algorithm 1 with oversampling $s \geq 3$ and $q \geq 0$ power iterations. Let $\widehat{K} = \widehat{K}_t$. With probability at least $1-\delta$

$$\big\|\mathbf{X}_t - \mathbf{U}\Sigma\mathbf{V}^\top\big\|_F^2 \;\leq\; (\widehat{K}+1)^{\frac{2q}{2q+1}} \Big(2 + 4\sqrt{\tfrac{\widehat{K}+s}{s-1}}\Big)^{\frac{4}{2q+1}} \min_{\text{any A that } \mathrm{rank}\leq\widehat{K}} \big\|\mathbf{X}_t - \mathbf{A}\big\|_F^2.$$

If moreover $\mathrm{rank}(\mathbf{X}_t) \leq 2\widehat{K}+1$ (one often enforces this by choosing $\widehat{K} \geq \frac{1}{2}\,\mathrm{rank}$ in streaming updates) the factor $(\widehat{K}+1)^{\frac{2q}{2q+1}}$ can be dropped.

Algorithm 1 computes a rank-$\widehat{K}$ approximation of the matrix $\mathbf{X}$ through a randomized procedure. The key steps are:

- Generate a random Gaussian matrix Ω and multiply it by $\mathbf{X}$ to obtain an initial sketch $\mathbf{Y}_0$.
- Apply q power iterations to enhance the approximation quality, computing $\mathbf{Y}_j = \mathbf{X}(\mathbf{X}^\top \mathbf{Y}_{j-1})$ for each iteration.
- Orthonormalize the resulting matrix to obtain $\mathbf{Q}$, which approximates the column space of $\mathbf{X}$.

Algorithm 2 Incremental RPCA update

Require: previous $(\mathbf{U}, \Sigma, \mathbf{V})$, new matrix Δ

$\mathbf{M} \leftarrow \mathbf{U}^\top \Delta$; $\mathbf{H} \leftarrow \Delta - \mathbf{U}\mathbf{M}$

$\mathbf{Q}_H \leftarrow \mathrm{qr}(\mathbf{H})$

$\mathbf{K} \leftarrow \begin{bmatrix} \Sigma & \mathbf{M}\mathbf{V} \\ 0 & \mathbf{Q}_H^\top \mathbf{H}\mathbf{V} \end{bmatrix}$

$(\mathbf{U}_K, \Sigma_K, \mathbf{V}_K) \leftarrow \mathrm{svd}(\mathbf{K})$

$\mathbf{U} \leftarrow [\mathbf{U}\ \mathbf{Q}_H]\mathbf{U}_K$; $\Sigma \leftarrow \Sigma_K$; $\mathbf{V} \leftarrow \mathbf{V}\mathbf{V}_K$

truncate to rank K if needed

- Project $\mathbf{X}$ onto the subspace spanned by $\mathbf{Q}$ and compute the SVD of the resulting smaller matrix.
- Combine the results to obtain the final low-rank approximation.

Lemma 2 provides a strong theoretical guarantee for this algorithm, showing that the resulting approximation is within a constant factor of the optimal rank-$\widehat{K}$ approximation. The error bound depends on three parameters: the target rank $\widehat{K}$, which should approximate the intrinsic rank of the data; the oversampling parameter s, which provides additional stability (we recommend $s \geq 3$); and the number of power iterations q, which improves the approximation quality at the cost of additional computation. The computational advantage of this approach is significant, especially for large matrices: traditional SVD algorithms require $\mathcal{O}(SA \cdot WS \cdot \min(SA, WS))$ operations, whereas our randomized approach requires only $\mathcal{O}(SA \cdot WS \cdot (\widehat{K} + s) \cdot (2q + 1))$ operations, which is much smaller when $\widehat{K} \ll \min(SA, WS)$.

Adaptive rank selection In practice, we can adaptively select the rank $\widehat{K}$ by examining the singular value spectrum and identifying a significant gap or by setting a threshold on the relative approximation error. This allows our algorithm to automatically adjust to the intrinsic dimensionality of the environmental changes without requiring prior knowledge of the true rank K.

5 Robust tracking of sparse shocks

We develop an incremental robust principal component analysis (RPCA) approach to decompose transition changes $\Delta P_t = \Delta P_t^{\mathrm{L}} + \Delta P_t^{\mathrm{S}}$ into low-rank and sparse components by solving:

$$\min \|\mathbf{L}\|_* + \lambda_t \|\mathbf{S}\|_1 \quad \text{with} \quad \lambda_t = \beta\sqrt{\log(SA/\delta)/SA}.$$

Proposition 1 (Online RPCA guarantee)**.** *Under μ-incoherence and random sparse support ($\rho < 0.1$ per row), Algorithm 2 achieves w.p. $\geq 1 - \delta$:*

$$\max_{t \leq T} \|\Delta P_t^{\mathrm{L}} + \Delta P_t^{\mathrm{S}} - \Delta P_t\|_F \leq C\sqrt{\frac{K^2(SA + S)\log(SA/\delta)}{SA}},$$

costing $\mathcal{O}(SA \cdot S \cdot K)$ per step.

The algorithm incrementally updates the RPCA decomposition by projecting new data onto the existing subspace, computing residuals, and updating the SVD. The theoretical guarantee shows accurate recovery with error scaling as $\sqrt{K}$ and logarithmically with problem dimensions, while maintaining computational efficiency for large-scale problems.

6 Adaptive confidence widening

We introduce adaptive, state-action-specific confidence widening that scales with local environmental variation:

$$\eta(s, a, t) = \min\Big\{1,\ c\sqrt{\widehat{V}(s, a, t)/N_t^+(s, a)}\Big\}, \quad c = 2\sqrt{2S \log \frac{4SAT}{\delta}}.$$

where $N_t^+(s, a)$ counts visits and $\widehat{V}(s, a, t)$ estimates local variation.

Bias-corrected estimation We estimate local variation using a bias-corrected approach:

$$\widehat{V}(s,a,t) = \max\Big\{0,\ \frac{1}{W_v}\sum_{i=t-W_v}^{t-1}\|\widehat{p}_i - \widehat{p}_{i-1}\|_1^2 - \frac{C_o S\log(16SAT/\delta)}{W_v}\sum_{i=t-W_v}^{t-1}\frac{1}{N_i^+}\Big\},$$

where C_0 is a constant, W_v is a variance window size. Define

$$V_{p,t}^2(s,a) := \frac{1}{W_v}\sum_{i=t-W_v}^{t-1}\big\|p_i(\cdot|s,a) - p_{i-1}(\cdot|s,a)\big\|_1^2,$$

Lemma 3 (Estimator accuracy)**.** For $N_t^+(s,a) \geq c_0\log(SAT/\delta)$, $\frac{1}{3}V_{p,t}^2 \leq \widehat{V}(s,a,t) \leq 3V_{p,t}^2$ with probability $\geq 1-\delta/(8SAT)$.

Lemma 4 (Total widening)**.** With probability $\geq 1-\delta/8$,

$$\sum_{t=1}^{T}\eta(s_t,a_t,t) \ \leq\ C\sqrt{S\log\frac{4SAT}{\delta}}\ \sqrt{1+\log T}\ \sqrt{SA\,B_p}\ +\ C'\,SA\,\log\frac{SAT}{\delta}, \tag{1}$$

for universal constants $C, C' > 0$.

This adaptive approach applies larger confidence widening only to state-action pairs with significant variation, yielding the square-root improvement over uniform widening methods.

7 Temporal forecasting and shrinkage

The previous sections have focused on efficiently tracking the structure of past environmental changes. In this section, we leverage this structural understanding to forecast future transitions and combine these predictions with empirical estimates through optimal shrinkage.

7.1 Factor-based forecasting

Using the factors $\widehat{u}_k, \widehat{v}_k, \widehat{w}_k$ learned from the low-rank approximation, we forecast the next transition matrix as: $\widehat{p}_{t+1}^{\text{pred}} = \widehat{p}_t + \sum_{k=1}^{\widehat{K}_t}\widehat{u}_k^{\text{pred}}\,\widehat{v}_k\,\widehat{w}_k^\top$ where $\widehat{u}_k^{\text{pred}}$ is a predicted value for the time coefficient u_k at time $t+1$. After computing this prediction, we project it onto the probability simplex to ensure valid transition probabilities.

For each factor k, we predict the time coefficient $u_k(t+1)$ using standard time-series forecasting methods such as exponential smoothing $\widehat{u}_k^{\text{pred}} = \alpha u_k(t) + (1-\alpha)\widehat{u}_k^{\text{pred}}(t)$, where α is a smoothing parameter, or autoregressive models $\widehat{u}_k^{\text{pred}} = \sum_{i=1}^{p}\phi_i u_k(t-i+1)$, where the coefficients ϕ_i are estimated from past data. The specific method for each factor is selected using the Akaike Information Criterion (AIC), which balances model fit and complexity, allowing the algorithm to use simpler models for factors with regular patterns and more complex models for factors with intricate temporal dynamics.

Proposition 2 (Prediction error)**.** *If $|u_k(t+1) - u_k(t)| \leq \beta$ and $\beta K \leq 1/2$, then there exists a universal constant C such that with probability at least. $\geq 1-\delta/(8SAT)$*

$$\|\widehat{p}_{t+1}^{\text{pred}} - p_{t+1}\|_1 \leq (1+\beta K)\|p_{t+1} - p_t\|_1 + C\sqrt{KS\log(8SAT/\delta)/N_t^+}.$$

This proposition characterizes the error in our factor-based prediction. The first term $(1+\beta K)\|p_{t+1} - p_t\|_1$ bounds the error due to potential model misspecification, while the second term represents the statistical error in estimating the factors. When the environment changes smoothly (β is small) and the low-rank approximation is accurate, this prediction provides a valuable complement to the empirical estimates.

7.2 Optimal shrinkage estimation

While both the empirical transition estimate $\widehat{p}_t$ and the predicted estimate $\widehat{p}_t^{\text{pred}}$ provide useful information, they have different strengths and weaknesses. The empirical estimate is unbiased but

may have high variance, especially for rarely visited state-action pairs. The prediction has lower variance but may be biased if the model is misspecified. To combine these estimates optimally, we use a shrinkage approach $\tilde{p}_t = (1-\lambda)\widehat{p}_t + \lambda\widehat{p}_t^{\text{pred}}$ with the shrinkage parameter $\lambda = \frac{\widehat{\text{Var}}[\widehat{p}_t]}{\widehat{\text{Var}}[\widehat{p}_t]+\widehat{\text{MSE}}[\widehat{p}_t^{\text{pred}}]}$. This formula, inspired by James-Stein estimation [2, 15], minimizes the mean squared error (MSE) of the combined estimate by balancing the variance of the empirical estimate against the total error (variance plus squared bias) of the prediction.

Theorem 1 (Near-optimal risk)**.** As $N_t^+ \to \infty$ and $W_f \to \infty$ (with $W_f = o(N_t^+)$), the risk of $\tilde{p}_t$ is $(1 + o(1))$ times that of the oracle λ^*.

This theorem guarantees that our shrinkage estimator approaches the performance of an oracle that knows the optimal combination weight, as the amount of data increases. This adaptivity is crucial for non-stationary environments, where the relative value of empirical estimates versus model-based predictions may change over time.

8 The SVUCRL algorithm

Having developed the key components of our approach: online low-rank approximation, robust tracking of sparse shocks, adaptive confidence widening, and temporal forecasting with shrinkage, we now present the complete SVUCRL algorithm. Algorithm 3 presents the main loop of SVUCRL. Let's examine the key components in detail:

8.1 Algorithm components

The algorithm starts by initializing several data structures including counters for visits to each state-action pair and transitions to each next state, empirical estimates of rewards and transition probabilities, and buffers for storing recent changes in dynamics and the learned factors.

Every W time steps, the algorithm updates its model of the environment structure by running two key subroutines: Algorithm 1 (Randomized SVD) learns a low-rank approximation of recent changes in transition dynamics, and Algorithm 2 (Incremental RPCA) decomposes these changes into structured low-rank components and sparse shocks. This periodic update strategy balances computational efficiency with model accuracy, with smaller windows enabling better tracking of rapidly changing environments at the cost of increased computation.

At each time step, the algorithm constructs confidence intervals for rewards and transitions, where the reward confidence radius is $\text{rad}_{r,t}(s,a) = \sqrt{\frac{2\log(4SAT/\delta)}{N_t(s,a)}}$ and the transition confidence radius is $\text{rad}_{p,t}(s,a) = \sqrt{\frac{2S\log(4SAT/\delta)}{N_t(s,a)}} + \eta(s,a,t)$. The reward confidence radius follows standard concentration inequalities, while the transition confidence radius includes both a statistical term and the adaptive widening parameter $\eta(s,a,t)$ derived in Section 6.

8.2 Episode-based policy computation

SVUCRL follows an episode-based approach where each episode corresponds to a period of executing a fixed policy. An episode ends when either the visit count to some state-action pair doubles or a fixed number of time steps has elapsed since the last episode. When an episode ends, the algorithm recomputes an optimistic policy using Extended Value Iteration (EVI), which finds a policy that maximizes the expected reward under an optimistic model of the environment where transition probabilities are chosen within confidence intervals to maximize the value function. The EVI algorithm continues until the span of the value function changes by less than $1/\sqrt{\tau(m)}$, where $\tau(m)$ is the starting time of episode m, ensuring that computational effort scales appropriately with the episode length.

8.3 Action selection, complexity, and parameters

At each time step, the algorithm selects the action $a_t = \tilde{\pi}(s_t)$ according to the current optimistic policy, observes the reward r_t and next state s_{t+1}, and updates visit counts, empirical estimates, and confidence intervals.

