Explore-then-Commit for Nonstationary Linear Bandits with Latent Dynamics

Sunmook Choi [1] Yahya Sattar [1] Yassir Jedra [2] Maryam Fazel [3] Sarah Dean [1]

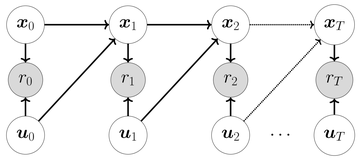
[1]Cornell University [2]Imperial College London [3]University of Washington

CLASSICAL BANDITS vs. OUR BANDITS

Setting: Bandits with Latent Dynamics

$$r_t = \boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{x}_t + z_t, \qquad \boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t + \boldsymbol{w}_t.$$

- Only rewards are observed; the state $\boldsymbol{x}_t$ is latent.
- $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are unknown and $\rho(\boldsymbol{A}) < 1$.
- Action $\boldsymbol{u}_t$ changes both current reward and future rewards.



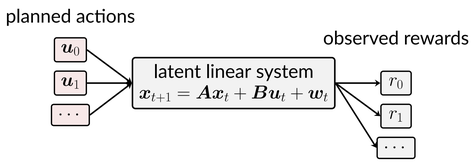
Objective & Benchmark

- **Goal**: choose $\boldsymbol{u}_t \in [-1, +1]^p$ to maximize $\mathbb{E}\left[\sum_{t=0}^{T} r_t\right]$
- **Regret** compares against the best open-loop sequence:

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} r_t^\star - \sum_{t=0}^{T} r_t^\pi\right].$$

Open-loop Policy

- **Open-loop**: Choose a full action sequence $\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_T$ without reacting to realized r_t.



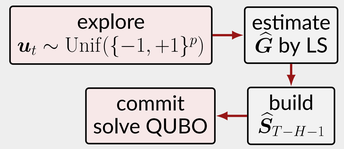
Optimal Open-loop Action Sequence from QUBO

Unrolling the dynamics gives a quadratic value:

$$\mathbb{E}\left[\sum_{t=0}^{T} r_t\right] = \frac{1}{2}\boldsymbol{u}_{0:T}^\top \boldsymbol{S}_T \boldsymbol{u}_{0:T}, \qquad \boldsymbol{u}_{0:T}^\star \in \arg\max_{\boldsymbol{u}_t \in [-1,1]^p} \frac{1}{2}\boldsymbol{u}_{0:T}^\top \boldsymbol{S}_T \boldsymbol{u}_{0:T}.$$

- An optimal sequence exists at a vertex: $\boldsymbol{u}_t^\star \in \{-1, +1\}^p, \forall t$.
- The resulting QUBO is NP-hard.

Algorithm: Explore-then-Commit



Require: Horizon T, exploration length H, truncation length L

1: Play random actions for $t = 0, \ldots, H$ and observe rewards.
2: Estimate the first L Markov parameters $\hat{\boldsymbol{G}}$.
3: Optimize the estimated commit objective

$$\max_{\boldsymbol{u}_{H+1:T} \in \{-1,+1\}^{p(T-H)}} \frac{1}{2}\boldsymbol{u}_{H+1:T}^\top \widehat{\boldsymbol{S}}_{T-H-1} \boldsymbol{u}_{H+1:T}.$$

4: Play the chosen sequence for $t = H+1, \ldots, T$.

Learning How Actions Affect Rewards

Markov Parameters characterize past action-reward relationship.

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{C}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{A}\boldsymbol{B} & \cdots & \boldsymbol{C}\boldsymbol{A}^{L-1}\boldsymbol{B} \end{bmatrix}$$

- We only consider first L of them due to stability assumption.

For $t \geq L$, the reward r_t depends on the past L actions [3]:

$$r_t = \mathbf{vec}(\boldsymbol{G})^\top (\bar{\boldsymbol{u}}_{t-1} \otimes \boldsymbol{u}_t) + \text{noise} + \mathcal{O}(\rho^L).$$

where $\bar{\boldsymbol{u}}_{t-1} = \begin{bmatrix} \boldsymbol{u}_{t-1}^\top & \boldsymbol{u}_{t-2}^\top & \cdots & \boldsymbol{u}_{t-L}^\top \end{bmatrix}^\top$. From $\{(\boldsymbol{u}_t, r_t)\}_{t=0}^{H}$,

$$\widehat{\boldsymbol{G}} = \arg\min_{\boldsymbol{G} \in \mathbb{R}^{p \times pL}} \sum_{t=L+1}^{H} \left(r_t - \mathbf{vec}(\boldsymbol{G})^\top (\bar{\boldsymbol{u}}_{t-1} \otimes \boldsymbol{u}_t)\right)^2.$$

Estimation guarantee With high probability,

$$H - L \gtrsim \tilde{\mathcal{O}}(p^2 L) \implies \|\hat{\boldsymbol{G}} - \boldsymbol{G}\|_F \lesssim \tilde{\mathcal{O}}\left(\sqrt{\frac{p^2 L}{H - L}}\right)$$

