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

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

Motivation & Problem Setup

Non-Stationary RL. Sequence of communicating MDPs $(\mathcal{S}, \mathcal{A}, p_t, r_t)_{t=1}^T$ with diameter $D_{\max}$ where transitions and rewards evolve over time.

Variation Budgets (quantify non-stationarity):

$$B_r = \sum_t \max_{s,a} |r_{t+1}(s,a) - r_t(s,a)|,$$
$$B_p = \sum_t \max_{s,a} \|p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a)\|_1.$$

Dynamic Regret (performance metric):

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

where ρ_t^* is the optimal average reward with transition p_t and mean reward r_t.

Key Challenge. Track changing optimal policies without discarding useful history; adapt quickly while maintaining tight confidence sets.

Structured Variation Model

Low-Rank Drift + Sparse Shocks. For transition change $\Delta P_t \in \mathbb{R}^{(SA)\times S}$:

$$\Delta P_t = \sum_{k=1}^K u_k(t)\, v_k\, w_k^\top + \epsilon_t$$

- $u_k(t)$: time weight of factor k
- v_k: pattern over state-action pairs
- w_k: reallocation pattern over next states
- ϵ_t: sparse localized shocks

Constraints: $\|w_k\|_1 \le 1$, $|v_k(s,a)| \le 1$, and

$$\sum_t \max_{s,a} \|\epsilon_t(s,a,\cdot)\|_1 \le \delta_B B_p$$

Why This Helps. Few drivers ($K \ll SA$) move many rows jointly $\Rightarrow$ uncertainty concentrates on K-dimensional subspace $\Rightarrow$ regret scales with $\sqrt{K}$ not $\sqrt{SA}$.

Key Contributions

1. **SVUCRL algorithm** exploiting low-rank drift structure + isolating sparse shocks
2. **Online low-rank tracking** via randomized SVD with Frobenius guarantees
3. **Incremental RPCA** for drift/shock separation with per-step error control
4. **Adaptive confidence widening** via bias-corrected local-variation estimator
5. **Factor forecasting + shrinkage** for low-variance transition centers
6. $\widetilde{O}(\sqrt{T})$ **regret** matching conjectured optimal rates

SVUCRL Algorithm (High-Level)

Inputs: windows W, W_v, W_f; confidence δ

Main Loop ($t = 1 \ldots T$):

1. **Act:** Play $\tilde{\pi}(s_t)$; observe r_t, s_{t+1}
2. **Update:** Empirical $\hat{r}, \hat{p}$; store $\Delta\widehat{P}_t$
3. **Structure** (every W steps):
   - Run **RSVD** on recent changes
   - Apply **RPCA** to separate drift/shocks
   - Extract factors $\{v_k, w_k\}$ and time weights
4. **Forecast:** One-step prediction of $u_k(t+1)$
5. **Shrink:** Combine forecast + empirical via James-Stein
6. **Widen:** Local variation $\Rightarrow$ adaptive $\eta(s,a,t)$
7. **Replan:** When episode ends, run EVI with optimistic model

Technical Components

1. Randomized SVD (low-rank drift)

- Track a rank-K approximation of recent transition changes over a sliding window.
- Near-best rank-K approximation in Frobenius norm:

$$\|X_t - U\Sigma V^\top\|_F^2 \le C \min_{\mathrm{rank}(A)\le K} \|X_t - A\|_F^2.$$

2. Incremental RPCA (drift vs. shocks)

- Decompose each change as $\Delta P_t = \hat{L}_t + \hat{S}_t$.
- $\hat{L}_t$: smooth low-rank drift; $\hat{S}_t$: sparse, localized shocks.
- Per-step cost $O(SASK)$ with high-probability Frobenius error control.