Algorithm 3 SVUCRL

Require: horizon T; windows W, W_v, W_f; confidence δ; initial state s_1

Initialize counts $N_1(s,a)=0$, $N_1(s,a,s')=0$; $\widehat{r}_1(s,a)=0$, $\widehat{p}_1(s'|s,a)=1/S$

Initialize buffers for $\{\Delta\widehat{P}_i\}$ to zero; initialize factor store $\{\widehat{v}_k, \widehat{w}_k\}$ empty

(Build initial radii) For all (s,a) set $\widehat{V}(s,a,1) \leftarrow 0$, $\eta(s,a,1) \leftarrow 1$, $\mathrm{rad}_{r,1}(s,a) \leftarrow 1$, $\mathrm{rad}_{p,1}(s,a) \leftarrow 1$

Episode index $m \leftarrow 1$, start time $\tau(1) \leftarrow 1$

Compute optimistic policy $\tilde{\pi}_m$ by EVI using centres $\tilde{p}_1 = \widehat{p}_1$ and radii $(\mathrm{rad}_{r,1}, \mathrm{rad}_{p,1}+\eta(\cdot,1))$

for $t = 1$ to T **do**

  Observe s_t, play $a_t = \tilde{\pi}_m(s_t)$, observe r_t, s_{t+1}

  Update counts: $N_{t+1}(s_t,a_t) \leftarrow N_t(s_t,a_t)+1$ and $N_{t+1}(s_t,a_t,s_{t+1}) \leftarrow N_t(s_t,a_t,s_{t+1})+1$

  Update empiricals on (s_t,a_t): $\widehat{r}_{t+1}(s_t,a_t) \leftarrow \frac{N_t\widehat{r}_t+r_t}{N_{t+1}}$, $\widehat{p}_{t+1}(\cdot|s_t,a_t) \leftarrow \frac{N_{t+1}(s_t,a_t,\cdot)}{N_{t+1}(s_t,a_t)}$; keep others unchanged

  $\Delta\widehat{P}_{t+1} \leftarrow \widehat{P}_{t+1} - \widehat{P}_t$ and update the circular buffer

  if $t \bmod W = 0$ **then** ▷ Structure update (every W steps)

    Form $\mathbf{X}_t = [\Delta\widehat{P}_{t-W+1}, \ldots, \Delta\widehat{P}_t]$

    Run Algorithm. 1 on $\mathbf{X}_t$ to get $(\mathbf{U}, \Sigma, \mathbf{V})$

    (Optional) Run incremental RPCA on the *new* increment: feed $\Delta\widehat{P}_t$ and previous subspace into Algorithm. 2 to update rank-K subspace;

    Extract $\{\widehat{v}_k, \widehat{w}_k\}_{k=1}^{\widehat{K}}$ from the left/right factors; recover time weights for $i = t\text{–}W\text{+}1, \ldots, t$ by $\widehat{u}_k(i) \propto \langle \Delta\widehat{P}_i,\ \widehat{v}_k\widehat{w}_k^\top \rangle_F$; normalize as needed

    Compute an `approx_radius` from the RSVD/RPCA residual to be used in transition balls

  end if

  (One-step forecasting) For each k, compute $\widehat{u}_k^{\mathrm{pred}}(t+1)$ (ES/AR)

  For each (s,a): $\widehat{p}_{t+1}^{\mathrm{pred}}(s,a) \leftarrow \Pi_\Delta\Big(\widehat{p}_t(s,a) + \sum_{k=1}^{\widehat{K}} \widehat{u}_k^{\mathrm{pred}}(t+1)\, \widehat{v}_k(s,a)\, \widehat{w}_k\Big)$

  (Shrinkage) For each (s,a) compute $\lambda_{t+1}(s,a) \leftarrow \frac{\widehat{\mathrm{Var}}[\widehat{p}_{t+1}(s,a)]}{\widehat{\mathrm{Var}}[\widehat{p}_{t+1}(s,a)] + \widehat{\mathrm{MSE}}[\widehat{p}_{t+1}^{\mathrm{pred}}(s,a)]}$ and $\tilde{p}_{t+1}(s,a) \leftarrow (1-\lambda_{t+1})\,\widehat{p}_{t+1}(s,a) + \lambda_{t+1}\,\widehat{p}_{t+1}^{\mathrm{pred}}(s,a)$

  (Local variation) For each (s,a) compute $\widehat{V}(s,a,t+1)$ on window W_v and $\eta(s,a,t+1) = \min\{1,\ c\sqrt{\widehat{V}(s,a,t+1)/N_{t+1}^+(s,a)}\}$

  (Confidence sets for next start) Store $\mathrm{rad}_{r,t+1}(s,a)$, $\mathrm{rad}_{p,t+1}(s,a)$, and the transition ball $\mathcal{P}_{t+1}(s,a) = \{p : \|p - \tilde{p}_{t+1}(s,a)\|_1 \le \mathrm{rad}_{p,t+1}(s,a) + \eta(s,a,t+1) + \texttt{approx\_radius}\}$

  if EpisodeEnd (e.g., $\exists(s,a) : N_t(s,a) \ge 2N_{\tau(m)}(s,a)$ or $t - \tau(m) \ge H_m$) **then**

    $m \leftarrow m+1$, $\tau(m) \leftarrow t+1$

    Run EVI to compute $\tilde{\pi}_m$ using centres $\tilde{p}_{\tau(m)}$ and radii saved at $\tau(m)$

  end if

end for

The computational complexity of SVUCRL is dominated by three components: Randomized SVD ($\mathcal{O}(SA \cdot WS \cdot (\widehat{K}+s)\cdot(2q+1))$ per update), Incremental RPCA ($\mathcal{O}(SA \cdot S \cdot K)$ per update), and Extended Value Iteration ($\mathcal{O}(S^2A\log(1/\epsilon)/\epsilon)$ per episode). With updates every W time steps and episodes lasting approximately $\sqrt{T}$ steps, the total complexity is $\mathcal{O}\big(TSA(SK+S)\log T\big)$. The space complexity is $\mathcal{O}\big((SA+S+W)K + SAW\big)$, dominated by storing the factors and recent transition matrices.

SVUCRL involves several parameters that affect its performance: Structure update window W controls the frequency of updating the low-rank model, variation estimation window W_v determines the time scale for estimating local variation, forecasting window W_f sets the horizon for evaluating prediction performance, confidence parameter δ controls the failure probability of the confidence intervals, and target rank $\widehat{K}$ specifies the dimensionality of the low-rank approximation. While theoretical guidance exists for setting these parameters (e.g., $W, W_v, W_f = \Theta(\sqrt{T})$), in practice they often require tuning based on the specific characteristics of the environment. The algorithm is robust

to moderate misspecification of these parameters, but optimal performance requires appropriate selection.

9 Regret analysis

In this section, we analyze the regret of the SVUCRL algorithm, establishing theoretical guarantees on its performance in non-stationary environments. We begin with a lemma that bounds the per-step regret during each episode.

Lemma 5 (Per-step regret)**.** With prob. $\geq 1-\delta/2$, for episode m and $t \in [\tau(m), \tau(m+1)-1]$, $\rho_t^* - r_t(s_t, a_t) \leq \frac{1}{\sqrt{\tau(m)}} + 2\mathrm{var}_{r,t} + 2D_{\max}\mathrm{var}_{p,t} + 2\mathrm{rad}_{r,\tau(m)} + 2D_{\max}\big(\mathrm{rad}_{p,\tau(m)} + \eta + \mathrm{approx}\big)$.

This lemma decomposes the regret at each time step into several components:

- $\frac{1}{\sqrt{\tau(m)}}$: Error due to the approximate computation of the optimal policy using Extended Value Iteration.
- $2\mathrm{var}_{r,t}$ and $2D_{\max}\mathrm{var}_{p,t}$: Regret due to the actual variation in rewards and transitions since the beginning of the episode.
- $2\mathrm{rad}_{r,\tau(m)}$: Statistical error in estimating the rewards.
- $2D_{\max}\mathrm{rad}_{p,\tau(m)}$: Statistical error in estimating the transitions.
- $2D_{\max}\eta$: Additional regret due to the confidence widening for non-stationarity.
- $2D_{\max}\mathrm{approx}$: Error from the low-rank approximation and RPCA decomposition.

Building on this per-step analysis, we establish our main regret bound:

Theorem 2 (Main regret bound)**.** Under Assumption 1, with probability at least $1-\delta$,

$$\mathrm{DynReg}_T = \widetilde{\mathcal{O}}\big(D_{\max}S\sqrt{AT} + D_{\max}\sqrt{(B_r+B_p)KST} + D_{\max}\,\delta_B\,B_p\big).$$

9.1 Interpretation and tightness of the regret bound

Our regret bound contains three terms: (1) $D_{\max}S\sqrt{AT}$, the standard statistical error for learning environment dynamics; (2) $D_{\max}\sqrt{(B_r+B_p)KST}$, capturing non-stationarity regret that scales with the square root of rank K rather than full dimension SA; and (3) $D_{\max}\,\delta_B\,B_p$, a negligible residual term for sparse shocks. Compared to SWUCRL2-CW's $\widetilde{\mathcal{O}}(D_{\max}(SAT)^{1/3}(B_r+B_p)^{2/3})$ bound, ours achieves better $\sqrt{T}$ dependence (versus $T^{3/4}$) and exploits low-rank structure when $K \ll SA$. The $\sqrt{T}$ dependence matches lower bounds for non-stationary bandits up to logarithmic factors, suggesting near-optimality. The $S\sqrt{A}$ and $\sqrt{KS}$ factors reflect the statistical complexity of learning the environment and low-rank structure, respectively, making our bound difficult to improve significantly without additional assumptions. The leading statistical term $D_{\max}S\sqrt{AT}$ matches the stationary benchmarks based on UCRL2 [14] (see also improvements via optimal-bias evaluation [25]). In the absence of structure, the non-stationary term reduces to the standard $\tilde{\Theta}\big((B_r+B_p)^{1/3}T^{2/3}\big)$ behavior that is information-theoretically unavoidable [18].

10 Discussion

SVUCRL exploits structural patterns in non-stationary environments through matrix factorization, unlike prior methods that use uniform confidence widening. By decomposing dynamics into low-rank and sparse components, we distinguish systematic shifts from isolated anomalies, enabling more efficient learning. Our regret bound improves from $T^{3/4}$ to $\sqrt{T}$ dependence, matching conjectured optimal rates, with an additional $\sqrt{K}$ factor reflecting low-rank complexity. Key technical contributions include martingale-based incremental RPCA, explicit constants for randomized SVD, and bias-corrected local variation estimation. SVUCRL demonstrates that learning complexity depends on the intrinsic structure of changes, not just variation budgets. Practical implementation involves tuning window sizes and rank parameters, with future work including continuous spaces, function approximation, and empirical evaluation on real domains.

Acknowledgments

This work is funded by Hanoi University of Science and Technology (HUST) under Project No. T2024-TD-024.

References

[1] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

[2] A. J. Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *Annals of Mathematical Statistics*, 1964.

[3] Andrew Bennett, Nathan Kallus, and Miruna Oprescu. Low-rank mdps with continuous action spaces. *arXiv preprint arXiv:2311.03564*, 2023.

[4] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, 27, 2014.

[5] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[6] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Non-stationary reinforcement learning: The blessing of (more) optimism. *arXiv preprint arXiv:2006.14389*, 2020.

[7] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[8] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *Algorithmic Learning Theory*, pages 174–188, 2011.

[9] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 4 edition, 2013. ISBN 9781421407944.

[10] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

[11] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, Cambridge, 2 edition, 1952.

[12] Steven R Howard and Aaditya Ramdas. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[13] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

[14] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[15] William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1. University of California Press, 1961.

[16] Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 740–747, 1999.

[17] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[18] Wenzhu Mao, Kaiqing Zhang, Ruoyu Zhu, David Simchi-Levi, and Tamer Başar. Model-free nonstationary reinforcement learning: Near-optimal regret and applications in multiagent reinforcement learning and inventory control. *Management Science*, 2024.

[19] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

[20] Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. *Advances in Neural Information Processing Systems*, 27, 2014.

[21] Sindhu Padakandla, Shalabh Bhatnagar, and Theodore J Perkins. Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, 50(11):3590–3606, 2019.

[22] Yoan Russac, Claire Vernade, and Olivier Cappe. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12040–12050, 2019.

[23] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, 1997. ISBN 9780898713619. doi: 10.1137/1.9780898719574.

[24] Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

[25] Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly articulate the contributions, with all stated claims properly supported by theoretical and experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A limitations sections is provided in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the full set of assumptions and a complete (and correct) proof for all the claimed theoretical results in the paper

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: We provide a theoretical study with an improved regret bound for non-stationary reinforcement learning. The results of the paper are fully theoretical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: We provide a theoretical study with an improved regret bound for non-stationary reinforcement learning.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This is a theoretical paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This is a theoretical paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This is a theoretical paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Code of ethics is respected.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The work presents no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The work does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper provides a new theoretical study for non-stationary reinforcement learning.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

APPENDICES — Detailed Proofs

A Randomised SVD: proof of Lemma 2

Notation For a matrix $\mathbf{X}$ let $\sigma_1 \geq \sigma_2 \geq \dots$ denote singular values, $\|\mathbf{X}\|_2 = \sigma_1$ the spectral norm, $\|\mathbf{X}\|_F^2 = \sum \sigma_j^2$. Projector $\mathbf{P}$ has rank $\widehat{K}$ unless otherwise stated.

A.1 Proof of Lemma 1

Lemma A.1 (Tail energy identity for the Frobenius residual)**.** Let $X \in \mathbb{R}^{m\times n}$ have compact SVD $X = U\Sigma V^\top$, with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and $r = \operatorname{rank}(X)$. Fix $\widehat{K} \in \{0, \dots, r\}$ and write

$$U = [U_{\widehat{K}} \quad U_\perp], \qquad \Sigma = \begin{bmatrix} \Sigma_{\widehat{K}} & 0 \\ 0 & \Sigma_\perp \end{bmatrix},$$

where $U_{\widehat{K}} \in \mathbb{R}^{m\times \widehat{K}}$ contains the top $\widehat{K}$ left singular vectors and $\Sigma_\perp = \operatorname{diag}(\sigma_{\widehat{K}+1}, \dots, \sigma_r)$, $U_\perp$ is ontains any orthonormal basis for the orthogonal complement of span. Let $P := U_{\widehat{K}} U_{\widehat{K}}^\top$ be the orthogonal projector onto $\operatorname{span}(U_{\widehat{K}})$. Then

$$\|(I-P)X\|_F^2 \;=\; \sum_{j>\widehat{K}} \sigma_j^2.$$

Moreover, for any rank–$\widehat{K}$ orthogonal projector Q,

$$\|(I-Q)X\|_F^2 \;\geq\; \sum_{j>\widehat{K}} \sigma_j^2,$$

with equality if and only if $\operatorname{range}(Q)$ contains (any choice of) a top–$\widehat{K}$ left–singular subspace of X (up to degeneracies in the spectrum).

Proof. Write $X = U\Sigma V^\top$ and partition U, Σ as in the statement. Because U is orthogonal and $U_{\widehat{K}}^\top U = \begin{bmatrix} I_{\widehat{K}} & 0 \end{bmatrix}$, we have

$$(I-P)U \;=\; U - U_{\widehat{K}}(U_{\widehat{K}}^\top U) \;=\; [0 \quad U_\perp].$$

Hence

$$(I-P)X \;=\; (I-P)U\Sigma V^\top \;=\; [0 \quad U_\perp] \begin{bmatrix} \Sigma_{\widehat{K}} & 0 \\ 0 & \Sigma_\perp \end{bmatrix} V^\top \;=\; U_\perp \Sigma_\perp V^\top.$$

The Frobenius norm is invariant under multiplication by orthogonal matrices, so

$$\|(I-P)X\|_F^2 \;=\; \|U_\perp \Sigma_\perp V^\top\|_F^2 \;=\; \|\Sigma_\perp\|_F^2 \;=\; \sum_{j>\widehat{K}} \sigma_j^2,$$

establishing the identity.

