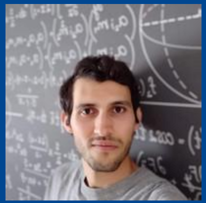
High-dimensional bandits:

when is SVD provably all you need?

Stefan Stojanovic (stesto@kth.se)

joint work with Y. Jedra, W. Reveillard, A. Proutiere



Winter Conference in Statistics 2024, Hemavan

Coming next

- Intro to bandits (finite-armed/linear)
- Bandits with hidden low dimensional structure
- Entry-wise matrix recovery using SVD

} find structure & reduce to simpler low dim. problem

Entry-wise guarantees for SVD provide framework for obtaining tightest known regret bounds $O(d^{3/4}\sqrt{T})$ for low-rank bandits!

Main references

A. Stojanovic, Stefan, Yassir Jedra, and Alexandre Proutiere. "Spectral entry-wise matrix estimation for low-rank reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).

B. Jedra, Yassir*, William Reveillard*, Stefan Stojanovic*, Alexandre Proutiere. "Low-Rank Bandits via Tight Two-to-Infinity Singular Subspace Recovery." *arXiv preprint arXiv:2402.15739* (2024).

What is not coming next

- Best arm identification
- Policy evaluation

} minimax optimal algorithms for low-rank bandits in B.

- Mathematical rigour
- Monologue?

Open problems

- Achieving upper regret bound of $O(\sqrt{dT})$
- Reconciling adaptivity and low-dimensional structure recovery

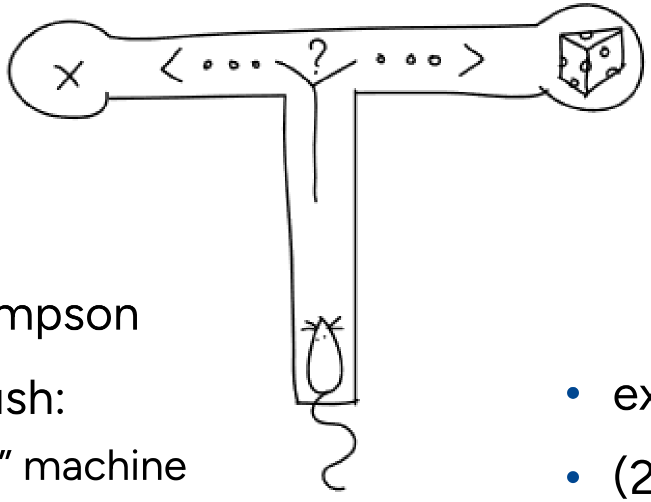
Main references

A. Stojanovic, Stefan, Yassir Jedra, and Alexandre Proutiere. "Spectral entry-wise matrix estimation for low-rank reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).

B. Jedra, Yassir*, William Reveillard*, Stefan Stojanovic*, Alexandre Proutiere. "Low-Rank Bandits via Tight Two-to-Infinity Singular Subspace Recovery." *arXiv preprint arXiv:2402.15739* (2024).

Sequential decision making with uncertainty

...or simply bandits

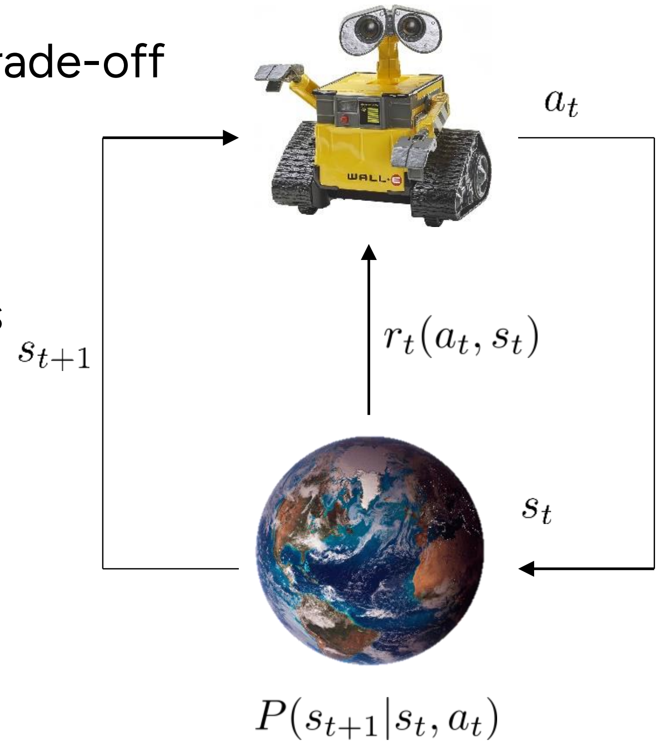


Stats

- (1933) William R. Thompson
- (1950) Mosteller & Bush:
  "two armed bandit" machine
- Robbins, Chernoff, Lai...
- treatments, mices and spaceships