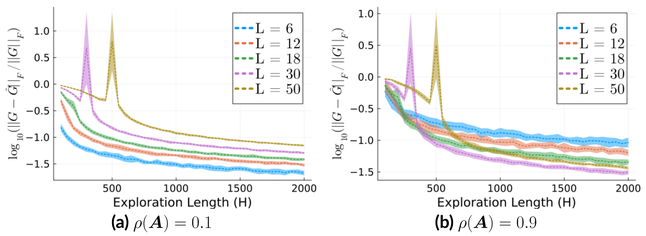


Figure 1. Markov Parameter Estimation Error

Regret Analysis Outline

Suboptimality Lemma. If $\|\widehat{\boldsymbol{G}} - \boldsymbol{G}\|_F \leq \epsilon$, the estimated QUBO chooses a near-optimal open-loop sequence:

$$\text{commit suboptimality} \lesssim p(T-H)\left(\epsilon + \rho^L\right).$$

Regret decomposition:

$$R_T(\pi) \lesssim \underbrace{pH}_{\text{exploration}} + \underbrace{pT\epsilon}_{\text{estimation}} + \underbrace{pT\rho^L}_{\text{truncation}} \quad \text{w.h.p.}$$

Taking $\epsilon \simeq \tilde{\mathcal{O}}(H^{-1/2})$ and $L = \Theta(\log T)$:

$$R_T(\pi) \lesssim p\left(H + \frac{T}{\sqrt{H}}\right).$$

Main Regret Guarantee

Theorem. With $H = \tilde{\mathcal{O}}(T^{2/3})$ and $L = \Theta(\log T)$, the explore-then-commit policy satisfies

$$R_T(\pi) = \tilde{\mathcal{O}}(pT^{2/3})$$

with high probability.

- H controls the explore/exploit tradeoff.
- L makes the truncation negligible.

Practical QUBO Approximation

The commit problem is an NP-hard QUBO

$$\max_{\boldsymbol{x} \in \{\pm 1\}^d} \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \quad \text{with } d = p(T-H).$$

Semidefinite Relaxation with Goemans-Williamson Rounding:

$$\max_{\boldsymbol{X} \succeq 0} \mathrm{tr}(\boldsymbol{W}\boldsymbol{X}) \quad \text{s.t. } \mathrm{rank}(\boldsymbol{X}) = 1, \ \boldsymbol{X}_{ii} = 1.$$

Solve SDP with Mosek [1]; then random-hyperplane round [2].

SignIter: Fast heuristic method; useful when the SDP is too large.

$$\boldsymbol{x}_i^{k+1} = \mathrm{sign}\left((\boldsymbol{W}\boldsymbol{x}^k)_i\right).$$

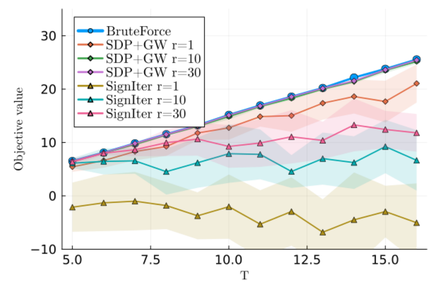


Figure 2. Small-dimensional comparison against brute force.

Numerical Results

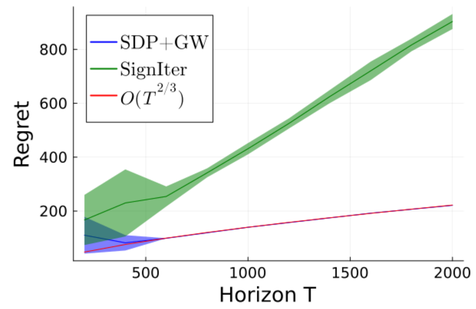


Figure 3. Regret of ETC using SDP+GW and SignIter.

- SDP+GW gives the stronger commit-phase benchmark.
- SignIter is scalable but can be more suboptimal.
- Empirical growth is consistent with theoretical regret.

Takeaways

- Regret is measured against the best open-loop action sequence.
- Random exploration learns the action-to-reward history.
- Approximate planning is enough for $\tilde{\mathcal{O}}(pT^{2/3})$ regret.

Future Work

- Adaptive closed-loop policies that use reward feedback.
- Faster QUBO solvers for larger horizons and actions.

References

[1] MOSEK ApS.
MOSEK Optimizer API for Julia 11.0.29, 2025.

[2] Michel X Goemans and David P Williamson.
Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming.
Journal of the ACM (JACM), 42(6):1115–1145, 1995.

[3] Yahya Sattar, Yassir Jedra, and Sarah Dean.
Learning linear dynamics from bilinear observations.
In 2025 American Control Conference (ACC), pages 3109–3115. IEEE, 2025.