Best-Arm Identification in Linear Bandits

Neural Information Processing Systems Foundation

Marta Soare, Alessandro Lazaric, Remi Munos

Inria — informatics mathematics

Motivation

- Product recommendation (e.g., movies)
- Adaptive hypothesis testing under linear assumption
- Optimization of a stochastic linear function

Setting

The linear stochastic bandit model

- Set of arms $\mathcal{X} \subseteq \mathbb{R}^d$, $|\mathcal{X}| = K$, $||x||_2 \le L, \forall x \in \mathcal{X}$.
- Linear reward model

$$r(x) = x^\top \theta^* + \varepsilon$$

with $\theta^* \in \mathbb{R}^d$ **unknown** parameter and noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
- The (unique) best arm in $\mathcal{X}$:

$$x^* = \arg\max_{x \in \mathcal{X}} x^\top \theta^*$$

The best-arm identification problem ($(0,\delta)$-PAC setting)

- $\hat{x}(n)$ – recommended best arm after n steps.
- Given a *fixed confidence* $\boldsymbol{\delta}$, design an *allocation strategy* and a *stopping criterion* such that:

$\boxed{\mathbb{P}(\hat{\mathbf{x}}(\mathbf{n}) = \mathbf{x}^*) \ge 1 - \delta \text{ and } \mathbf{n} \text{ as small as possible.}}$

Tools

Ordinary Least-Squares estimate

- Sequence of arms $\mathbf{x}_n = (x_1, \ldots, x_n) \in \mathcal{X}^n$
- Sequence of rewards $(r_1, \ldots, r_n)$
- OLS estimate, $A_{\mathbf{x}_n} = \sum_{t=1}^n x_t x_t^\top$, $b_{\mathbf{x}_n} = \sum_{t=1}^n x_t r_t$

$$\hat{\theta}_n = A_{\mathbf{x}_n}^{-1} b_{\mathbf{x}_n}$$

Prediction errors

- Fixed sequence and OLS estimate (w.p. $1-\delta$)

$$|x^\top \theta^* - x^\top \hat{\theta}_n| \le c||x||_{A_{\mathbf{x}_n}^{-1}} \sqrt{\log_n(K/\delta)}$$

- Adaptive sequence (Thm.2 in [1]) for η-regularized OLS (w.p. $1-\delta$)

$$|x^\top(\theta^* - \hat{\theta}_n^\eta)| \le ||x||_{A_{\mathbf{x}_n}^{\eta,-1}} \left(\sigma \sqrt{d \log\left(\frac{1 + nL^2/\eta}{\delta}\right)} + \eta^{1/2} ||\theta^*|| \right)$$

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances Neural Information Processing Systems 24*, 2011.

[2] E. Even-Dar, S. Mannor, and Y.Mansour. PAC bounds for multi-armed bandit and markov decision processed. In *Computational Learning Theory*, 2002.

[3] V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances Neural Information Processing Systems 25*, 2012.

Acknowledgements

Université de Lille — ComPLACS — Microsoft Research — DeepMind

Gaps and Directions

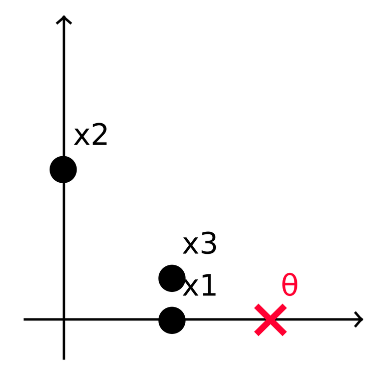
Value gaps

- For any pair $x, x' \in \mathcal{X}$, $\Delta(x, x') = (x - x')^\top \theta^*$
- Smallest gap $\Delta_{\min} = \min_{x \in \mathcal{X}} \Delta(x^*, x)$

The smaller the gaps the more difficult the problem.

Example

- x_1 and x_3 are very close, while x_2 is clearly suboptimal
- Only direction $y = x_1 - x_3$ (i.e., θ_2^*) must be estimated accurately
- x_2 provides *much* information about direction y (θ_2^*)

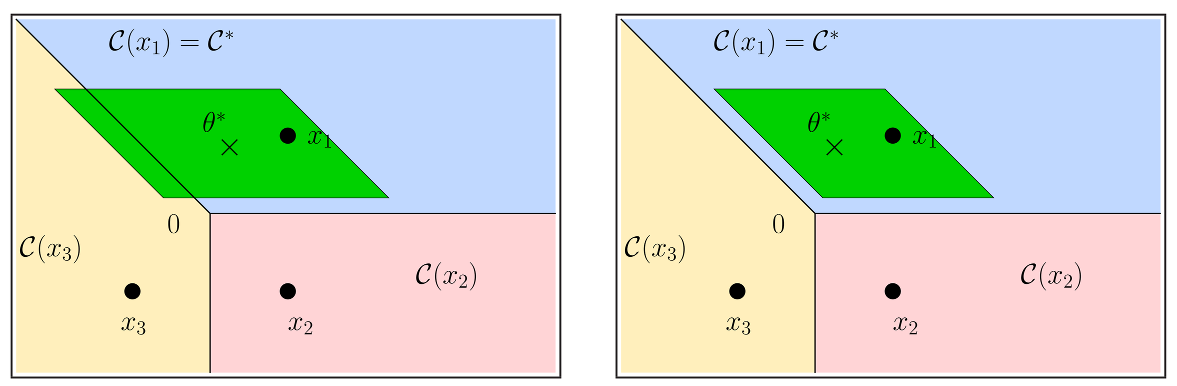


Use one arm to learn about the others (exploit the linear structure)!