RL

- exploration-exploitation trade-off
- (2015) used in AlphaGo
- advert placement
- recommendation services (Spotify, Netflix...)
- big tech



Main source for introduction: Lattimore, Tor, and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Stochastic bandits with finitely many arms

- How to choose arms based on history to minimize the regret?

for $t = 1, 2, \ldots, T$ **do**
  Choose an arm a_t based on history $(a_1, r_1, \ldots, a_{t-1}, r_{t-1})$;
  Observe noisy reward $r_t = \mu_{a_t} + \eta_t$;
end

Output: regret $R_T = \max_a \sum_{t=1}^{T} (\mu_a - \mu_{a_t})$

- Exploration-exploitation trade-off: choose each arm *lagom* number of times

Stochastic linear bandits

- Arms have feature representations $a \in \mathcal{A} \subset \mathbb{R}^d$
- Obtain rewards $r_t = \langle \theta^\star, a_t \rangle + \eta_t$
- Minimize regret $R_T = \max_{a \in \mathcal{A}} \sum_{t=1}^{T} \langle \theta^\star, a - a_t \rangle$
- First idea:

$$\hat{\theta}_t = (\lambda I + \sum_{i=1}^{t} a_i a_i^\top)^{-1} \sum_{i=1}^{t} a_i r_i$$

$$a_t = \arg\max_{a \in \mathcal{A}} \langle \hat{\theta}_t, a \rangle$$

- Issue: too greedy, no exploration

Digression: linear regression

$$y_i = \langle \theta^\star, x_i \rangle + \eta_t$$

- Data fixed and independent!
- Regularized least squares estimator

$$\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - \langle \theta, x_i \rangle)^2 + \lambda \|\theta\|_2^2$$

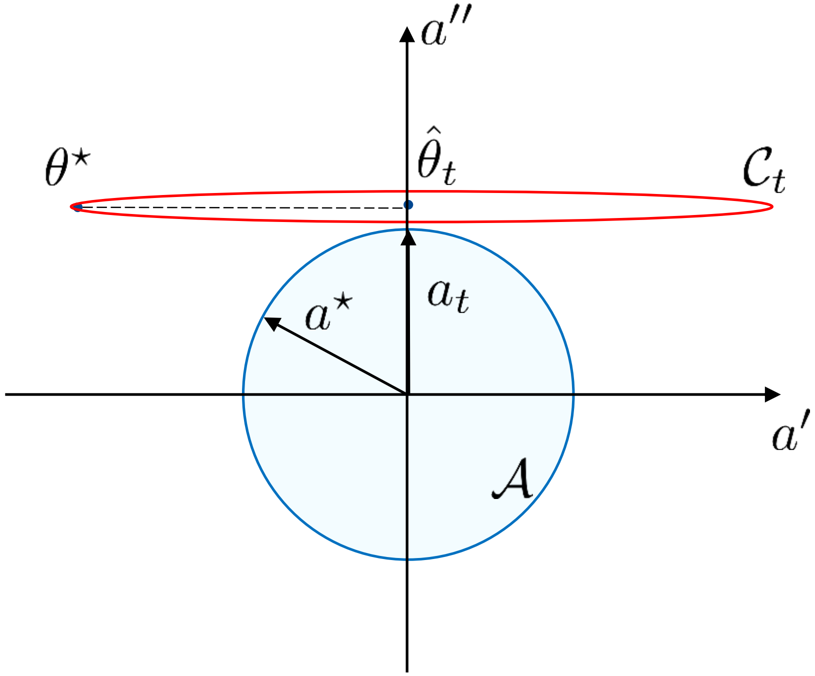
$$\hat{\theta}_n = (\lambda I + \sum_{i=1}^{n} x_i x_i^\top)^{-1} \sum_{i=1}^{n} x_i y_i$$

Optimism is key to success!

- First idea: choose $\hat{\theta}_t = \underbrace{(\lambda I + \sum_{i=1}^{t} a_i a_i^\top)}_{V_t}{}^{-1} \sum_{i=1}^{t} a_i r_i$

$$a_t = \arg\max_{a \in \mathcal{A}} \langle \hat{\theta}_t, a \rangle$$

- Optimism under uncertainty (LinUCB):
  - Confidence ellipsoid $\mathcal{C}_t = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_{t-1}}^2 \leq \beta_t\}$
  - Choose action $a_t = \arg\max_{a \in \mathcal{A}} \underbrace{\max_{\theta \in \mathcal{C}_t} \langle \theta, a \rangle}_{\text{upper confidence bound}}$
  - With high probability $\theta^\star \in \mathcal{C}_t, \forall t$ and volume of $\mathcal{C}_t$ shrinks
  - Optimal in many ways
- Takeaway: optimism encourages exploration