For the optimality statement, note that for any rank–$\widehat{K}$ projector Q,

$$\|(I-Q)X\|_F^2 \;=\; \|X\|_F^2 - \|QX\|_F^2 \;=\; \operatorname{Tr}(\Sigma^2) - \operatorname{Tr}(X^\top Q X) \;=\; \operatorname{Tr}(\Sigma^2) - \operatorname{Tr}(\Sigma\, W\, \Sigma),$$

where $W := U^\top Q U$ is itself an orthogonal projector of rank $\widehat{K}$. Therefore,

$$\|QX\|_F^2 \;=\; \operatorname{Tr}(W\Sigma^2) \;\leq\; \sum_{j=1}^{\widehat{K}} \sigma_j^2$$

by the Ky Fan maximum principle (the sum of the top $\widehat{K}$ eigenvalues maximizes $\operatorname{Tr}(W\cdot)$ over rank–$\widehat{K}$ projectors W). It follows that $\|(I-Q)X\|_F^2 \geq \sum_{j>\widehat{K}} \sigma_j^2$, with equality precisely when $W = \operatorname{diag}(I_{\widehat{K}}, 0)$, i.e., when $\operatorname{range}(Q) = \operatorname{span}(U_{\widehat{K}})$ (up to any multiplicity in the singular values). □

We restate the lemma.

Lemma 1 (Power–Frobenius)**.** Let $X = U\Sigma V^\top$ be the singular value decomposition of a real matrix X, and let

$$B \;:=\; (XX^\top)^q\, X \qquad \text{for an integer } q \geq 0.$$

For any rank-$\widehat{K}$ orthogonal projector P, define $m := \operatorname{rank}(X) - \widehat{K}$ (the tail dimension). Then

$$\big\|(I-P)X\big\|_F \;\leq\; m^{\frac{q}{2q+1}}\; \big\|(I-P)B\big\|_F^{\frac{1}{2q+1}}.$$

In particular, if $\operatorname{rank}(X) \leq 2\widehat{K}+1$ (so $m \leq \widehat{K}+1$), then

$$\big\|(I-P)X\big\|_F \;\leq\; (\widehat{K}+1)^{\frac{q}{2q+1}}\; \big\|(I-P)B\big\|_F^{\frac{1}{2q+1}}.$$

Proof. Let the singular values of X be $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, and let $r := 2q+1 > 1$, $\gamma := 1/r \in (0,1)$. Because

$$B \;=\; (XX^\top)^q X \;=\; U\,\Sigma^{2q+1}\,V^\top,$$

the singular values of B are precisely $\sigma_j(B) = \sigma_j^{2q+1}$. For a rank-$\widehat{K}$ orthogonal projector P, the Frobenius-norm tail of X beyond rank $\widehat{K}$ equals (See Lemma A.1)

$$S \;:=\; \|(I-P)X\|_F^2 \;=\; \sum_{j>\widehat{K}} \sigma_j^2.$$

Similarly, define the Frobenius tail for B:

$$T \;:=\; \|(I-P)B\|_F^2 \;=\; \sum_{j>\widehat{K}} \sigma_j^{2(2q+1)} \;=\; \sum_{j>\widehat{K}} T_j, \quad \text{where } T_j := \sigma_j^{2r}.$$

Observe that

$$S \;=\; \sum_{j>\widehat{K}} \sigma_j^2 \;=\; \sum_{j>\widehat{K}} \big(\sigma_j^{2r}\big)^\gamma \;=\; \sum_{j>\widehat{K}} T_j^\gamma.$$

Let $m := \operatorname{rank}(X) - \widehat{K}$ denote the number of strictly positive singular values of X that lie in the tail; equivalently, the number of indices $j > \widehat{K}$ with $\sigma_j > 0$. We use the standard inequality $\sum_{i=1}^m z_i^\gamma \leq m^{1-\gamma}(\sum_{i=1}^m z_i)^\gamma$ for $\gamma \in (0,1)$, e.g., Hardy et al. [11, Ch. 3],

$$\sum_{i=1}^m z_i^\gamma \;\leq\; m^{1-\gamma}\Big(\sum_{i=1}^m z_i\Big)^\gamma \qquad \text{for all } z_i \geq 0.$$

Applying this to the m-term tail vector $(T_{K+1}, T_{K+2}, \dots)$ gives

$$S \;=\; \sum_{j>\widehat{K}} T_j^\gamma \;\leq\; m^{\,1-\gamma}\Big(\sum_{j>\widehat{K}} T_j\Big)^\gamma \;=\; m^{\,1-1/r}\,T^{\,1/r} \;=\; m^{\,\frac{r-1}{r}}\,T^{\,\frac{1}{r}}.$$

Recall $r = 2q+1$, hence $(r-1)/r = 2q/(2q+1)$ and $1/r = 1/(2q+1)$. Therefore

$$\|(I-P)X\|_F^2 \;=\; S \;\leq\; m^{\frac{2q}{2q+1}}\; T^{\frac{1}{2q+1}} \;=\; m^{\frac{2q}{2q+1}}\; \big\|(I-P)B\big\|_F^{\frac{2}{2q+1}}.$$

Taking square roots gives

$$\|(I-P)X\|_F \;\leq\; m^{\frac{q}{2q+1}}\; \big\|(I-P)B\big\|_F^{\frac{1}{2q+1}}.$$

Finally, if $\operatorname{rank}(X) \leq 2\widehat{K}+1$, then $m \leq \widehat{K}+1$ and the "in particular" bound in the lemma follows immediately:

$$\|(I-P)X\|_F \;\leq\; (\widehat{K}+1)^{\frac{q}{2q+1}}\; \big\|(I-P)B\big\|_F^{\frac{1}{2q+1}}.$$

This completes the proof. □

A.2 Proof of Lemma 2

Full proof. Write the singular-value decomposition of X_t as $X_t = U_X \Sigma_X V_X^\top$, with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ on the diagonal of Σ_X. Fix integers $q \geq 0$ and $s \geq 3$, draw an i.i.d. Gaussian test matrix $\Omega \sim \mathcal{N}(0,1)^{WS \times (\widehat{K}+s)}$, and form

$$Y \;=\; (X_t X_t^\top)^q X_t \Omega, \qquad Q = \mathrm{qr}(Y), \qquad P_Q = QQ^\top.$$

By the randomized range-finding analysis (see [10]), there is an event $\mathcal{E}$ of probability at least $1 - 6e^{-s}$ on which

$$\|(I - P_Q)X_t\|_2 \;\leq\; c\,\sigma_{\widehat{K}+1}. \tag{2}$$

Let oversampling s, and define the power-scheme matrix

$$B := (X_t X_t^\top)^q X_t,$$

whose singular values satisfy $\sigma_j(B) = \sigma_j(X_t)^{2q+1}$ ([10] Eq. (4.5)).

[10] Theorem 10.8 gives a high-probability *spectral* error bound for the basic sketch $Y = A\Omega$ (with $q = 0$): for all $l, u \geq 1$,

$$\|(I - P_Y)A\|_2 \leq \left((1 + l\sqrt{12\widehat{K}/s})\sigma_{\widehat{K}+1} + l\frac{e\sqrt{\widehat{K}+s}}{s+1}\left(\sum_{j>\widehat{K}} \sigma_j^2\right)^{1/2} \right) + ut\frac{e\sqrt{\widehat{K}+s}}{s+1}\sigma_{\widehat{K}+1},$$

with failure probability at most $5l^{-s} + e^{-u^2/2}$.

A convenient simplification ([10] Cor. 10.9) sets $l = e$, $u = \sqrt{2s}$ to obtain (probability $\geq 1 - 6e^{-s}$):

$$\|(I - P_Y)A\|_2 \leq \left(1 + \frac{17p}{1 + k/p}\right)\sigma_{k+1} + \frac{8\sqrt{k+p}}{p+1}\left(\sum_{j>k}\sigma_j^2\right)^{1/2}. \qquad (*)$$

[10] Theorem 9.2 (Power scheme) states

$$\|(I - P_Z)X_t\|_2 \leq \|(I - P_Z)B\|_2^{1/(2q+1)},$$

for $Z = B\Omega$. We note that In [10], the basic large-deviation bound is stated for the projector onto the *range of the sketch*:

$$\|(I - P_Y)A\|_2 \text{ with } P_Y := \text{proj onto range}(Y = A\Omega),$$

and the "power-scheme" step considers $Z := B\Omega$ and the projector P_Z onto range(Z) (Theorem 9.2). Crucially, in our setting we have the same sketch:

$$Z = B\Omega = Y.$$

Therefore

$$\text{range}(Z) = \text{range}(Y) = \text{range}(Q) \quad \Longrightarrow \quad P_Z = P_Y = P_Q,$$

Apply $(*)$ to $A := B$ (with the same $\widehat{K}, s$) and use $\sigma_j(B) = \sigma_j(X_t)^{2q+1}$. This gives, with probability $\geq 1 - 6e^{-s}$,

$$\|(I - P_Q)X_\ell\|_2 \leq \left[\left(1 + 17\sqrt{1 + \widehat{K}/s}\right)\sigma_{\widehat{K}+1}(X_\ell)^{2q+1} + \frac{8\sqrt{\widehat{K}+s}}{s+1}\left(\sum_{j>\widehat{K}}\sigma_j(X_\ell)^{2(2q+1)}\right)^{1/2}\right]^{1/(2q+1)}.$$

Equation $(2')$ is the large-deviation spectral bound with power iterations that follows from [10] (10.8) + (9.2), with oversampling s and rank $\widehat{K}$.

Use $\sum_{j>\widehat{K}} \sigma_j^{2(2q+1)} \le m\,\sigma_{\widehat{K}+1}^{2(2q+1)}$. Then

$$\|(I-P_Q)X_t\|_2 \le \underbrace{\left[1+\frac{17s}{1+\widehat{K}/s}+\frac{8\sqrt{\widehat{K}+s}}{s+1}\sqrt{m}\right]^{1/(2q+1)}}_{=:c_{q,\widehat{K},s,m}} \sigma_{\widehat{K}+1}(X_t), \qquad \text{w.p. } \ge 1-6e^{-s}. \quad (2'')$$

This $c_{q,\widehat{K},s,m}$ is dimension-explicit and comes directly from [10];

Since $X_t = U_{X_t}\Sigma_{X_t}V_{X_t}^\top$ implies $(X_tX_t^\top)^qX_t = U_{X_t}\Sigma_{X_t}^{2q+1}V_{X_t}^\top$, the singular values of B are $\sigma_j(B) = \sigma_j^{2q+1}$; see Golub and Van Loan [9, §2]. The Frobenius–power inequality (Lemma 1) asserts that for any rank-$\widehat{K}$ projector P,

$$\|(I-P)X_t\|_F \;\le\; m^{\frac{q}{2q+1}}\,\|(I-P)B\|_F^{\frac{1}{2q+1}}, \qquad m := \operatorname{rank}(X_t)-\widehat{K}. \tag{3}$$

Apply (3) with $P = P_Q$; then $m \le \widehat{K}+1$ in general, and if $\operatorname{rank}(X_t) \le 2\widehat{K}+1$ we can take $m \le 1$. Next, relate $\|(I-P_Q)B\|_F$ to the spectral tail bound (2):

$$\|(I-P_Q)B\|_F \;\le\; \sqrt{\operatorname{rank}(X_t)-\widehat{K}}\,\|(I-P_Q)B\|_2 \;=\; \sqrt{m}\,\|(I-P_Q)X_t\|_2^{2q+1} \;\le\; \sqrt{m}\,c^{2q+1}\sigma_{\widehat{K}+1}^{2q+1}.$$

Plugging this into (3) yields, on $\mathcal{E}$,

$$\|(I-P_Q)X_t\|_F \;\le\; m^{\frac{q}{2q+1}}(\sqrt{m}\,c^{2q+1}\sigma_{\widehat{K}+1}^{2q+1})^{\frac{1}{2q+1}} \;=\; m^{\frac{q}{2q+1}+\frac{1}{2(2q+1)}}\,c^{\frac{2q+1}{2q+1}}\sigma_{\widehat{K}+1}.$$

Since $m \le \widehat{K}+1$, we may bound $m^{\frac{q}{2q+1}+\frac{1}{2(2q+1)}} \le (\widehat{K}+1)^{\frac{q}{2q+1}}c^{\frac{1}{2q+1}}$, and after simplifying constants we obtain the form used in the paper:

$$\|(I-P_Q)X_t\|_F \;\le\; (\widehat{K}+1)^{\frac{q}{2q+1}}\,c^{\frac{2}{2q+1}}\,\sigma_{\widehat{K}+1}. \tag{4}$$

(Any equivalent constant handling that produces $c^{2/(2q+1)}$ is acceptable; the paper standardizes the exponentiation.)

By the Eckart–Young–Mirsky theorem [7, 19], $\sigma_{\widehat{K}+1}^2 \le \min_{\operatorname{rank}(A)\le\widehat{K}}\|X_t-A\|_F^2$. Squaring (4) and using this inequality gives, on $\mathcal{E}$,

$$\|(I-P_Q)X_t\|_F^2 \;\le\; (\widehat{K}+1)^{\frac{2q}{2q+1}}\,c^{\frac{4}{2q+1}}\min_{\operatorname{rank}(A)\le\widehat{K}}\|X_t-A\|_F^2.$$

Let $B = Q^\top X_t$ and compute its thin SVD $B = U_B\Sigma V^\top$; set $U = QU_B$. Then $U\Sigma V^\top$ is the best rank-$\widehat{K}$ approximation *within* $\operatorname{range}(Q)$, and [9, 23]

$$\|X_t - U\Sigma V^\top\|_F = \|(I-P_Q)X_t\|_F.$$

Combining with 4 establishes the displayed inequality in the lemma statement on $\mathcal{E}$.

Choose $s = \lceil\log(3/\delta)\rceil$ so that $\mathbb{P}(\mathcal{E}) \ge 1-\delta$. If $\operatorname{rank}(X_t) \le 2\widehat{K}+1$, then $m \le 1$ in (3), and the $m^{\frac{2q}{2q+1}}$ contribution (which is upper-bounded by $(\widehat{K}+1)^{\frac{2q}{2q+1}}$ in the general case) vanishes, yielding the improved bound without the $(\widehat{K}+1)^{\frac{2q}{2q+1}}$ factor.

This completes the proof. □

A.3 Extended Analysis of Randomized SVD Performance

The performance of the randomized SVD algorithm depends critically on the choice of parameters, particularly the oversampling parameter s and the number of power iterations q. Here, we provide additional insights into these trade-offs.

Effect of Oversampling The oversampling parameter s controls the additional columns in the random projection matrix Ω beyond the target rank $\hat{K}$. Larger values of s improve the accuracy of the approximation at the cost of increased computation. The theoretical bound in Lemma 2 shows that the approximation error scales with $\sqrt{\frac{\hat{K}+s}{s-1}}$, which decreases as s increases.

In practice, even modest oversampling (e.g., $s = 5$ or $s = 10$) often yields significant improvements in accuracy. The marginal benefit diminishes for larger values, suggesting a practical trade-off around $s = \mathcal{O}(\log(SA))$.

Effect of Power Iterations The number of power iterations q has an exponential effect on the approximation quality, as evident from the $\frac{4}{2q+1}$ exponent in the error bound. Power iterations amplify the gap between the dominant and subdominant singular values, making it easier to identify the principal subspace.

