Many-Armed Bandits with High-Dimensional Contexts under a Low-Rank Structure

Nima Hamidi
Stanford University

Mohsen Bayati
Stanford University

Kapil Gupta
Airbnb

Formal setting

1. Each arm i corresponds to an **unknown** vector $B_i \in \mathbb{R}^d$.
2. At time t, a **context vector** $X_t \in \mathbb{R}^d$ is revealed to the policy.
3. The policy π selects action $a_t \in [k]$.
4. The **reward** is given by $y_t = \langle B_{a_t}, X_t \rangle + \varepsilon_t$.

Formal setting

1. Each arm i corresponds to an **unknown** vector $B_i \in \mathbb{R}^d$.
2. At time t, a **context vector** $X_t \in \mathbb{R}^d$ is revealed to the policy.
3. The policy π selects action $a_t \in [k]$.
4. The **reward** is given by $y_t = \langle B_{a_t}, X_t \rangle + \varepsilon_t$.

We further assume:

1. X_t's are i.i.d.
2. ε_t's are independent mean-zero sub-Gaussian.
3. $(X_t) \perp\!\!\!\perp (\varepsilon_t)$
4. B is of rank r.

Cumulative regret

Definition

We define the **cumulative regret** of a given policy as follows:

$$R_T = \sum_{t=1}^{T} \left[\max_{1 \le i \le k} \langle B_{t,i}, X_t \rangle - \langle B_{t,a_t}, X_t \rangle \right].$$

Policies with smaller (expected) regrets are desired.

Theoretical guarantees

- OLS-Bandit: $O(d^2 k^3 \log(T))$ [Alexander Goldenshluger]
- Lasso-Bandit: $O(s^2 k^3 \log(T)^2)$ [Hamsa Bastani]
- REAL-Bandit: $O(r^2 (k + d) \log(T)^2)$

REAL-Bandit

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

REAL-Bandit

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

REAL-Bandit

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2	3	1	1	3	1	2	3	1	2	3	2	3	1	2	2	2	3	?

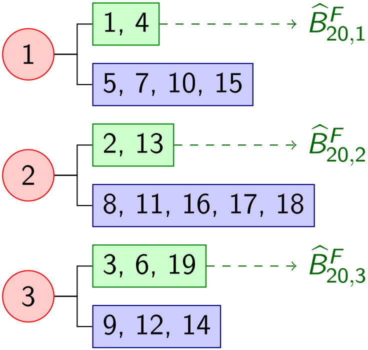
REAL-Bandit

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2	3	1	1	3	1	2	3	1	2	3	2	3	1	2	2	2	3	?

- 1
  - 1, 4
  - 5, 7, 10, 15
- 2
  - 2, 13
  - 8, 11, 16, 17, 18
- 3
  - 3, 6, 19
  - 9, 12, 14

REAL-Bandit

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2	3	1	1	3	1	2	3	1	2	3	2	3	1	2	2	2	3	?



REAL-Bandit

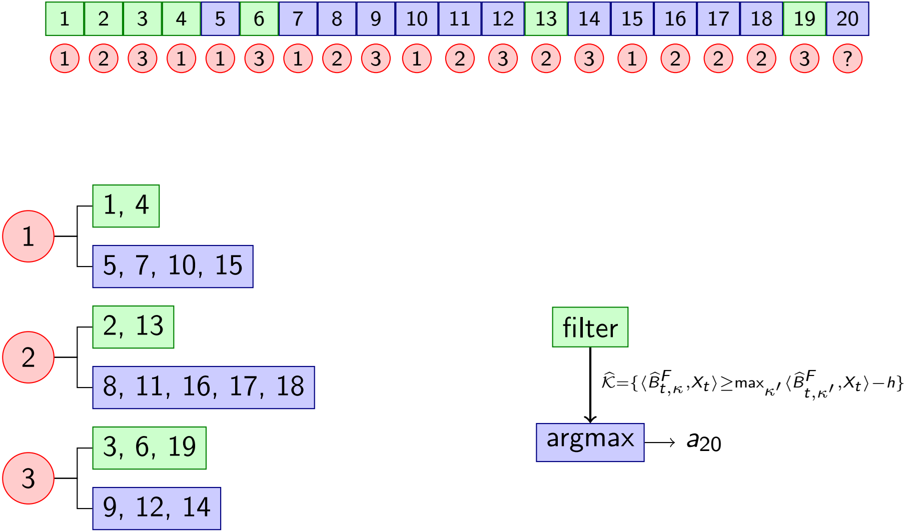
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2	3	1	1	3	1	2	3	1	2	3	2	3	1	2	2	2	3	?