Low-rank stochastic bandits

- High dimensional setting: number of samples $\ll$ number of arms
- Structural assumptions: sparsity, block structures, **low-rank**
- Instead of $\theta^\star \in \mathbb{R}^d$, assume $\Theta^\star \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(\Theta^\star) = r \ll d$
- But why low-rank bandits?

for $t = 1, 2, \ldots, T$ **do**
  Choose an arm pair (i_t, j_t);
  Observe noisy reward $r_t = \Theta^\star_{i_t, j_t} + \eta_t$;
end
Output: regret $R_T = \max_{(i,j)} \sum_{t=1}^{T} (\Theta^\star_{i,j} - r_t)$

pick row →

$$\begin{bmatrix} \Theta^\star_{1,1} & \cdots & \Theta^\star_{1,d} \\ & \ddots & \\ \Theta^\star_{d,1} & \cdots & \Theta^\star_{d,d} \end{bmatrix}$$

↑ pick column

- Number of arms: d^2, number of samples $T \ll d^2$
- Trace regression $Y_t = \mathrm{Tr}(\Theta^\star X_t^\top) + \eta_t$ with $X_t = e_{i_t} e_{j_t}^\top$

“Många bäckar små gör en stor å”, reverse?

$$\Theta^\star = U \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix} U^\top = U\Sigma U^\top$$

- Recover a low-rank matrix? Use SVD!
- Empirical estimate $\widetilde{\Theta}$ after projection: $\widehat{\Theta} = \widehat{U}\widehat{\Sigma}\widehat{U}^\top$
- Condition number $\kappa = \dfrac{\sigma_{\max}}{\sigma_{\min}}$ and incoherence: $\mu = \sqrt{\frac{d}{r}}\|U\|_{2\to\infty} = \sqrt{\frac{d}{r}}\max_{i\in[d]}\|U_{i,:}\|_2$
- Is global (Frobenius norm) recovery sufficient?

Entry-wise recovery guarantees

- For $T = \widetilde{\Omega}(d \cdot \mathrm{poly}(\mu, \kappa, r))$ w.h.p:

$$\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty} = \widetilde{O}\left(\frac{1}{\sqrt{T}}\mathrm{poly}(\mu, \kappa, r)\right)$$

- Even in independent setting difficult to analyse $\|(\widetilde{\Theta} - \Theta^\star)\widehat{U}\|_{2\to\infty}$ because $\widetilde{\Theta}, \widehat{U}$ are dependent
- Our case: approximate entries by independent Compound Poisson random variables to remove dependences

(mental) Fika break

- Questions?
- Wake-up time!
- Recap for forgetful:

Linear Bandits

- No structure
- Need to try all arms
- Known optimal algorithms
- Regret $O(\sqrt{\dim(\mathcal{A})T})$

low-dimensional singular subspace guarantees

$$\|U - \widehat{U}(\widehat{U}^\top U)\|$$

Low-rank Bandits

- Low-rank rewards matrix $\Theta^\star$
- Sample deficient regime:
  - Cannot try all arms
  - Shared information
- Unknown optimal algorithms
- Goal: regret $O(\sqrt{dT})$ while $\dim(\mathcal{A}) = d^2$

Cautious approach:
reduction to *almost* low-dimensional linear bandits*

- Projection on singular vectors subspace: $\mathrm{proj}(U) = U(U^\top U)^{-1}U^\top = UU^\top$

$$\Theta^\star_{i,j} = e_i^\top \widehat{U}(\widehat{U}^\top \Theta^\star \widehat{U})\widehat{U}^\top e_j + e_i^\top \widehat{U}(\widehat{U}^\top \Theta^\star \widehat{U}_\perp)\widehat{U}_\perp^\top e_j + e_i^\top \widehat{U}_\perp(\widehat{U}_\perp^\top \Theta^\star \widehat{U})\widehat{U}^\top e_j + e_i^\top \widehat{U}_\perp(\widehat{U}_\perp^\top \Theta^\star \widehat{U}_\perp)\widehat{U}_\perp^\top e_j.$$