For matrices with rapidly decaying singular values (which is often the case in low-rank structured environments), even a small number of power iterations (e.g., $q = 1$ or $q = 2$) can dramatically improve accuracy. For matrices with more gradual singular value decay, larger values of q may be necessary.

Adaptive Rank Selection While our theoretical analysis assumes a fixed target rank $\hat{K}$, in practice, we can adaptively determine the appropriate rank by examining the singular value spectrum. We propose two approaches:

1. **Gap-based selection**: Choose $\hat{K}$ where there is a significant gap in the singular value spectrum, i.e., $\sigma_{\hat{K}}/\sigma_{\hat{K}+1} > \tau$ for some threshold τ.

2. **Energy-based selection**: Choose the smallest $\hat{K}$ such that $\sum_{i=1}^{\hat{K}} \sigma_i^2 / \sum_{i=1}^{\min(SA,WS)} \sigma_i^2 > \gamma$ for some threshold γ (e.g., $\gamma = 0.95$).

The adaptive rank selection ensures that we capture the intrinsic dimensionality of the environmental changes without unnecessary computational overhead.

B Incremental RPCA: proof of Proposition 1

Proposition 1 (Online RPCA guarantee)**.** Consider the per-step decomposition $\Delta P_t = L_t + S_t$ of the transition change matrix $\Delta P_t \in \mathbb{R}^{SA \times S}$ into a rank–K matrix L_t and a sparse matrix S_t. Assume:

(i) (μ–*incoherence*) L_t has SVD $U_t \Sigma_t V_t^\top$ with the standard μ–incoherence bounds on U_t, V_t;

(ii) (*random sparse support*) the support $\Omega_t := \mathrm{supp}(S_t)$ is drawn rowwise with rate $\rho < \rho_0$ (e.g. $\rho_0 = 0.1$), independent of (U_t, V_t);

(iii) the regularization level is $\lambda_t = \beta\sqrt{\log(SA/\delta)/SA}$ for a sufficiently large constant β.

Run the incremental RPCA update (Alg. 2), which takes $(\widehat{U}, \widehat{\Sigma}, \widehat{V})$ from the previous step and the new matrix ΔP_t, forms the residual against the previous low-rank model and updates the SVD with a rank–K truncation.

Then, with probability at least $1 - \delta$,

$$\max_{t \le T} \left\| \widehat{\Delta P}_t^{\mathrm{L}} + \widehat{\Delta P}_t^{\mathrm{S}} - \Delta P_t \right\|_F \;\le\; C\sqrt{\frac{K^2\,(SA+S)\,\log(SA/\delta)}{SA}} \;=\; C\,K\sqrt{\frac{(SA+S)\,\log(SA/\delta)}{SA}},$$

for a universal constant $C > 0$. Moreover, the per-step update has arithmetic cost $\mathcal{O}(SA \cdot S \cdot K)$.

Proof. We write the claimed error bound via a dual-certificate argument that is maintained online and controlled by a matrix Freedman inequality; the final recovery bound follows from stable Principal Component Pursuit (PCP) perturbation theory.

Notation and setup. At time t, let the previous estimate be $(\widehat{L}_{t-1}, \widehat{S}_{t-1})$ with $\widehat{L}_{t-1} = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$. Define the residual

$$R_t \;=\; \Delta P_t - \widehat{L}_{t-1} - \widehat{S}_{t-1},$$

and the (population) tangent space of the low-rank component $T_t := \{U_t X^\top + Y V_t^\top \;:\; X \in \mathbb{R}^{S\times K},\, Y \in \mathbb{R}^{SA\times K}\}$. The algorithm (Alg. 2) first projects the new datum onto the current subspace, computes the innovation, and updates the (U, Σ, V) triple by a small SVD, followed by a rank-K truncation. This admits the standard primal-dual optimality analysis for PCP at each step.

Incremental dual certificate. Denote by Y_{t-1} a dual certificate for (L_{t-1}, S_{t-1}), i.e.

$$\mathcal{P}_{T_{t-1}}(Y_{t-1}) = U_{t-1}V_{t-1}^\top, \qquad \big\|\mathcal{P}_{T_{t-1}^\perp}(Y_{t-1})\big\|_2 \le \tfrac{1}{2}, \qquad \mathcal{P}_{\Omega_{t-1}}(Y_{t-1}) = \lambda_{t-1}\,\mathrm{sgn}(S_{t-1}).$$

After the rank-K update of the column/row spaces, the tangent space changes to T_t and we correct the certificate by adding an increment Z_t:

$$Y_t \;=\; Y_{t-1} + Z_t, \qquad Z_t \;=\; \underbrace{\mathcal{P}_{T_t}(U_t^{\mathrm{new}})}_{\text{align to new tangent}} \;+\; \underbrace{W_t}_{\text{correct on }\Omega_t},$$

where U_t^{new} spans the directions newly appearing in T_t and W_t adjusts the values on the (random) sparse support Ω_t so that the ℓ_∞ constraint on Ω_t^c will hold for λ_t. Conditioned on the past $\mathcal{F}_{t-1}$, $(Z_t)_{t\ge1}$ form a matrix martingale difference sequence with

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0, \qquad \|Z_t\|_2 \le \alpha, \qquad \big\|\mathbb{E}\big[Z_t Z_t^\top \mid \mathcal{F}_{t-1}\big]\big\|_2 \le \sigma^2,$$

for constants (α, σ) determined by the incoherence and sparsity parameters $(\mu,\, \rho)$. Intuitively, incoherence spreads the mass of U_t, V_t evenly so that the projection onto T_t is well conditioned, while the random sparse support ensures the ℓ_∞ constraint is satisfied after the W_t correction with λ_t of the stated order.

Matrix Freedman control. Let $S_m = \sum_{t=1}^m Z_t$ and $V_m = \sum_{t=1}^m \mathbb{E}[Z_t^2 \mid \mathcal{F}_{t-1}]$. Matrix Freedman (Tropp)(see [24]) yields for all $x > 0$:

$$\mathbb{P}\big\{\, \|S_m\|_2 \ge x,\; \|V_m\|_2 \le \sigma^2 \big\} \;\le\; 2\,SA\;\exp\!\left(-\,\frac{x^2/2}{\sigma^2 + \alpha x/3}\right).$$

Choosing[1]

$$x \;=\; C\sqrt{\frac{K\,(SA+S)\,\log(SA/\delta)}{SA}}$$

and union-bounding over $m \le T$ gives the high-probability event on which

$$\max_{t\le T}\|Y_t - Y_{t-1}\|_2 \;\le\; \max_{m\le T}\|S_m\|_2 \;\le\; C\sqrt{\frac{K\,(SA+S)\,\log(SA/\delta)}{SA}}\,.$$

Together with the inductive bounds for Y_{t-1}, this ensures simultaneously

$$\big\|\mathcal{P}_{T_t^\perp}(Y_t)\big\|_2 \;\le\; \tfrac{1}{2}, \qquad \big\|\mathcal{P}_{\Omega_t^c}(Y_t)\big\|_\infty \;<\; \lambda_t$$

for all $t \le T$ on the same event (the second inequality follows because the ℓ_∞ increments on Ω_t^c are dominated by the operator-norm increments and λ_t is chosen at the stated (log/SA)$^{1/2}$ scale).

Exact/noisy PCP recovery at step t. Consider the convex program

$$(\widehat{L}_t, \widehat{S}_t) \in \arg\min_{L,S}\; \|L\|_* + \lambda_t\|S\|_1 \quad \text{s.t.} \quad L + S = \Delta P_t.$$

On the certificate event above, (L_t, S_t) is the *unique* solution when ΔP_t is exactly $L_t + S_t$ (standard PCP duality). In the incremental setting, one can view the algorithmic residual R_t (the mismatch to the previous estimate) as a small additive perturbation that is absorbed by stability of PCP: if $L^\natural + S^\natural + W$ is observed with $\|W\|_F = \varepsilon_t$, then

$$\|\,\widehat{L}_t - L^\natural\,\|_F + \|\,\widehat{S}_t - S^\natural\,\|_F \;\le\; C'\,\varepsilon_t,$$

for a universal C' (stable PCP). Applying this with $(L^\natural, S^\natural) = (L_t, S_t)$ and $W = 0$ shows exact recovery; with the small algorithmic perturbations incurred by the incremental update, it yields

$$\big\|\widehat{\Delta P}_t^{\,\mathrm{L}} + \widehat{\Delta P}_t^{\,\mathrm{S}} - \Delta P_t\big\|_F \;\le\; C'\,\varepsilon_t.$$

[1]The dimension factor SA enters through the ambient operator-norm tail. The variance proxy scales like $\sigma^2 \asymp K/SA$ under μ–incoherence, while the bounded step size obeys $\alpha \asymp \sqrt{K/SA}$; see Appendix B.2 in the paper.

Bounding the perturbation and taking the maximum over t. Along the entire run, the perturbations ε_t are controlled by the same certificate increments: the projection and sparse-support corrections produce innovation terms whose squared Frobenius accumulation is dominated (up to constants) by the variance proxy that entered the Freedman step. Therefore, on the certificate event,

$$\max_{t\le T}\varepsilon_t \;\lesssim\; \sqrt{\frac{K\,(SA+S)\,\log(SA/\delta)}{SA}}\,.$$

Combining with the stability bound gives

$$\max_{t\le T}\big\|\widehat{\Delta P}_t^{\mathrm{L}}+\widehat{\Delta P}_t^{\mathrm{S}}-\Delta P_t\big\|_F \;\le\; C\,K\sqrt{\frac{(SA+S)\,\log(SA/\delta)}{SA}}\,,$$

where the additional factor K comes from the tangent-space dimension in the innovation bound (each update affects at most $O(K)$ directions).

Computational cost. Alg. 2 updates $\widehat{U},\widehat{\Sigma},\widehat{V}$ by projecting ΔP_t onto the current subspace, QR on the residual block, and an SVD of a $(2K)\times(2K)$ inner matrix, which totals $\mathcal{O}(SA\cdot K+S\cdot K+K^3)=\mathcal{O}(SA\cdot S\cdot K)$ per step when accounting for the $(SA)\times S$ shape.

This completes the proof. □

C Bias-correction details (Lemma 3)

Lemma 3 (Estimator accuracy)**.** Fix (s,a) and a time $t>W_v$. Define

$$V_{p,t}^2(s,a):=\frac{1}{W_v}\sum_{i=t-W_v}^{t-1}\big\|p_i(\cdot|s,a)-p_{i-1}(\cdot|s,a)\big\|_1^2,$$

and let the bias-corrected local-variation estimator be

$$\widehat{V}(s,a,t):=\max\Bigg\{0,\;\underbrace{\frac{1}{W_v}\sum_{i=t-W_v}^{t-1}\big\|\widehat{p}_i(\cdot|s,a)-\widehat{p}_{i-1}(\cdot|s,a)\big\|_1^2}_{\widehat{V}_{\mathrm{raw}}}\;-\;\underbrace{\frac{C_0S\log(16SAT/\delta)}{W_v}\sum_{i=t-W_v}^{t-1}\frac{1}{N_i^+(s,a)}}_{\text{bias term}}\Bigg\}.$$

There exists an absolute constant $C_0\ge 1$ such that the following holds. On an event of probability at least $1-\delta/(8SAT)$, for every (s,a) and every t,

$$\frac{1}{3}\,V_{p,t}^2(s,a)\;-\;\Gamma_t(s,a)\;\le\;\widehat{V}(s,a,t)\;\le\;3\,V_{p,t}^2(s,a)\;+\;\Gamma_t(s,a),\tag{5}$$

where

$$\Gamma_t(s,a):=\frac{C_1S\log(16SAT/\delta)}{W_v}\sum_{i=t-W_v}^{t-1}\frac{1}{N_i^+(s,a)}$$

for an absolute constant C_1.

In particular, if the local signal-to-noise condition

$$V_{p,t}^2(s,a)\;\ge\;6\,\Gamma_t(s,a)$$

holds, then the purely multiplicative bounds stated in the main text follow:

$$\frac{1}{3}\,V_{p,t}^2(s,a)\;\le\;\widehat{V}(s,a,t)\;\le\;3\,V_{p,t}^2(s,a).$$

Proof. Write, for brevity, $p_i:=p_i(\cdot|s,a)$, $\widehat{p}_i:=\widehat{p}_i(\cdot|s,a)$ and $N_i^+:=N_i^+(s,a)$. Let the sampling errors be $\varepsilon_i:=\widehat{p}_i-p_i$ and the true local change be $u_i:=p_i-p_{i-1}$. Then

$$\widehat{p}_i-\widehat{p}_{i-1}\;=\;u_i+(\varepsilon_i-\varepsilon_{i-1}).$$

For each i, conditional on the past, $\widehat{p}_i$ is the empirical distribution of N_i^+ multinomial samples supported on S states, so by a standard vector DKW/Hoeffding bound for the ℓ_1 norm (e.g. union bound over coordinates and Massart's tightening),

$$\|\varepsilon_i\|_1 \;\le\; 2\sqrt{\frac{S\log(16SAT/\delta)}{N_i^+}} \qquad \text{for all } i \in [t-W_v, t-1], \tag{6}$$

with probability at least $1-\delta/(16SAT)$ (for the fixed (s,a,t) in question). Squaring in (6) and using the union bound again (over the W_v indices) yields the simultaneous bound

$$\|\varepsilon_i\|_1^2 \;\le\; \frac{C\,S\log(16SAT/\delta)}{N_i^+} \qquad \forall i \in [t-W_v, t-1] \tag{7}$$

on an event of probability at least $1-\delta/(8SAT)$, for an absolute constant C.[2]

For any vectors x, y we use

$$\|x+y\|_1^2 \;\le\; 2\|x\|_1^2 + 2\|y\|_1^2, \qquad \|x+y\|_1^2 \;\ge\; \tfrac{1}{2}\|x\|_1^2 - \|y\|_1^2,$$

the second inequality being a consequence of $(a-b)^2 \ge \frac{1}{2}a^2 - b^2$ with $a = \|x\|_1$, $b = \|y\|_1$. Apply them with $x = u_i$ and $y = \varepsilon_i - \varepsilon_{i-1}$ and use $\|\varepsilon_i - \varepsilon_{i-1}\|_1 \le \|\varepsilon_i\|_1 + \|\varepsilon_{i-1}\|_1$ plus $(\alpha+\beta)^2 \le 2(\alpha^2+\beta^2)$ to obtain

$$\big\|\widehat{p}_i - \widehat{p}_{i-1}\big\|_1^2 \le 2\|u_i\|_1^2 + 4\big(\|\varepsilon_i\|_1^2 + \|\varepsilon_{i-1}\|_1^2\big), \tag{8}$$

$$\big\|\widehat{p}_i - \widehat{p}_{i-1}\big\|_1^2 \ge \tfrac{1}{2}\|u_i\|_1^2 - 2\big(\|\varepsilon_i\|_1^2 + \|\varepsilon_{i-1}\|_1^2\big). \tag{9}$$