Sets of directions

$$\mathcal{Y} = \{y = x - x'\}, \qquad \mathcal{Y}^* = \{y = x^* - x\}$$

Illustration



Optimality cones

$$\mathcal{C}(x) = \cap_{x' \in \mathcal{X}} \{\theta \in \mathbb{R}^d, (x - x')^\top \theta \ge 0\}$$

Confidence set

$$\mathcal{S}(\mathbf{x}_n) = \left\{ \theta, \forall y, y^\top(\theta^* - \theta) \le c||y||_{A_{\mathbf{x}_n}^{-1}} \sqrt{\log(K^2/\delta)} \right\}$$

$\mathcal{XY}$-Oracle

Intuition: select arms so that the confidence set shrinks into one optimality cone as soon as possible.

Stopping rule	Allocation rule
$\exists x \in \mathcal{X} \text{ s.t. } \mathcal{S}(\mathbf{x}_n) \subset \mathcal{C}(x)$	$\mathbf{x}_n^* = \arg\min_{\mathbf{x}_n} \max_{y \in \mathcal{Y}^*} \frac{\lVert y \rVert_{A_{\mathbf{x}_n}^{-1}}}{\Delta(y)}$

Oracle sample complexity	Complexity of Linear Best-Arm Identification
$N^* = c^2 H_{\text{LB}} \log_n(K^2/\delta)$	$H_{\text{LB}} = \min_{\lambda \in \mathcal{D}^k} \max_{y \in \mathcal{Y}^*} \frac{\lVert y \rVert_{\Lambda_\lambda^{-1}}^2}{\Delta^2(y)}$

N^* is the lower-bound on the sample complexity of any fixed allocation strategy

Remarks

$$\max_{y \in \mathcal{Y}^*} \frac{||y||^2}{L\Delta_{\min}^2} \le H_{\text{LB}} \le \frac{4d}{\Delta_{\min}^2} \qquad H_{\text{MAB}} \le H_{\text{LB}} \le 2H_{\text{MAB}}$$

Static allocations

G-Optimal Design: estimate θ^* *uniformly well* over all arms

$$\mathbf{x}_n^G = \arg\min_{\mathbf{x}_n} \max_{x \in \mathcal{X}} ||x||_{A_{\mathbf{x}_n}^{-1}}$$

$\mathcal{XY}$-Design: estimate the value of the gaps *uniformly well* over all the directions in $\mathcal{Y}$

$$\mathbf{x}_n^{\mathcal{XY}} = \arg\min_{\mathbf{x}_n} \max_{y \in \mathcal{Y}} ||y||_{A_{\mathbf{x}_n}^{-1}}$$

Empirical stopping criterion:

$$\boxed{\begin{aligned} &\exists x \in \mathcal{X}, \forall x' \in \mathcal{X}, \forall \theta \in \hat{\mathcal{S}}(\mathbf{x}_n) \\ &\qquad (x - x')^\top(\hat{\theta}_n - \theta) \le \widehat{\Delta}_n(x, x') \end{aligned}}$$

Sample complexity $O(d/\mathbf{\Delta}^2_{\min})$.

From $\mathcal{Y}$ to $\mathcal{Y}^*$

The minimum number of steps needed by the static $\mathcal{XY}$-allocation to discard all *suboptimal* directions.

$$M^* = \min\{n \in \mathbb{N}, \forall x \ne x^*, \forall x' \ne x^*, \\ S^*(\mathbf{x}_n^{\mathcal{XY}}) \cap (\mathcal{C}(x) \cap \mathcal{C}(x')) = \emptyset\}.$$

$\mathcal{XY}$-adaptive Algorithm

Input: $\mathcal{X} \in \mathbb{R}^d$; confidence δ;
Phase length given by an α improvement;
Set $j = 1$; $\hat{\mathcal{X}}_j = \mathcal{X}$; $\hat{\mathcal{Y}}_1 = \mathcal{Y}$; $n = 0$;