- Or equivalently: $\Theta^\star_{i,j} = \langle \theta, \phi_{i,j} \rangle$ with $\phi_{i,j} = \begin{bmatrix} \mathrm{vec}(\widehat{U}^\top e_i e_j^\top \widehat{U}) \\ \mathrm{vec}(\widehat{U}^\top e_i e_j^\top \widehat{U}_\perp) \\ \mathrm{vec}(\widehat{U}_\perp^\top e_i e_j^\top \widehat{U}) \\ \mathrm{vec}(\widehat{U}_\perp^\top e_i e_j^\top \widehat{U}_\perp) \end{bmatrix}, \quad \theta = \begin{bmatrix} \mathrm{vec}(\widehat{U}^\top \Theta^\star \widehat{U}) \\ \mathrm{vec}(\widehat{U}^\top \Theta^\star \widehat{U}_\perp) \\ \mathrm{vec}(\widehat{U}_\perp^\top \Theta^\star \widehat{U}) \\ \mathrm{vec}(\widehat{U}_\perp^\top \Theta^\star \widehat{U}_\perp) \end{bmatrix}.$

- Subspace recovery implies: $\|\widehat{U}_\perp^\top \Theta^\star \widehat{U}_\perp\|_F \leq \|U - \widehat{U}(\widehat{U}^\top U)\|_F^2 \|\Theta^\star\|_2$
- Algorithm idea:
  - Find the structure: explore uniformly at random to obtain $\widehat{U}$
  - Exploit the structure: use (Sup)LinUCB with

$$\hat{\theta}_\tau = (\Lambda + \textstyle\sum_{t=1}^\tau \phi_{i_t,j_t}\phi_{i_t,j_t}^\top)^{-1} \sum_{t=1}^\tau \phi_{i_t,j_t} r_t \qquad \Lambda = \mathrm{diag}(\underbrace{\lambda, \lambda, \dots, \lambda_\perp, \lambda_\perp \dots}_{\text{regularize less first r(2d-r) entries}})$$

  - Regret satisfies $R_T = \widetilde{O}(\mathrm{poly}(\mu, \kappa, r) d\sqrt{T})$

*Jun, Kwang-Sung, et al. "Bilinear bandits with low-rank structure." *International Conference on Machine Learning*. PMLR, 2019.

Indifferent approach:
reduction to misspecified linear bandits

- Can we leverage instead our tight entry-wise guarantees?
- Define $\varepsilon_{i,j} = e_i^\top \widehat{U}_\perp (\widehat{U}_\perp^\top \Theta^\star \widehat{U}_\perp) \widehat{U}_\perp^\top e_j$
- Reduction to misspecified linear bandits:

$$\Theta^\star_{i,j} = \langle \theta, \phi_{i,j} \rangle + \varepsilon_{i,j} \text{ with } \phi_{i,j} = \begin{bmatrix} \mathrm{vec}(\widehat{U}^\top e_i e_j^\top \widehat{U}) \\ \mathrm{vec}(\widehat{U}^\top e_i e_j^\top \widehat{U}_\perp) \\ \mathrm{vec}(\widehat{U}_\perp^\top e_i e_j^\top \widehat{U}) \end{bmatrix}, \quad \theta = \begin{bmatrix} \mathrm{vec}(\widehat{U}^\top \Theta^\star \widehat{U}) \\ \mathrm{vec}(\widehat{U}^\top \Theta^\star \widehat{U}_\perp) \\ \mathrm{vec}(\widehat{U}_\perp^\top \Theta^\star \widehat{U}) \end{bmatrix}.$$

- Subspace recovery implies: $\max_{i,j} |\varepsilon_{i,j}| \le \|U - \widehat{U}(\widehat{U}^\top U)\|^2_{2\to\infty} \|\Sigma^\star\|_2$
- Algorithm idea:
  - Find the structure: explore uniformly at random to obtain $\widehat{U}$
  - Exploit the structure: use (Sup)LinUCB with

$$\hat{\theta}_\tau = (\lambda I + \textstyle\sum_{t=1}^\tau \phi_{i_t,j_t} \phi_{i_t,j_t}^\top)^{-1} \sum_{t=1}^\tau \phi_{i_t,j_t} r_t$$

  - Regret satisfies $R_T = \widetilde{O}(\mathrm{poly}(\mu, \kappa, r) d^{3/4} \sqrt{T})$
- First algorithm achieving not-trivial tightness in dimension!

So, when is SVD all you need?

- For regret minimization (/ best arm identification / policy evaluation) in low-rank bandits with
  - incoherent subspaces
  - known rank
  - contexts (if nearly uniformly distributed)
  - not too large dimension
- Lowest reported upper bound on regret, but probably still not optimal

Discovering structure adaptively is hard!

- Proposed algorithms explore uniformly at random
- Can we do adaptive exploration + low-rank structure recovery?

Main references

A. Stojanovic, Stefan, Yassir Jedra, and Alexandre Proutiere. "Spectral entry-wise matrix estimation for low-rank reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).

B. Jedra, Yassir*, William Reveillard*, Stefan Stojanovic*, Alexandre Proutiere. "Low-Rank Bandits via Tight Two-to-Infinity Singular Subspace Recovery." *arXiv preprint arXiv:2402.15739* (2024).