1
- 1, 4
- 5, 7, 10, 15 $\dashrightarrow \widehat{B}^{A}_{20,1}$

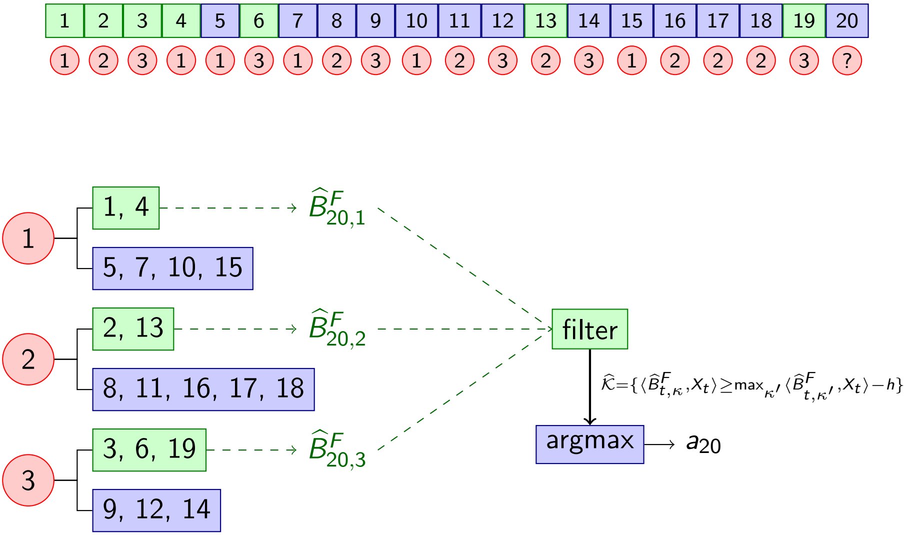
2
- 2, 13
- 8, 11, 16, 17, 18 $\dashrightarrow \widehat{B}^{A}_{20,2}$

3
- 3, 6, 19
- 9, 12, 14 $\dashrightarrow \widehat{B}^{A}_{20,3}$

REAL-Bandit

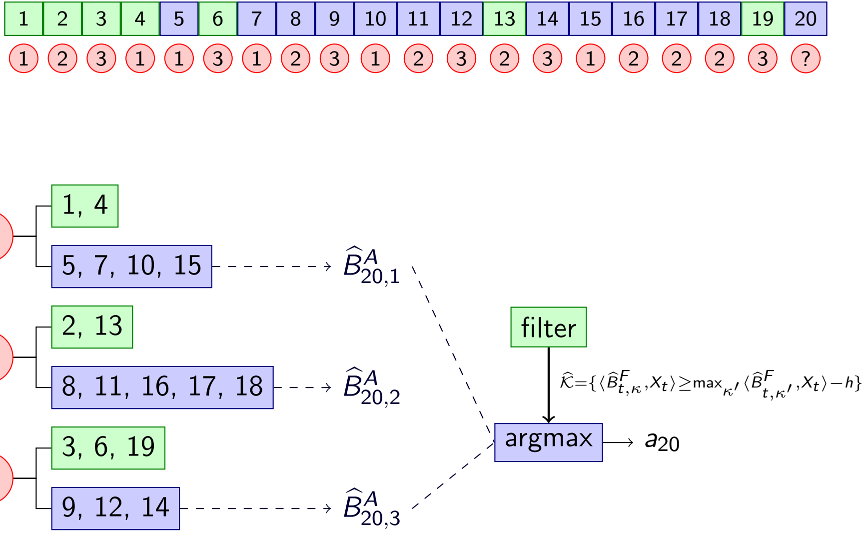

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1 2 3 1 1 3 1 2 3 1 2 3 2 3 1 2 2 2 3 ?
1
1, 4
5, 7, 10, 15
2
2, 13
8, 11, 16, 17, 18
3
3, 6, 19
9, 12, 14
filter
$\widehat{\mathcal{K}}=\{\langle \widehat{B}^F_{t,\kappa}, x_t\rangle \geq \max_{\kappa'} \langle \widehat{B}^F_{t,\kappa'}, x_t\rangle - h\}$
argmax
a_{20}

REAL-Bandit



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1 2 3 1 1 3 1 2 3 1 2 3 2 3 1 2 2 2 3 ?
1
1, 4
5, 7, 10, 15
2
2, 13
8, 11, 16, 17, 18
3
3, 6, 19
9, 12, 14
$\widehat{B}^F_{20,1}$
$\widehat{B}^F_{20,2}$
$\widehat{B}^F_{20,3}$
filter
$\widehat{\mathcal{K}}=\{\langle \widehat{B}^F_{t,\kappa}, X_t\rangle \geq \max_{\kappa'} \langle \widehat{B}^F_{t,\kappa'}, X_t\rangle - h\}$
argmax
a_{20}