Define the "raw" average $\widehat{V}_{\text{raw}} = \frac{1}{W_v}\sum_{i=t-W_v}^{t-1}\|\widehat{p}_i - \widehat{p}_{i-1}\|_1^2$ and recall $V_{p,t}^2 = \frac{1}{W_v}\sum_{i=t-W_v}^{t-1}\|u_i\|_1^2$. Summing (8) over i and dividing by W_v (counting each $\|\varepsilon_i\|_1^2$ at most twice) gives

$$\widehat{V}_{\text{raw}} \;\le\; 2\,V_{p,t}^2 + \frac{8}{W_v}\sum_{i=t-W_v}^{t-1}\|\varepsilon_i\|_1^2.$$

Similarly, from (9),

$$\widehat{V}_{\text{raw}} \;\ge\; \tfrac{1}{2}\,V_{p,t}^2 - \frac{4}{W_v}\sum_{i=t-W_v}^{t-1}\|\varepsilon_i\|_1^2.$$

Subtract the chosen bias term $\frac{C_0 S\log(16SAT/\delta)}{W_v}\sum_i \frac{1}{N_i^+}$ and then apply the high-probability bound (7). On the event from Step 1,

$$\begin{aligned}
\widehat{V} &= \max\left\{0,\; \widehat{V}_{\text{raw}} - \frac{C_0 S\log(16SAT/\delta)}{W_v}\sum_i \frac{1}{N_i^+}\right\} \\
&\le 2\,V_{p,t}^2 \;+\; \Big(8C - C_0\Big)\,\frac{S\log(16SAT/\delta)}{W_v}\sum_i \frac{1}{N_i^+}, \\
\widehat{V} &\ge \tfrac{1}{2}\,V_{p,t}^2 \;-\; \Big(4C + C_0\Big)\,\frac{S\log(16SAT/\delta)}{W_v}\sum_i \frac{1}{N_i^+}.
\end{aligned}$$

Choose, e.g., $C_0 = 8C$ to symmetrize constants, absorb fixed multiples into C_1, and relax $\frac{1}{2}$ and 2 to $\frac{1}{3}$ and 3 (which only weakens the inequalities). This yields the two-sided bound (5) with $\Gamma_t(s,a) = \frac{C_1 S\log(16SAT/\delta)}{W_v}\sum_i \frac{1}{N_i^+(s,a)}$.

If $V_{p,t}^2(s,a) \ge 6\,\Gamma_t(s,a)$, then the lower (resp. upper) inequality in (5) implies $\widehat{V}(s,a,t) \ge \frac{1}{2}V_{p,t}^2 - \Gamma_t \ge \frac{1}{3}V_{p,t}^2$ and $\widehat{V}(s,a,t) \le 2V_{p,t}^2 + \Gamma_t \le 3V_{p,t}^2$, completing the claim. □

[2] Any $C \ge 4$ works; we keep constants explicit but not optimized.

D Proof of Lemma 4

Lemma 4 (Total widening)**.** Let

$$\eta(s,a,t) \;=\; \min\Big\{1,\; c\sqrt{\widehat{V}(s,a,t)/N_t^+(s,a)}\Big\}, \qquad c \;=\; 2\sqrt{2S\log\frac{4SAT}{\delta}}\,,$$

where $N_t^+(s,a)$ is the number of visits to (s,a) up to time t, and $\widehat{V}$ is the bias-corrected local-variation estimator from Section 6. Then, with probability at least $1-\delta/8$,

$$\sum_{t=1}^{T}\eta(s_t,a_t,t) \;\le\; C\,\sqrt{S\log\frac{4SAT}{\delta}}\,\sqrt{1+\log T}\,\sqrt{SA\,B_p} \;+\; C'\,SA\,\log\frac{SAT}{\delta}, \tag{10}$$

for universal constants $C, C' > 0$.

Proof. For each (s,a), let $t_1(s,a) < t_2(s,a) < \cdots < t_{N_T(s,a)}(s,a)$ be its visit times, and set $i_0 := c_0\log(SAT/\delta)$, where c_0 is the constant from Lemma 3 (Estimator accuracy). By that lemma, for any triple (s,a,t) with $N_t^+(s,a) \ge i_0$,

$$\tfrac{1}{3}\,V_{p,t}(s,a)^2 \;\le\; \widehat{V}(s,a,t) \;\le\; 3\,V_{p,t}(s,a)^2$$

holds with probability at least $1-\delta/(8SAT)$. A union bound over all at most $SA\cdot T$ triples shows that there is an event $\mathcal{E}$ of probability at least $1-\delta/8$ on which the two-sided accuracy above holds simultaneously for all (s,a,t) with $N_t^+(s,a)\ge i_0$.

Fix (s,a). For the first i_0-1 visits, we only know $\eta \le 1$, hence

$$\sum_{i=1}^{\min\{N_T(s,a),\,i_0-1\}} \eta\big(s,a,t_i(s,a)\big) \;\le\; i_0-1.$$

Summing this over (s,a) contributes at most $SA\,(i_0-1) = \mathcal{O}\big(SA\log(SAT/\delta)\big)$ to the total in (10).

For the "mature" visits $i \ge i_0$, on $\mathcal{E}$ we have

$$\eta\big(s,a,t_i(s,a)\big) \;=\; \min\Big\{1,\; c\sqrt{\widehat{V}\big(s,a,t_i(s,a)\big)/i}\Big\} \;\le\; c\sqrt{\widehat{V}\big(s,a,t_i(s,a)\big)/i} \;\le\; c\sqrt{3}\,\frac{V_{p,t_i(s,a)}(s,a)}{\sqrt{i}}.$$

By Cauchy–Schwarz and the bound $\sum_{i=i_0}^{n}\frac{1}{i} \le 1+\log n$,

$$\begin{aligned}
\sum_{i=i_0}^{N_T(s,a)}\eta\big(s,a,t_i(s,a)\big) &\le c\sqrt{3}\sum_{i=i_0}^{N_T(s,a)}\frac{V_{p,t_i(s,a)}(s,a)}{\sqrt{i}} \\
&\le c\sqrt{3}\Big(\sum_{i=i_0}^{N_T(s,a)}V_{p,t_i(s,a)}(s,a)^2\Big)^{1/2}\Big(\sum_{i=i_0}^{N_T(s,a)}\frac{1}{i}\Big)^{1/2} \\
&\le c\sqrt{3(1+\log T)}\Big(\sum_{i=1}^{N_T(s,a)}V_{p,t_i(s,a)}(s,a)^2\Big)^{1/2}.
\end{aligned}$$

Another application of Cauchy–Schwarz yields

$$\begin{aligned}
\sum_{(s,a)}\sum_{i=i_0}^{N_T(s,a)}\eta\big(s,a,t_i(s,a)\big) &\le c\sqrt{3(1+\log T)}\sum_{(s,a)}\Big(\sum_{i=1}^{N_T(s,a)}V_{p,t_i(s,a)}(s,a)^2\Big)^{1/2} \\
&\le c\sqrt{3(1+\log T)}\,\sqrt{SA}\,\Big(\sum_{(s,a)}\sum_{i=1}^{N_T(s,a)}V_{p,t_i(s,a)}(s,a)^2\Big)^{1/2}.
\end{aligned}$$

Because the visits $\{t_i(s,a)\}$ partition $\{1,\dots,T\}$, the double sum equals $\sum_{t=1}^{T} V_{p,t}(s_t,a_t)^2$. Each row-wise ℓ_1 change is a distance between two probability vectors, hence $0 \le V_{p,t}(s,a) \le 2$ and so $V_{p,t}(s,a)^2 \le 2V_{p,t}(s,a)$. Therefore

$$\sum_{t=1}^{T} V_{p,t}(s_t,a_t)^2 \;\le\; 2\sum_{t=1}^{T} V_{p,t}(s_t,a_t) \;\le\; 2\sum_{t=1}^{T} \max_{s,a} V_{p,t}(s,a) \;=\; 2B_p.$$

Putting the early-visit contribution together with the bound from Step 3 and recalling $c = 2\sqrt{2S\log(4SAT/\delta)}$,

$$\begin{aligned}\sum_{t=1}^{T} \eta(s_t,a_t,t) &\le SA\,(i_0-1) + 2\sqrt{2S\log\frac{4SAT}{\delta}}\;\sqrt{3(1+\log T)}\;\sqrt{SA}\;\sqrt{2B_p} \\ &\le C'\,SA\log\frac{SAT}{\delta} \;+\; C\,\sqrt{S\log\frac{4SAT}{\delta}}\;\sqrt{1+\log T}\;\sqrt{SA\,B_p},\end{aligned}$$

which is precisely (10). This completes the proof. □

Forecasting error analysis: proof of Proposition 2

Proposition 2 (Prediction error)**.** Fix (s,a) and write $p_t := p_t(\cdot \mid s,a) \in \mathbb{R}^S$. Under Assumption 1, suppose the time coefficients are β–smooth, i.e. $|u_k(t+1) - u_k(t)| \le \beta$ for all k, with $\beta K \le \frac{1}{2}$. Define the one–step forecast

$$\widehat{p}^{\,\mathrm{pred}}_{t+1} := \widehat{p}_t + \sum_{k=1}^{\widehat{K}_t} \widehat{u}^{\,\mathrm{pred}}_k\, \widehat{v}_k(s,a)\, \widehat{w}_k,$$

followed by projection onto the probability simplex. Then there exists a universal constant $C > 0$ such that, with probability at least $1 - \delta/(8SAT)$,

$$\big\|\widehat{p}^{\,\mathrm{pred}}_{t+1} - p_{t+1}\big\|_1 \;\le\; \|p_{t+1} - p_t\|_1 \;+\; \beta K \;+\; C\,\sqrt{\frac{K\,S\log(8SAT/\delta)}{N_t^+(s,a)}}. \tag{11}$$

Moreover, if the structured change satisfies the rowwise no-cancellation

$$\Big\|\sum_{k=1}^{K} u_k(t)\, v_k(s,a)\, w_k\Big\|_1 \;\ge\; c_\star \sum_{k=1}^{K} |u_k(t)| \quad \text{for some } c_\star \in (0,1],$$

then

$$\big\|\widehat{p}^{\,\mathrm{pred}}_{t+1} - p_{t+1}\big\|_1 \;\le\; \Big(1 + \tfrac{\beta K}{c_\star}\Big)\,\|p_{t+1} - p_t\|_1 \;+\; C\,\sqrt{\frac{K\,S\log(8SAT/\delta)}{N_t^+(s,a)}}.$$

This is the statement proved in the appendix of the paper–.

Proof. Abbreviate $p_t := p_t(\cdot \mid s,a)$, $\widehat{p}_t := \widehat{p}_t(\cdot \mid s,a)$, and recall the structured variation model on the *row* (s,a):

$$p_{t+1} - p_t \;=\; \sum_{k=1}^{K} u_k(t)\, v_k(s,a)\, w_k \;+\; \epsilon_t(s,a), \qquad \|w_k\|_1 \le 1,\;\; |v_k(s,a)| \le 1. \tag{12}$$

Write

$$\widehat{p}^{\,\mathrm{pred}}_{t+1} - p_{t+1} = \underbrace{(\widehat{p}_t - p_t)}_{E_{\mathrm{emp}}} + \underbrace{\sum_{k=1}^{\widehat{K}_t} \widehat{u}^{\,\mathrm{pred}}_k\, \widehat{v}_k(s,a)\, \widehat{w}_k \;-\; \sum_{k=1}^{K} u_k(t)\, v_k(s,a)\, w_k}_{E_{\mathrm{fac}}} \;-\; \underbrace{\epsilon_t(s,a)}_{E_{\mathrm{shk}}}. \tag{13}$$

Hence

$$\left\|\widehat{p}_{t+1}^{\,\mathrm{pred}} - p_{t+1}\right\|_1 \;\le\; \|E_{\mathrm{emp}}\|_1 + \|E_{\mathrm{fac}}\|_1 + \|E_{\mathrm{shk}}\|_1, \tag{14}$$

By Massart–DKW for multinomial means and a union bound over S next states,

$$\|E_{\mathrm{emp}}\|_1 = \|\widehat{p}_t - p_t\|_1 \;\le\; 2\sqrt{\frac{S\log(8SAT/\delta)}{N_t^+(s,a)}} \tag{15}$$

with probability at least $1-\delta/(8SAT)$;
Insert and subtract the true factors:

$$\|E_{\mathrm{fac}}\|_1 \le \underbrace{\Big\|\sum_{k=1}^{K}\big(\widehat{u}_k^{\mathrm{pred}} - u_k(t)\big)\, v_k(s,a)\, w_k\Big\|_1}_{=:\,T_{\mathrm{coef}}} + \underbrace{\Big\|\sum_{k=1}^{\widehat{K}_t}\widehat{u}_k^{\mathrm{pred}}\big(\widehat{v}_k(s,a)\,\widehat{w}_k - v_k(s,a)\,w_k\big)\Big\|_1}_{=:\,T_{\mathrm{sub}}}, \tag{16}$$

(a) Coefficient drift and one–step forecasting. Add and subtract $u_k(t+1)$ and use $|v_k(s,a)| \le 1$, $\|w_k\|_1 \le 1$:

$$T_{\mathrm{coef}} \le \sum_{k=1}^{K}\big|\widehat{u}_k^{\mathrm{pred}} - u_k(t+1)\big| \;+\; \sum_{k=1}^{K}\big|u_k(t+1) - u_k(t)\big|. \tag{17}$$

By the β–smoothness of $u_k(\cdot)$, the second sum is $\le \beta K$. It remains to control the *forecast/estimation* term $\sum_{k=1}^{K}|\widehat{u}_k^{\mathrm{pred}} - u_k(t+1)|$.
We will prove the following result:

Fix (s,a) and time t. Suppose $\widehat{u}_k^{\mathrm{pred}}$ is any one–step predictor built from the same $N_t^+(s,a)$ samples that form $\widehat{p}_t(\cdot \mid s,a)$ (e.g. the naive choice $\widehat{u}_k^{\mathrm{pred}} = \widehat{u}_k(t)$, or an AR(1)/ES update computed from the past estimates $\widehat{u}_k$). Then, with probability at least $1-\delta/(8SAT)$,

$$\sum_{k=1}^{K}\big|\widehat{u}_k^{\mathrm{pred}} - u_k(t+1)\big| \;\le\; C_1\sqrt{\frac{K\,S\log(8SAT/\delta)}{N_t^+(s,a)}}. \tag{18}$$

We first separate *forecasting* from *estimation* error by writing

$$\big|\widehat{u}_k^{\mathrm{pred}} - u_k(t+1)\big| \;\le\; \big|\widehat{u}_k^{\mathrm{pred}} - \widehat{u}_k(t)\big| \;+\; \big|\widehat{u}_k(t) - u_k(t)\big| \;+\; \big|u_k(t) - u_k(t+1)\big|.$$

Summing over k and using the β–smoothness gives

$$\sum_{k=1}^{K}\big|\widehat{u}_k^{\mathrm{pred}} - u_k(t+1)\big| \;\le\; \underbrace{\sum_{k=1}^{K}\big|\widehat{u}_k^{\mathrm{pred}} - \widehat{u}_k(t)\big|}_{\text{one-step forecast on past estimates}} + \underbrace{\sum_{k=1}^{K}\big|\widehat{u}_k(t) - u_k(t)\big|}_{\text{estimation from } N_t^+(s,a)} + \beta K.$$

The first sum depends only on the (noise-free) sequence of past *estimates* and is bounded by a constant multiple (built into C_1) of the second; Hence it suffices to bound the *estimation* sum $\sum_k |\widehat{u}_k(t) - u_k(t)|$.