3. Bias-Corrected Local Variation

- Short window W_v to estimate how fast $p_t(\cdot \mid s,a)$ moves.
- Bias-corrected estimator $\widehat{V}(s,a,t)$ removes sampling noise.
- Total widening:

$$\sum_t \eta(s_t, a_t, t) \le \widetilde{O}\big(\sqrt{SA\,B_p}\big).$$

Complexity & Parameters

Time Complexity: $\mathcal{O}\big(TSA(SK+S)\log T\big)$
Space Complexity: $O((SA+S+W)K + SAW)$
Recommended Settings (high-level):

- Windows: $W = \Theta(\sqrt{T})$, $W_v = \Theta(\log T)$, $W_f = \Theta(\sqrt{W})$
- RSVD: randomized SVD with a small fixed number of power iterations
- Rank: chosen adaptively from spectrum (e.g. 95% energy cutoff)
- Episodes: standard UCRL2-style doubling rule

Main Theoretical Result

Dynamic Regret Bound (w.h.p. $1-\delta$):

$$\mathrm{DynReg}_T = \widetilde{\mathcal{O}}\Big(D_{\max} S\sqrt{AT} + D_{\max}\sqrt{(B_r{+}B_p)\,K\,S\,T} + D_{\max}\,\delta_B\,B_p\Big)$$

Three Terms Explained:

1. $D_{\max} S\sqrt{AT}$: the standard statistical error for learning environment dynamics.
2. $D_{\max}\sqrt{(B_r{+}B_p)KST}$: **Non-stationarity** regret with $\sqrt{K}$ instead of full SA
3. $D_{\max}\delta_B B_p$: **Residual** for sparse shocks (negligible if δ_B small)

Key Improvement: $\sqrt{T}$ dependence vs. prior $T^{3/4}$ bounds; $\sqrt{K}$ vs. $\sqrt{SA}$ when drift is low-rank.

Shrinkage & Forecasting

Factor Forecasting:

- For each factor k, forecast $u_k(t+1)$ using simple time-series models (exponential smoothing / AR models).
- Select the best forecasting model per factor on a short validation window.

James–Stein Shrinkage (key idea):

$$\tilde{p}_{t+1} = (1-\lambda_t)\,\hat{p}_{t+1} + \lambda_t\,\hat{p}_{t+1}^{\mathrm{pred}}.$$

- $\hat{p}_{t+1}$: empirical transition estimate; $\hat{p}_{t+1}^{\mathrm{pred}}$: model-based forecast.
- λ_t trades off empirical variance vs. forecast bias and is estimated from data.
- As samples grow, risk approaches that of the best (oracle) combination.

Comparison to Prior Work

SWUCRL2-CW [Cheung et al. 2020]:

$$\widetilde{O}(D_{\max}(B_r+B_p)^{1/4} S^{2/3} A^{1/2} T^{3/4})$$

SVUCRL (Ours):

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

Improvements:

- $\sqrt{T}$ vs. $T^{3/4}$ dependence
- $\sqrt{K}$ vs. $S^{2/3}A^{1/2}$ in non-stationary term
- Matches $\sqrt{T}$ lower bounds (up to logs)
- Exploits structure when $K \ll SA$

When SVUCRL Works Best

Ideal Conditions:

- Drift is truly low-rank ($K \ll SA$)
- Occasional localized shocks (sparse ϵ_t)
- Smooth factor evolution (good forecasting)
- Sufficient visits for shrinkage

Parameter Guidelines:

- Choose $W, W_v, W_f \propto \sqrt{T}$, then adjust based on observed change rate
- Use a small fixed number of power iterations and mild oversampling in RSVD
- Adaptive rank via 95% energy or a clear spectral gap
- Episode-doubling for EVI triggers (as in UCRL2)

Limitations & Future Work

Current Limitations:

1. **Model mismatch:** Full-rank drift or dominant shocks reduce advantage
2. **Assumptions:** Requires incoherence for RPCA; sparse support constraints
3. **Tuning:** Windows, rank, RSVD parameters need selection
4. **Scale:** Very large (S, A) may need function approximation
5. **Theory-practice gap:** No empirical validation yet

Key Takeaways

1. **Structure matters:** Exploiting low-rank drift improves regret from $T^{3/4}$ to $\sqrt{T}$
2. **Dimension reduction:** $\sqrt{K}$ factor vs. $\sqrt{SA}$ enables scaling
3. **Optimal rates:** Matches conjectured lower bounds.
4. **Practical tools:** RSVD, RPCA, shrinkage, adaptive widening all contribute
5. **Open question:** Empirical performance in real applications

References & Contact

Key References:

- Cheung et al. (2020): SWUCRL2-CW
- Halko et al. (2011): Randomized SVD
- Candès et al. (2011): Robust PCA
- Jaksch et al. (2010): UCRL2

Contact: tuandq@soict.hust.edu.vn
Acknowledgments: Funded by HUST Project No. T2024-TD-024