REAL-Bandit



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1 2 3 1 1 3 1 2 3 1 2 3 2 3 1 2 2 2 3 ?
1
1, 4
5, 7, 10, 15
$\widehat{B}^A_{20,1}$
2
2, 13
8, 11, 16, 17, 18
$\widehat{B}^A_{20,2}$
3
3, 6, 19
9, 12, 14
$\widehat{B}^A_{20,3}$
filter
$\widehat{\mathcal{K}}=\{\langle \widehat{B}^F_{t,\kappa}, x_t\rangle \geq \max_{\kappa'} \langle \widehat{B}^F_{t,\kappa'}, x_t\rangle - h\}$
argmax
a_{20}

REAL-Estimator

- Use any low-rank estimator, such as

$$\bar{B} := \arg\min_B \frac{\|Y - \mathfrak{X}(B)\|_2^2}{n} + \lambda \|B\|_*$$

such that the following holds with high probability

$$\left\|\bar{B} - B\right\|_F^2 \leq C\sigma^2 \frac{dr}{n}.$$

- This bound leads to extra $\sqrt{k}$ in the regret bound.

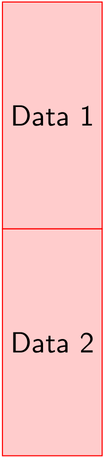
REAL-Estimator

- Let $\bar{B}$ be defined as in the previous slide.
- Run the following "*row-enhancement*" procedure.
- This procedure eliminates extra $\sqrt{k}$ factor in the regret.

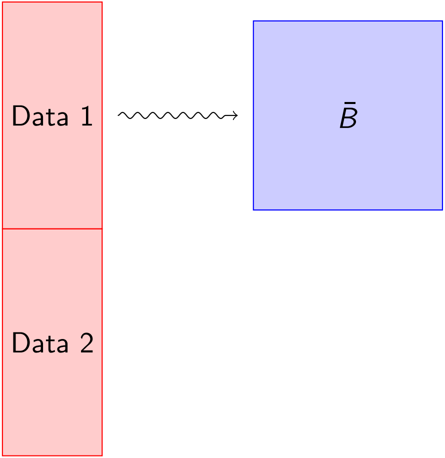
Input: matrix $\bar{B}_{k\times d}$, observations $(X_1, Y_1), \cdots, (X_n, Y_n)$

1: Compute SVD $\bar{B} = UDV^T$.
2: Let V_r^T be the matrix containing r top rows of V^T.
3: Let $\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^d} \sum_{i=1}^n (Y_i - X_i V_r \beta)^2$.
4: Then, output $\widehat{B}_\kappa = (V_r\hat{\beta})^T$.