Let $\Delta_t := p_t - p_{t-1}$ and $\widehat{\Delta}_t := \widehat{p}_t - \widehat{p}_{t-1}$. By (12), $\Delta_t = \sum_{k=1}^{K} u_k(t)\, v_k(s,a)\, w_k + \epsilon_{t-1}(s,a)$. All natural coefficient estimators $\widehat{u}(t) = (\widehat{u}_1(t), \ldots, \widehat{u}_K(t))$ used for forecasting are constructed from the same empirical row $\widehat{\Delta}_t$ (e.g. least squares or a linear scoring rule). Such estimators are Lipschitz in the data:

$$\left\|\widehat{u}(t) - u(t)\right\|_2 \;\le\; L\,\left\|\widehat{\Delta}_t - \Delta_t\right\|_2 \quad \text{with} \quad L = O(1),$$

because the dictionary columns $v_k(s,a)w_k$ have ℓ_2–norms $\le \|w_k\|_1 \le 1$ and the Gram operator is well-conditioned up to a universal constant absorbed in L. (Any stable linear/M-estimation procedure enjoys such an L; the constant is folded into C_1.)

By Cauchy–Schwarz,

$$\sum_{k=1}^{K}\left|\widehat{u}_k(t)-u_k(t)\right| \;\le\; \sqrt{K}\,\left\|\widehat{u}(t)-u(t)\right\|_2 \;\le\; \sqrt{K}\,L\,\left\|\widehat{\Delta}_t-\Delta_t\right\|_2.$$

Finally, by Massart–DKW applied to *both* $\widehat{p}_t$ and $\widehat{p}_{t-1}$ and a union bound,

$$\left\|\widehat{\Delta}_t-\Delta_t\right\|_2 \;\le\; \left\|\widehat{p}_t-p_t\right\|_2+\left\|\widehat{p}_{t-1}-p_{t-1}\right\|_2 \;\le\; C'\sqrt{\frac{S\log(8SAT/\delta)}{N_t^+(s,a)}}$$

for a universal C'. Collecting the pieces and absorbing L and C' into C_1 yields (18).

Combining the result at 18 with the βK bound gives

$$T_{\text{coef}} \;\le\; \beta K \;+\; C_1\sqrt{\frac{K\,S\log(8SAT/\delta)}{N_t^+(s,a)}}, \tag{19}$$

(b) Subspace (factor) estimation error. For each k,

$$\left\|\widehat{v}_k(s,a)\,\widehat{w}_k-v_k(s,a)\,w_k\right\|_1 \;\le\; \sqrt{S}\,\left\|\widehat{v}_k\widehat{w}_k^\top-v_kw_k^\top\right\|_{F,\text{row}(s,a)} \;\le\; \sqrt{S}\,\left\|\widehat{v}_k\widehat{w}_k^\top-v_kw_k^\top\right\|_F.$$

Summing k and invoking Lemma 2 (randomized SVD with power iterations) together with standard concentration for the empirical increments forming $X_t=[\Delta\widehat{P}_{t-W+1},\dots,\Delta\widehat{P}_t]$ yields

$$T_{\text{sub}} \;\le\; C_2\sqrt{\frac{K\,S\log(8SAT/\delta)}{N_t^+(s,a)}}, \tag{20}$$

matching equation (20) in the paper– and using the RSVD constants detailed earlier (Appendix A).

From (12), $\|E_{\text{shk}}\|_1=\|\epsilon_t(s,a)\|_1\le\|p_{t+1}-p_t\|_1$ since $p_{t+1}-p_t$ decomposes into the structured part plus $\epsilon_t(s,a)$ in ℓ_1.

Using (14), (15), (19), and (20), and absorbing the purely statistical term $\|E_{\text{emp}}\|_1$ into the $C\sqrt{KS\log/N_t^+}$ term (by enlarging C), we obtain

$$\left\|\widehat{p}_{t+1}^{\text{pred}}-p_{t+1}\right\|_1 \;\le\; \|p_{t+1}-p_t\|_1 \;+\; \beta K \;+\; C\sqrt{\frac{K\,S\log(8SAT/\delta)}{N_t^+(s,a)}},$$

which is exactly (11) and agrees with (11) in the appendix.

If additionally $\left\|\sum_k u_k(t)v_k(s,a)w_k\right\|_1\ge c_\star\sum_k|u_k(t)|$, then $\beta K\le(\beta K/c_\star)\,\|p_{t+1}-p_t\|_1$, so the additive βK term is dominated by a factor $(\beta K/c_\star)\|p_{t+1}-p_t\|_1$. □

E Shrinkage optimality: proof of Theorem 1

Theorem 1 (Near-optimal risk)**.** Let $\widehat{p}_t\in\Delta^{S-1}$ be the empirical transition estimate from N_t^+ samples for a fixed (s,a) at time t, and let $\widehat{p}_t^{\text{pred}}$ be any (possibly biased) forecast built from past data only. For $\lambda\in[0,1]$ define the shrinkage estimator $\widetilde{p}_t(\lambda)=(1-\lambda)\widehat{p}_t+\lambda\,\widehat{p}_t^{\text{pred}}$ and its ℓ_2-risk $R_t(\lambda):=\mathbb{E}\big[\|\widetilde{p}_t(\lambda)-p_t\|_2^2\big]$. Assume:

(A1) **Asymptotic orthogonality:** $\mathbb{E}\Big[\langle\widehat{p}_t-p_t,\;\widehat{p}_t^{\text{pred}}-p_t\rangle\Big]=o(1/N_t^+)$ (e.g. holds if the forecast uses only data independent of the N_t^+ samples that form $\widehat{p}_t$; sample splitting suffices).

(A2) **Bounded forecast risk:** $b_t:=\mathbb{E}\Big[\|\widehat{p}_t^{\text{pred}}-p_t\|_2^2\Big]$ is finite and bounded away from 0 along the considered times ($\inf_t b_t>0$ is enough).

(A3) **Consistent plug-in estimators:**

$$\widehat{a}_t \;:=\; \frac{1-\|\widehat{p}_t\|_2^2}{N_t^+} \;\xrightarrow{\;p\;}\; a_t := \mathbb{E}\big[\|\widehat{p}_t - p_t\|_2^2\big] = \frac{1-\|p_t\|_2^2}{N_t^+},$$

and, with a window $W_f \to \infty$,

$$\widehat{b}_t \;:=\; \frac{1}{W_f}\sum_{i=t-W_f}^{t-1}\Big(\|\widehat{p}_i^{\,\mathrm{pred}} - \widehat{p}_i\|_2^2 - \tfrac{1-\|\widehat{p}_i\|_2^2}{N_i^+}\Big) \;\xrightarrow{\;p\;}\; b_t.$$

Let the data-driven weight be $\widehat{\lambda}_t := \widehat{a}_t/(\widehat{a}_t+\widehat{b}_t)$ and the oracle weight be $\lambda_t^* := a_t/(a_t+b_t)$. Then, as $N_t^+ \to \infty$ and $W_f \to \infty$ (no rate relation between them is needed),

$$\frac{R_t(\widehat{\lambda}_t)}{R_t(\lambda_t^*)} \;=\; 1 + o(1).$$

Proof. **Step 1** Write $X_t := \widehat{p}_t - p_t$ and $Y_t := \widehat{p}_t^{\,\mathrm{pred}} - p_t$. By definition,

$$R_t(\lambda) = \mathbb{E}\big[\|(1-\lambda)X_t + \lambda Y_t\|_2^2\big] = (1-\lambda)^2 a_t + \lambda^2 b_t + 2\lambda(1-\lambda)c_t,$$

where $a_t = \mathbb{E}\|X_t\|_2^2$, $b_t = \mathbb{E}\|Y_t\|_2^2$ and $c_t = \mathbb{E}\langle X_t, Y_t\rangle$. Assumption (A1) gives $c_t = o(1/N_t^+)$, hence c_t is negligible relative to $a_t = \Theta(1/N_t^+)$. Therefore the minimizer is

$$\lambda_t^* = \frac{a_t - c_t}{a_t + b_t - 2c_t} = \frac{a_t}{a_t+b_t} + o(1/N_t^+)$$

and the oracle risk satisfies

$$R_t(\lambda_t^*) = \frac{(a_t-c_t)(b_t-c_t)}{a_t+b_t-2c_t} = \frac{a_t b_t}{a_t+b_t}\big(1+o(1)\big) \sim a_t \qquad (N_t^+ \to \infty),$$

since b_t is bounded away from 0 by (A2). In particular, $R_t(\lambda_t^*) = \Theta(1/N_t^+)$.

Step 2 Define $g(a,b) := a/(a+b)$. By (A3), $\widehat{a}_t \to a_t$ and $\widehat{b}_t \to b_t$ in probability, with $\widehat{a}_t - a_t = O_p(N_t^{-3/2})$ (delta method for $\widehat{a}_t = (1-\|\widehat{p}_t\|_2^2)/N_t^+$) and $\widehat{b}_t - b_t = O_p(W_f^{-1/2})$ (window average). A first-order expansion of g at (a_t, b_t) yields

$$\widehat{\lambda}_t - \lambda_t^* = \frac{\partial g}{\partial a}(a_t,b_t)\,(\widehat{a}_t - a_t) + \frac{\partial g}{\partial b}(a_t,b_t)\,(\widehat{b}_t - b_t) + o_p\Big(|\widehat{a}_t - a_t| + |\widehat{b}_t - b_t|\Big).$$

Because $\frac{\partial g}{\partial a} = \frac{b}{(a+b)^2} = \Theta(1)$ and $\frac{\partial g}{\partial b} = -\frac{a}{(a+b)^2} = \Theta(a_t) = \Theta(1/N_t^+)$,

$$\widehat{\lambda}_t - \lambda_t^* = O_p\Big(N_t^{-3/2}\Big) + O_p\Big((N_t^+)^{-1}W_f^{-1/2}\Big) = o_p\Big(N_t^{-1/2}\Big).$$

In particular, $\widehat{\lambda}_t \to \lambda_t^*$ in probability.[3]

Step 3 Since R_t is twice differentiable and $R_t'(\lambda_t^*) = 0$,

$$R_t(\widehat{\lambda}_t) - R_t(\lambda_t^*) = \tfrac{1}{2}\,R_t''(\xi_t)\,(\widehat{\lambda}_t - \lambda_t^*)^2, \qquad \xi_t \in \mathrm{conv}\big\{\widehat{\lambda}_t, \lambda_t^*\big\}.$$

Moreover, $R_t''(\lambda) = 2(a_t+b_t) - 4c_t = 2(b_t + o(1))$, hence $R_t''(\xi_t) = \Theta(1)$ by (A2) and (A1). Combining with Step 2,

$$R_t(\widehat{\lambda}_t) - R_t(\lambda_t^*) = O_p\Big(N_t^{-3}\Big) + O_p\Big((N_t^+)^{-2}W_f^{-1}\Big).$$

Finally, divide by $R_t(\lambda_t^*) = \Theta(1/N_t^+)$ from Step 1:

$$\frac{R_t(\widehat{\lambda}_t)}{R_t(\lambda_t^*)} - 1 = O_p\Big(N_t^{-2}\Big) + O_p\Big((N_t^+)^{-1}W_f^{-1}\Big) = o(1)$$

as soon as $W_f \to \infty$ (no relative rate to N_t^+ is needed). This proves the claim. □

[3] If one replaces $\widehat{b}_t$ by the uncorrected $\frac{1}{W_f}\sum_i \|\widehat{p}_i^{\,\mathrm{pred}} - \widehat{p}_i\|_2^2$, its limit is $b_t + a_t$; then $\widehat{\lambda}_t \to a_t/(a_t+b_t+a_t)$ differs from λ_t^* by $O(a_t) = O(1/N_t^+)$, hence still $\widehat{\lambda}_t - \lambda_t^* = o_p(N_t^{-1/2})$, giving the same conclusion.

Remark (on the plug-in MSE). The windowed proxy $\frac{1}{W_f}\sum_{i=t-W_f}^{t-1}\|\widehat{p}_i^{\text{pred}} - \widehat{p}_i\|_2^2$ converges to $b_t + a_t$ because $\mathbb{E}\|\widehat{p}_i - p_i\|_2^2 = a_t$ and the cross-term is $o(1)$ by (A1). Subtracting the known multinomial variance proxy $(1 - \|\widehat{p}_i\|_2^2)/N_i^+$ yields the consistent $\widehat{b}_t$ used in (A3). Using the uncorrected proxy leaves the theorem unchanged, since the induced bias in $\widehat{\lambda}_t$ is $O(a_t) = O(1/N_t^+)$ and the ratio $R_t(\widehat{\lambda}_t)/R_t(\lambda_t^*)$ still tends to 1.

F Full regret proof

Episode notation Episode m starts at $\tau(m)$, ends at $\tau(m+1) - 1$, and follows optimistic policy $\tilde{\pi}_m$.

F.1 Decomposition

For $t \in$ episode m

$$\rho_t^* - r_t \;\le\; \underbrace{(\rho_t^* - \tilde{\rho}_m)}_{A_t} + \underbrace{(\tilde{\rho}_m - \tilde{r}_m(s_t, a_t))}_{B_t} + \underbrace{(\tilde{r}_m - r_t)}_{C_t}.$$

Term $B_t \le 1/\sqrt{\tau(m)}$ by value-iteration tolerance. Terms A_t and C_t are bounded by variation $\text{var}_{\{r,p\}}$, statistical radii, widening η, and approximation approx exactly as in Lemma 5.

F.2 Summation over $t \le T$

1. Doubling episodes $\Rightarrow \sum B_t \le 2\sqrt{T \log T}$.
2. Reward/transition variation budget $\Rightarrow \sum \text{var}_{r,t} \le B_r$ and $\sum \text{var}_{p,t} \le B_p$.
3. Statistical radii: $\sum \text{rad}_r \le \widetilde{O}(\sqrt{SAT})$ and $\sum \text{rad}_p \le \widetilde{O}(\sqrt{SAT})$.
4. Widening: Lemma 4.
5. Approximation: RPCA + low-rank gives $O(\delta_B B_p + \sqrt{KT \log T})$.