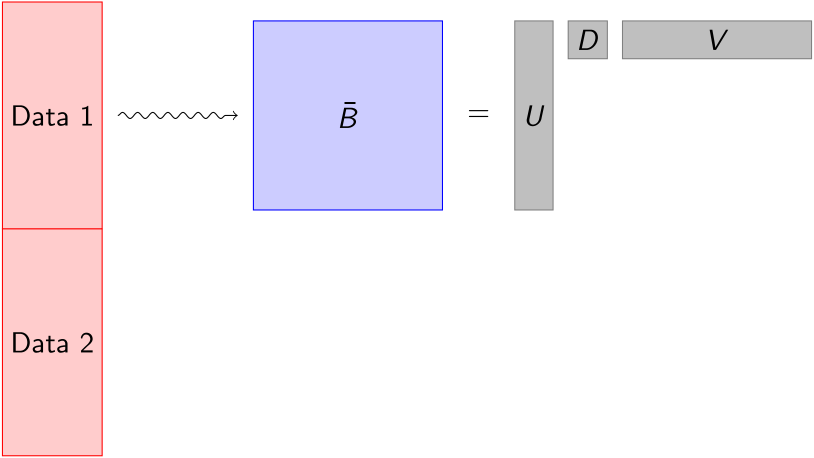
REAL-Estimator



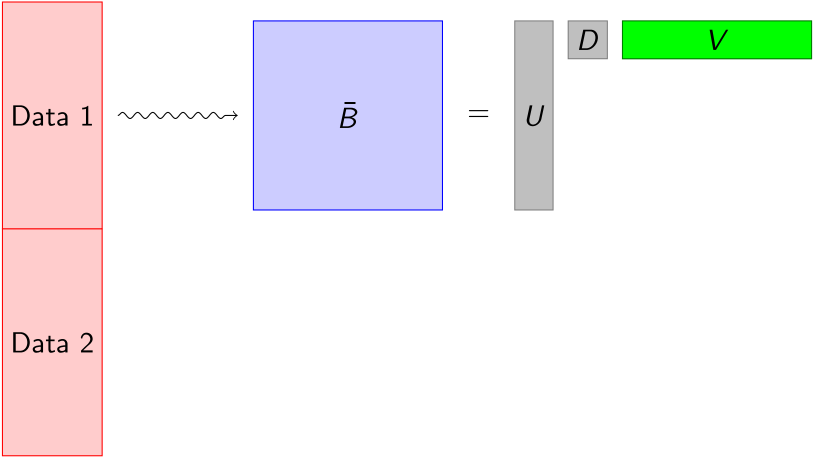
REAL-Estimator



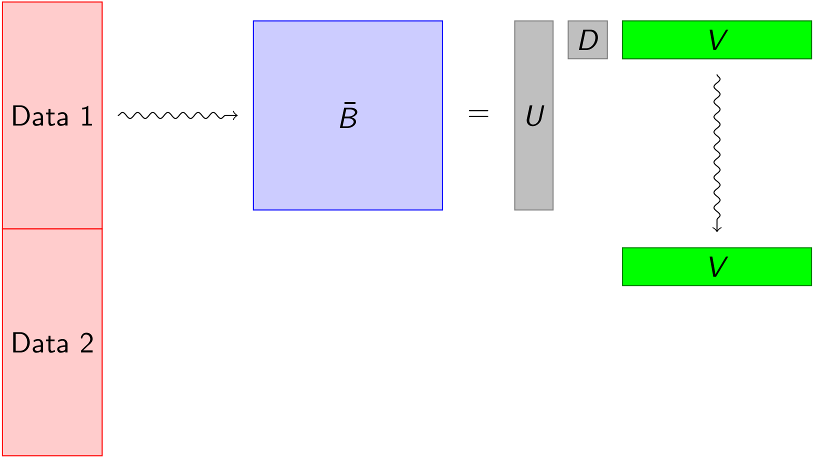
REAL-Estimator



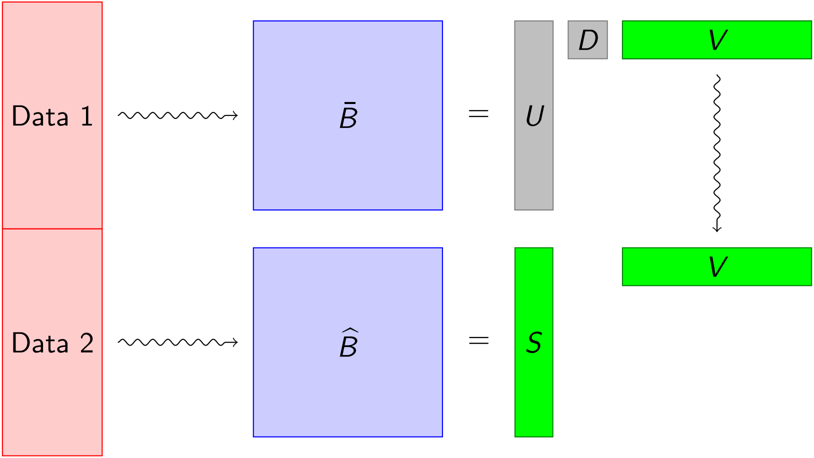
REAL-Estimator



REAL-Estimator

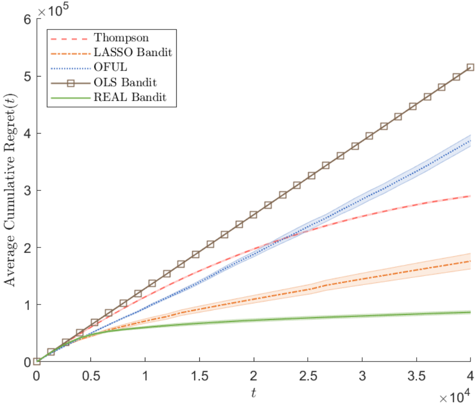


REAL-Estimator



Simulations

- B: 200×201 of rank 3,
- SD of noise (σ): 1,
- Context vectors (X_t): vectors of length 201 with i.i.d. standard normal entries.



References

Alexander Goldenshluger and Assaf Zeevi
A linear response bandit problem
Stochastic Systems 3.1 (2013): 230-261.

Hamsa Bastani and Mohsen Bayati
Online decision-making with high-dimensional covariates
(2015).

Thank you