Multiply the transition-related terms by $D_{\max}$, collect logarithms into $\widetilde{\mathcal{O}}$, and obtain Theorem 2. □

F.3 Detailed Regret Decomposition: Detail proof of Lemma 5

Lemma 5 (Per-step regret)**.** Fix an episode m with start time $\tau = \tau(m)$ and policy $\tilde{\pi}_m$ returned by EVI on the optimistic model constructed at time τ. Let $t \in [\tau, \tau(m+1) - 1]$. Define

$$\text{var}_{r,t} \;:=\; \max_{s,a} \big|r_t(s,a) - r_\tau(s,a)\big|, \qquad \text{var}_{p,t} \;:=\; \max_{s,a} \big\|p_t(\cdot|s,a) - p_\tau(\cdot|s,a)\big\|_1,$$

and let $\text{rad}_{r,\tau}$, $\text{rad}_{p,\tau}$ be the reward/transition statistical radii at time τ, $\eta = \eta(s_t, a_t, t)$ the adaptive widening, and approx the model-approximation slack. If EVI stops with tolerance $\epsilon_\tau := 1/\sqrt{\tau}$, then on a high-probability event of probability at least $1 - \delta/2$, for the action $a_t = \tilde{\pi}_m(s_t)$ we have

$$\rho_t^* - r_t(s_t, a_t) \;\le\; \epsilon_\tau + 2\,\text{var}_{r,t} + 2D_{\max}\,\text{var}_{p,t} + 2\,\text{rad}_{r,\tau} + 2D_{\max}\big(\text{rad}_{p,\tau} + \eta + \text{approx}\big).$$

Proof. **Good event and optimism.** At episode start τ we form confidence sets (Algorithm 3) around the shrinkage centre $\tilde{p}_\tau(\cdot|s,a)$:

$$\mathcal{C}_\tau(s,a;t) \;:=\; \Big\{p \in \Delta^{S-1} :\; \|p - \tilde{p}_\tau(\cdot|s,a)\|_1 \le \text{rad}_{p,\tau}(s,a) + \eta(s,a,t) + \text{approx}\Big\},$$

and reward intervals $[\underline{r}_\tau, \overline{r}_\tau]$ with half-width $\text{rad}_{r,\tau}$. By standard concentration (multinomial for p, Hoeffding for r) and the construction of η, there is an event $\mathcal{E}$ of probability $\ge 1 - \delta/2$ on which for all (s,a) and all $t \ge \tau$:

$$r_\tau(s,a) \in [\underline{r}_\tau(s,a), \overline{r}_\tau(s,a)], \quad p_t(\cdot|s,a) \in \mathcal{C}_\tau(s,a;t).$$

Let $\widetilde{M}_m = (\tilde{r}_m, \tilde{p}_m)$ be the optimistic MDP built at τ by picking $\tilde{r}_m(s,a) \in [\underline{r}_\tau, \overline{r}_\tau]$ and $\tilde{p}_m(\cdot|s,a) \in \mathcal{C}_\tau(s,a;t)$ so as to maximize the value (EVI). On $\mathcal{E}$ the true MDP at time τ lies in the (unwidened) sets, hence the optimism principle implies

$$\rho_\tau^* \;\le\; \tilde{\rho}_m, \tag{21}$$

where $\tilde{\rho}_m$ is the optimal average reward in $\widetilde{M}_m$.

A Lipschitz bound in average reward. $\rho^\pi(M)$ denotes the average (per-step) reward, also called the *gain*, of policy π in the MDP $M = (r, p)$. Formally, if $P^\pi(s, s') = p(s' \mid s, \pi(s))$ and $r^\pi(s) = r(s, \pi(s))$, then

$$\rho^\pi(M) = \lim_{T\to\infty} \frac{1}{T}\, \mathbb{E}_\pi\left[\sum_{t=0}^{T-1} r^\pi(s_t)\right] = \sum_s d^\pi(s)\, r^\pi(s),$$

where $d^\pi(s)$ denotes the stationary distribution of the Markov chain induced by policy π. For any communicating MDPs $M = (r, p)$ and $M' = (r', p')$ of diameter at most $D_{\max}$ and any stationary policy π,

$$\left|\rho^\pi(M) - \rho^\pi(M')\right| \;\le\; \|r - r'\|_\infty \;+\; D_{\max}\, \|p - p'\|_{1,\infty}. \tag{22}$$

Indeed, take a bias function h' for (M', π) with span $\mathrm{sp}(h') \le D_{\max}$ (obtainable by the standard hitting-time construction in communicating MDPs). The Poisson equation gives

$$\rho^\pi(M') + h'(s) = r'(s, \pi(s)) + p'(\cdot|s, \pi(s))^\top h'$$

for all s. Then

$$r(s,\pi(s)) + p(\cdot|s,\pi(s))^\top h' - h'(s) - \rho^\pi(M') = \underbrace{r(s,\pi(s)) - r'(s,\pi(s))}_{\le \|r-r'\|_\infty} + \underbrace{(p-p')(\cdot|s,\pi(s))^\top h'}_{\le \|p-p'\|_{1,\infty}\,\mathrm{sp}(h')}.$$

Taking the supremum over s and the infimum over h' (with $\mathrm{sp}(h') \le D_{\max}$) yields

$$\rho^\pi(M) \le \rho^\pi(M') + \|r - r'\|_\infty + D_{\max}\|p - p'\|_{1,\infty},$$

and exchanging M, M' proves (22).

From ρ_t^* to $\tilde{\rho}_m$. Apply (22) with $M_t = (r_t, p_t)$ and $M_\tau = (r_\tau, p_\tau)$ under the *optimal* policy at time t to obtain

$$\rho_t^* - \rho_\tau^* \;\le\; \mathrm{var}_{r,t} \;+\; D_{\max}\, \mathrm{var}_{p,t}. \tag{23}$$

Combining (23) with optimism (21) yields

$$\rho_t^* - \tilde{\rho}_m \;\le\; \mathrm{var}_{r,t} \;+\; D_{\max}\, \mathrm{var}_{p,t}. \tag{24}$$

EVI residual and one-step domination. Let $\tilde{h}_m$ be the bias function produced by EVI together with $\tilde{\pi}_m$ for the optimistic model. By the EVI stopping rule with tolerance $\epsilon_\tau = 1/\sqrt{\tau}$, for every state s,

$$\tilde{r}_m\big(s, \tilde{\pi}_m(s)\big) \;+\; \max_{p \in \mathcal{C}_\tau(s, \tilde{\pi}_m(s); t)} p^\top \tilde{h}_m \;-\; \tilde{h}_m(s) \;\ge\; \tilde{\rho}_m \;-\; \epsilon_\tau. \tag{25}$$

Evaluate (25) at $s = s_t$ and note that, on $\mathcal{E}$, the *true* row $p_t(\cdot|s_t, a_t)$ belongs to $\mathcal{C}_\tau(s_t, a_t; t)$. Hence

$$\tilde{\rho}_m - \tilde{r}_m(s_t, a_t) \;\le\; \epsilon_\tau \;+\; \big(p_t(\cdot|s_t, a_t)\big)^\top \tilde{h}_m \;-\; \tilde{h}_m(s_t). \tag{26}$$

Replace $\tilde{r}_m$ by r_t: reward part. Add and subtract $r_t(s_t, a_t)$ in (26) to get

$$\tilde{\rho}_m - r_t(s_t, a_t) \;\le\; \epsilon_\tau \;+\; \big(\tilde{r}_m - r_t\big)(s_t, a_t) \;+\; \big(p_t(\cdot|s_t, a_t)\big)^\top \tilde{h}_m - \tilde{h}_m(s_t).$$

Because $\tilde{r}_m(s, a) \in [\underline{r}_\tau(s, a), \overline{r}_\tau(s, a)]$ and $r_\tau(s, a)$ lies in the same interval, we have $\big|\tilde{r}_m(s, a) - r_\tau(s, a)\big| \le 2\,\mathrm{rad}_{r,\tau}(s, a)$; by definition of $\mathrm{var}_{r,t}$, $|r_\tau(s, a) - r_t(s, a)| \le \mathrm{var}_{r,t}$. Therefore

$$\big(\tilde{r}_m - r_t\big)(s_t, a_t) \;\le\; 2\,\mathrm{rad}_{r,\tau} \;+\; \mathrm{var}_{r,t}. \tag{27}$$

Replace $\tilde{p}_m$ by p_t: transition part. Insert and subtract $\tilde{p}_m(\cdot|s_t, a_t)$:

$$\big(p_t - \tilde{p}_m\big)^\top \tilde{h}_m \;+\; \tilde{p}_m^\top \tilde{h}_m \;-\; \tilde{h}_m(s_t).$$

The last two terms are nonpositive by (25) (they are upper-bounded by ϵ_τ already accounted for), so it suffices to bound the deviation term $\big|\big(p_t - \tilde{p}_m\big)^\top \tilde{h}_m\big| \le \mathrm{sp}(\tilde{h}_m)\, \|p_t - \tilde{p}_m\|_1 \le D_{\max}\|p_t - \tilde{p}_m\|_1$. By

construction, both $p_t(\cdot|s_t, a_t)$ and $\tilde{p}_m(\cdot|s_t, a_t)$ lie in the *same* ball $\|p - \tilde{p}_\tau\|_1 \le \mathrm{rad}_{p,\tau} + \eta + \mathrm{approx}$ around the centre $\tilde{p}_\tau(\cdot|s_t, a_t)$; hence

$$\|p_t - \tilde{p}_m\|_1 \;\le\; 2\big(\mathrm{rad}_{p,\tau} + \eta + \mathrm{approx}\big). \tag{28}$$

(We may additionally add $\mathrm{var}_{p,t}$, via $\|p_t - \tilde{p}_m\|_1 \le \|p_t - p_\tau\|_1 + \|p_\tau - \tilde{p}_m\|_1$, which only increases the bound; we keep the symmetric $2(\cdot)$ form induced by the common centre.)

Therefore

$$\big(p_t(\cdot|s_t, a_t)\big)^\top \tilde{h}_m - \tilde{h}_m(s_t) \;\le\; D_{\max} \cdot 2\big(\mathrm{rad}_{p,\tau} + \eta + \mathrm{approx}\big). \tag{29}$$

Collect the pieces. Combine (27)–(29) into (26):

$$\tilde{\rho}_m - r_t(s_t, a_t) \;\le\; \epsilon_\tau \;+\; 2\,\mathrm{rad}_{r,\tau} \;+\; \mathrm{var}_{r,t} \;+\; 2D_{\max}\big(\mathrm{rad}_{p,\tau} + \eta + \mathrm{approx}\big).$$

Finally add (24):

$$\rho_t^* - r_t(s_t, a_t) \;\le\; \epsilon_\tau + 2\,\mathrm{var}_{r,t} + 2D_{\max}\,\mathrm{var}_{p,t} + 2\,\mathrm{rad}_{r,\tau} + 2D_{\max}\big(\mathrm{rad}_{p,\tau} + \eta + \mathrm{approx}\big),$$

which is the claimed inequality.

□

F.4 Summation Analysis

We now analyze the sum of each term over all time steps $t \le T$.

Value Iteration Error Using the doubling nature of the episodes and the fact that episode lengths are at most $\sqrt{T}$, we have:

$$\begin{aligned}
\sum_{t=1}^{T} B_t &= \sum_{m=1}^{M} \sum_{t=\tau(m)}^{\tau(m+1)-1} \frac{1}{\sqrt{\tau(m)}} \\
&= \sum_{m=1}^{M} \frac{\tau(m+1) - \tau(m)}{\sqrt{\tau(m)}} \\
&\le \sum_{m=1}^{M} \frac{2\tau(m)}{\sqrt{\tau(m)}} \\
&= 2\sum_{m=1}^{M} \sqrt{\tau(m)} \\
&\le 2\sqrt{T} \cdot M
\end{aligned}$$

Since the number of episodes M is at most $\mathcal{O}(\log T)$ due to the doubling condition, we get $\sum B_t \le 2\sqrt{T \log T}$.

Variation Terms For the reward variation, we have:

$$\begin{aligned}
\sum_{t=1}^{T} \mathrm{var}_{r,t} &= \sum_{t=1}^{T} \max_{s,a} |r_t(s,a) - r_{\tau(m)}(s,a)| \\
&\le \sum_{t=1}^{T} \sum_{i=\tau(m)}^{t-1} \max_{s,a} |r_{i+1}(s,a) - r_i(s,a)| \\
&\le \sum_{i=1}^{T-1} \max_{s,a} |r_{i+1}(s,a) - r_i(s,a)| \cdot |\{t : i \ge \tau(m(t)), i < t\}|
\end{aligned}$$

Each transition i contributes to at most one episode, and by the definition of the variation budget, we have $\sum_{i=1}^{T-1} \max_{s,a} |r_{i+1}(s,a) - r_i(s,a)| \le B_r$. Therefore, $\sum_{t=1}^{T} \mathrm{var}_{r,t} \le B_r$.

A similar argument applies to the transition variation, giving $\sum_{t=1}^{T} \mathrm{var}_{p,t} \le B_p$.

Statistical Radii The statistical radius for rewards is defined as:

$$\mathrm{rad}_{r,t}(s,a) = \sqrt{\frac{2\log(4SAT/\delta)}{N_t(s,a)}}$$

Summing over all time steps and state-action pairs:

$$\begin{aligned}
\sum_{t=1}^{T} \mathrm{rad}_{r,\tau(m)}(s_t,a_t) &= \sum_{t=1}^{T} \sqrt{\frac{2\log(4SAT/\delta)}{N_{\tau(m)}(s_t,a_t)}} \\
&\leq \sqrt{2\log(4SAT/\delta)} \sum_{(s,a)} \sum_{n=1}^{N_T(s,a)} \frac{1}{\sqrt{n}} \\
&\leq \sqrt{2\log(4SAT/\delta)} \sum_{(s,a)} 2\sqrt{N_T(s,a)} \\
&\leq 2\sqrt{2\log(4SAT/\delta)} \sum_{(s,a)} \sqrt{N_T(s,a)}
\end{aligned}$$

By Cauchy-Schwarz:

$$\begin{aligned}
\sum_{(s,a)} \sqrt{N_T(s,a)} &\leq \sqrt{SA} \cdot \sqrt{\sum_{(s,a)} N_T(s,a)} \\
&= \sqrt{SA} \cdot \sqrt{T}
\end{aligned}$$

Therefore, $\sum_{t=1}^{T} \mathrm{rad}_{r,\tau(m)}(s_t,a_t) \leq \mathcal{O}(\sqrt{SAT\log(SAT/\delta)})$. A similar analysis applies to the transition radius, giving $\sum_{t=1}^{T} \mathrm{rad}_{p,\tau(m)}(s_t,a_t) \leq \mathcal{O}(\sqrt{S^2AT\log(SAT/\delta)})$.

Adaptive Widening By Lemma 4, we have:

$$\sum_{t=1}^{T} \eta(s_t,a_t,t) = \widetilde{\mathcal{O}}(D_{\max}\sqrt{(B_r+B_p)KST})$$

This bound exploits the low-rank structure of the environmental changes, resulting in a significant improvement over the uniform widening approach.

Approximation Error The approximation error comes from two sources: the randomized SVD and the incremental RPCA.

For the randomized SVD, by Lemma 2, the Frobenius norm error of the low-rank approximation is bounded by:

$$\|\mathbf{X}_t - \mathbf{U}\Sigma\mathbf{V}^T\|_F^2 \leq C_1 \min_{\mathrm{rank}\leq\hat{K}_t} \|\mathbf{X}_t - \mathbf{A}\|_F^2$$

where $C_1 = (2+4\sqrt{(\hat{K}_t+s)/(s-1)})^{4/(2q+1)}$.

For the incremental RPCA, by Proposition 1, the error in recovering the low-rank and sparse components is bounded by:

$$\max_{t\leq T} \|\Delta P_t^{\mathrm{L}} + \Delta P_t^{\mathrm{S}} - \Delta P_t\|_F \leq C_2\sqrt{\frac{K^2(SA+S)\log(SA/\delta)}{SA}}$$

where C_2 is a constant.

For the sparse component, we have the bound $\sum_t \max_{s,a} \|\epsilon_t(s,a,\cdot)\|_1 \leq \delta_B B_p$ from Assumption 1.

Combining these sources of error and summing over all time steps, we get:

$$\sum_{t=1}^{T} \mathrm{approx}_t = \mathcal{O}(\delta_B B_p + \sqrt{KT\log(T)})$$

F.5 Final Regret Bound

Combining all the terms and multiplying the transition-related terms by $D_{\max}$, we get:

$$\begin{aligned}\mathrm{DynReg}_T &= \sum_{t=1}^{T}(\rho_t^* - r_t(s_t, a_t)) \\ &\leq 2\sqrt{T\log T} + 2B_r + 2D_{\max}B_p + \mathcal{O}(\sqrt{SAT\log(SAT/\delta)}) + \mathcal{O}(D_{\max}\sqrt{S^2AT\log(SAT/\delta)}) \\ &\quad + \widetilde{\mathcal{O}}(D_{\max}\sqrt{(B_r+B_p)KST}) + \mathcal{O}(D_{\max}\delta_B B_p + D_{\max}\sqrt{KT\log(T)})\end{aligned}$$

The dominant terms are the statistical error $\mathcal{O}(D_{\max}\sqrt{SAT\log(SAT/\delta)})$ and the adaptive widening $\widetilde{\mathcal{O}}(D_{\max}\sqrt{(B_r+B_p)KST})$. Collecting the logarithmic factors into $\widetilde{\mathcal{O}}$, we get the regret bound stated in Theorem 2:

$$\mathrm{DynReg}_T = \widetilde{\mathcal{O}}(D_{\max}\sqrt{SAT} + D_{\max}\sqrt{(B_r+B_p)KST} + D_{\max}\delta_B B_p)$$

F.6 Optimality of the Regret Bound

The regret bound we obtain is nearly optimal in several aspects:

Dependence on T The $\sqrt{T}$ dependence matches the lower bound for non-stationary bandits with variation budget constraints, which is $\Omega(\sqrt{BT})$ where B is the variation budget. This suggests that our algorithm achieves the optimal rate in terms of the time horizon.

Dependence on state-action space The first term $D_{\max}\sqrt{SAT}$ matches the lower bound for reinforcement learning in stationary environments, which is $\Omega(D_{\max}\sqrt{SAT})$. This indicates that our algorithm achieves the optimal dependence on the state and action space sizes in the absence of non-stationarity.

Dependence on variation budgets The second term $D_{\max}\sqrt{(B_r+B_p)KST}$ shows that the regret scales with the square root of the variation budgets, which is optimal under the standard model of non-stationarity.

Dependence on rank K The dependence on the rank K is an improvement over previous algorithms that did not exploit low-rank structure. The factor $\sqrt{K}$ replaces the factor $\sqrt{SA}$ in the non-stationary term, resulting in a significant reduction in regret when $K \ll SA$.

Residual term The residual term $D_{\max}\delta_B B_p$ accounts for the sparse shock component in our model. This term can be made arbitrarily small by setting δ_B to a small value, at the cost of potentially increasing the rank K to capture more of the variation.

F.7 Comparison to Previous Results

Our regret bound improves upon the regret bounds of previous algorithms for non-stationary reinforcement learning:

SWUCRL2-CW The sliding-window algorithm with uniform confidence widening achieves a regret bound of $\widetilde{\mathcal{O}}(D_{\max}(SAT)^{1/3}(B_r+B_p)^{2/3})$ or $\widetilde{\mathcal{O}}(D_{\max}S\sqrt{AT} + D_{\max}\sqrt{SAT(B_r+B_p)})$. Our algorithm improves the dependence on T from $T^{3/4}$ to $\sqrt{T}$ and reduces the dependence on the state-action space from SA to K in the non-stationary term.

Bandit-based approaches Non-stationary bandit algorithms typically achieve regret bounds of the form $\widetilde{\mathcal{O}}(\sqrt{KBT})$ where K is the number of arms and B is the variation budget. Our algorithm generalizes this to the reinforcement learning setting while maintaining the optimal dependence on the time horizon and variation budgets.

In summary, our regret bound represents a significant improvement over existing results for non-stationary reinforcement learning, particularly in environments with low-rank structure in the dynamics changes.

G Detailed algorithm implementation

G.1 Confidence interval construction

The confidence intervals for rewards and transitions are constructed as follows:

Reward confidence interval For each state-action pair (s, a), we define the confidence interval for the reward at time t as:

$$[\underline{r}_t(s,a), \overline{r}_t(s,a)] = [\widehat{r}_t(s,a) - \text{rad}_{r,t}(s,a), \widehat{r}_t(s,a) + \text{rad}_{r,t}(s,a)]$$

where $\widehat{r}_t(s,a)$ is the empirical average reward for (s,a) up to time t, and the confidence radius is:

$$\text{rad}_{r,t}(s,a) = \sqrt{\frac{2\log(4SAT/\delta)}{N_t(s,a)}}$$

Transition confidence interval For the transition probabilities, we define the confidence set at time t as:

$$\mathcal{P}_t(s,a) = \{p : \|p - \tilde{p}_t(\cdot|s,a)\|_1 \leq \text{rad}_{p,t}(s,a) + \eta(s,a,t)\}$$

where $\tilde{p}_t(\cdot|s,a)$ is the shrinkage estimator defined in Section 7, and the confidence radius has two components:

- $\text{rad}_{p,t}(s,a) = \sqrt{\frac{2S\log(4SAT/\delta)}{N_t(s,a)}}$ accounts for statistical uncertainty
- $\eta(s,a,t) = \min\{1, c\sqrt{\widehat{V}(s,a,t)/N_t^+(s,a)}\}$ accounts for non-stationarity

G.2 Extended Value Iteration

The Extended Value Iteration (EVI) algorithm computes an optimistic policy as follows:

Algorithm 4 Extended Value Iteration

Require: Confidence sets $\{[\underline{r}_t(s,a), \overline{r}_t(s,a)]\}$, $\{\mathcal{P}_t(s,a)\}$, tolerance ϵ

Initialize $V_0(s) = 0$ for all $s \in \mathcal{S}$
$span \leftarrow \infty$
while $span > \epsilon$ **do**
  for $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $Q_k(s,a) \leftarrow \overline{r}_t(s,a) + \max_{p \in \mathcal{P}_t(s,a)} \sum_{s'} p(s')V_k(s')$
    end for
    $V_{k+1}(s) \leftarrow \max_a Q_k(s,a)$
    $\pi(s) \leftarrow \arg\max_a Q_k(s,a)$
  end for
  $span \leftarrow \max_s V_{k+1}(s) - \min_s V_{k+1}(s)$
end while
return π, $span$

The inner maximization $\max_{p \in \mathcal{P}_t(s,a)} \sum_{s'} p(s')V_k(s')$ can be solved efficiently by assigning as much probability as possible to the states with the highest values, subject to the constraint that p must be within distance $\text{rad}_{p,t}(s,a) + \eta(s,a,t)$ of $\tilde{p}_t(\cdot|s,a)$ in ℓ_1 norm.

G.3 Factor tracking and forecasting

The algorithm maintains a buffer of recent transition changes and periodically updates the low-rank model. The key steps are:

Buffer update At each time step, we update the empirical transition estimates and compute the change:

$$\Delta\widehat{P}_t = \widehat{P}_t - \widehat{P}_{t-1}$$

This change is added to a circular buffer of size W.

Low-rank model update Every W time steps, we:

1. Form the matrix $\mathbf{X}_t = [\Delta\widehat{P}_{t-W+1}, \ldots, \Delta\widehat{P}_t]$
2. Run Algorithm 1 (Randomized SVD) to obtain factors $\mathbf{U}, \Sigma, \mathbf{V}$
3. Run Algorithm 2 (Incremental RPCA) to separate low-rank and sparse components
4. Extract time-varying coefficients $\widehat{u}_k(t-W+1), \ldots, \widehat{u}_k(t)$ for each factor k

Forecasting For each factor k, we:

1. Fit multiple time-series models to the sequence $\widehat{u}_k(t-W+1), \ldots, \widehat{u}_k(t)$:
   - Exponential smoothing: $\widehat{u}_k^{\text{ES}}(t+1) = \alpha u_k(t) + (1-\alpha)\widehat{u}_k^{\text{ES}}(t)$
   - AR(1): $\widehat{u}_k^{\text{AR1}}(t+1) = \phi_0 + \phi_1 u_k(t)$
   - AR(2): $\widehat{u}_k^{\text{AR2}}(t+1) = \phi_0 + \phi_1 u_k(t) + \phi_2 u_k(t-1)$
2. Select the model with the lowest AIC
3. Generate the prediction $\widehat{u}_k^{\text{pred}}(t+1)$

Shrinkage estimation To compute the shrinkage weight λ for combining empirical and predicted estimates:

1. Estimate the variance of the empirical transition probabilities:

$$\widehat{\text{Var}}[\widehat{p}_t] \approx \frac{\widehat{p}_t(1-\widehat{p}_t)}{N_t^+}$$

2. Estimate the MSE of the prediction based on recent performance:

$$\widehat{\text{MSE}}[\widehat{p}_t^{\text{pred}}] \approx \frac{1}{W_f} \sum_{i=t-W_f}^{t-1} (\widehat{p}_i^{\text{pred}} - \widehat{p}_i)^2$$

3. Compute the shrinkage weight:

$$\lambda = \frac{\widehat{\text{Var}}[\widehat{p}_t]}{\widehat{\text{Var}}[\widehat{p}_t] + \widehat{\text{MSE}}[\widehat{p}_t^{\text{pred}}]}$$

4. Combine the estimates:

$$\tilde{p}_t = (1-\lambda)\widehat{p}_t + \lambda\widehat{p}_t^{\text{pred}}$$

G.4 Implementation Optimizations

Several optimizations can improve the computational efficiency of SVUCRL:

Sparse matrix operations For large state spaces, the transition matrices are often sparse. Using sparse matrix operations can significantly reduce memory usage and computation time. The randomized SVD and incremental RPCA algorithms can be adapted to work with sparse matrices, exploiting the sparsity structure.

Lazy updates Since the low-rank model is updated only every W time steps, many intermediate computations can be deferred. For example, the empirical transition matrices can be updated incrementally, and the full matrix is only formed when needed for the model update.

Parallel computation Many parts of the algorithm can be parallelized:

- The randomized SVD algorithm can leverage parallel matrix-matrix multiplications
- The confidence interval constructions for different state-action pairs can be done in parallel
- The forecasting of different factors can be computed independently

Adaptive rank selection Instead of using a fixed rank $\hat{K}$, we can adaptively determine the rank based on the singular value spectrum:

$$\hat{K}_t = \min\left\{k : \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{\min(SA,WS)} \sigma_i^2} \geq \gamma\right\}$$

where γ is a threshold (e.g., $\gamma = 0.95$).

Efficient EVI implementation The Extended Value Iteration can be optimized by:

- Caching the optimistic transitions for each state-action pair
- Using priority queue-based updates to focus computation on states with significant value changes
- Warm-starting each EVI run with the value function from the previous episode

G.5 Parameter Selection Guidelines

The performance of SVUCRL depends on several parameters. We provide guidelines for setting these parameters:

Structure update window W The window size W controls the frequency of updating the low-rank model. It should be large enough to provide sufficient data for learning the factors, but small enough to track changes in the environment. A reasonable choice is $W = \Theta(\sqrt{T})$.

Variation estimation window W_v The window W_v determines the time scale for estimating local variation. It should be chosen based on the expected rate of change in the environment. For environments with smooth changes, larger values (e.g., $W_v = \Theta(\sqrt{T})$) are appropriate. For more volatile environments, smaller values (e.g., $W_v = \Theta(\log T)$) may be better.

Forecasting window W_f The window W_f sets the horizon for evaluating prediction performance. It should be large enough to provide reliable MSE estimates but small enough to adapt to changing prediction accuracy. A reasonable choice is $W_f = \Theta(W_v)$.

Power iterations q The number of power iterations in the randomized SVD affects the accuracy of the low-rank approximation. For most applications, $q = 1$ or $q = 2$ provides a good balance between accuracy and computation. For matrices with slowly decaying singular values, larger values may be necessary.

Oversampling s The oversampling parameter in the randomized SVD should be set to $s \geq 3$. Larger values improve accuracy at the cost of computation. A typical choice is $s = 5$ or $s = 10$.

Confidence parameter δ The confidence parameter δ controls the failure probability of the confidence intervals. It should be set to a small value, typically $\delta = 0.1/T$ or $\delta = 0.01/T$.

Target rank $\hat{K}$ If not using adaptive rank selection, a conservative choice is $\hat{K} = \min\{10, \sqrt{SA}\}$. This captures most of the structure while keeping the computation manageable.

These guidelines provide a starting point for parameter selection, but the optimal values may depend on the specific characteristics of the environment. In practice, a parameter sweep or online adaptation may be necessary to achieve the best performance.

H Limitations

Despite its theoretical appeal, **SVUCRL** has several important limitations that warrant future investigation:

1. **Low–rank assumption**. Our regret guarantees hinge on Assumption 1, i.e. that *most* non-stationarity lies in a rank–$K \ll SA$ subspace. Highly entangled or full-rank drift can break the $\sqrt{KST}$ term and lead to vacuous bounds.
2. **Sparse–shock model**. The incremental RPCA step presumes that abrupt changes are sparse across state–action pairs. Large-scale shocks (e.g. global re-parameterisations) violate this sparsity and may induce large approximation errors, inflating confidence widths.
3. **Parameter sensitivity**. Windows (W, W_v, W_f), oversampling s, power iterations q and the shrinkage threshold all require tuning. Poorly chosen values can negate the theoretical gains and incur additional regret; an adaptive, provably robust selection rule is still missing.
4. **Computational overhead**. Although §8 exploits randomized SVD and streaming updates, the per-update cost is $\mathcal{O}(TSA(SK + S)\log T)$—substantial for very large S or dense transition tensors. Scaling to high-dimensional continuous spaces will need function approximation or sketching techniques beyond the present scope.

These caveats highlight directions for extending SVUCRL towards more realistic and large-scale reinforcement-learning settings.