Stopping rule

while $|\hat{\mathcal{X}}_j| > 1$ **do**
  Start a new phase: $j = j + 1$, $t = 1$; $A_0 = I$
  while $\rho^j / t \ge \alpha \rho^{j-1}(\mathbf{x}_{n_{j-1}}^{j-1})/n_{j-1}$ **do**
    Allocation rule
    $x_t = \arg\min_{x \in X} \max_{y \in \hat{\mathcal{Y}}_j} y^\top (A + xx^\top)^{-1} y$
    Update $A_t = A_t + x_t x_t^\top$; $t = t + 1$; $n = n + 1$
    $\rho^j = \max_{y \in \hat{\mathcal{Y}}_j} y^\top A_t^{-1} y$
  end while
  $b = \sum_{s=1}^t x_s r_s$; $\hat{\theta}_j = A_t^{-1} b$
  Recompute the set of potential optimal arms:
  $\hat{\mathcal{X}}_j = \{x, \nexists x' : ||x - x'||_{A_t^{-1}} \sqrt{\log_n(K^2/\delta)} \le \hat{\Delta}_j(x', x)\}$
  Recompute the set of directions of interest:
  $\hat{\mathcal{Y}}_j = \{y = (x - x'); x, x' \in \hat{\mathcal{X}}_j\}$
end while

Recommendation rule

Return $\hat{x}(n)$ - the only arm remaining in $\hat{\mathcal{X}}_j$.

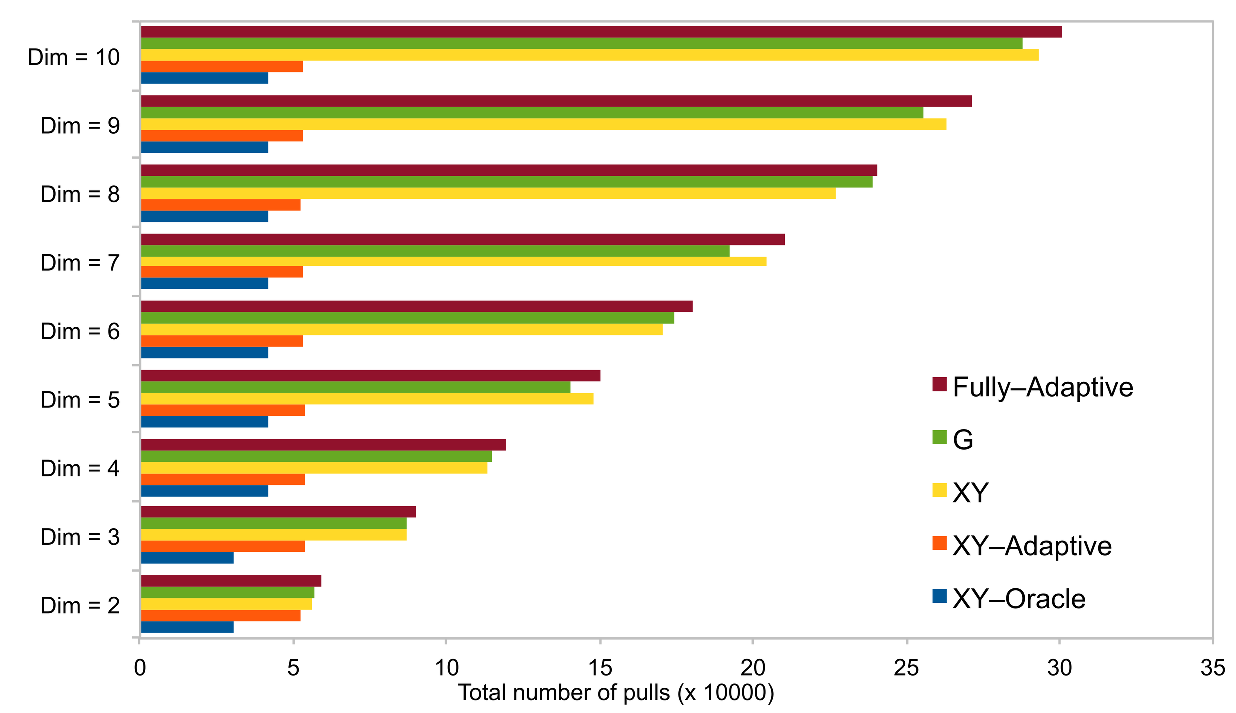
Main result

Theorem 1. *If the $\mathcal{XY}$-adaptive allocation strategy is implemented with a β-approximate method then*

$$\mathbb{P}\left[N \le \frac{(1+\beta)\max\{M^*, \frac{16}{\alpha}N^*\}}{\log(1/\alpha)} \log\left(\frac{c\sqrt{\log_n(K^2/\delta)}}{\Delta_{\min}}\right) \wedge (\hat{x}_N = x^*)\right] \ge 1 - \delta.$$

The bound holds for any $(1+\beta)$-*approximate allocation strategy*: e.g., continuous relaxation, greedy incremental allocation.

Experiments – Sample Complexity and arm pulls



Arm	$\mathcal{XY}$-oracle	$\mathcal{XY}$-adapt.	$\mathcal{XY}$	G	Fully-adapt.
x_1	207	263	29523	28014	740
x_2	41440	52713	29524	28015	149220
x_3	2	3	29524	28015	1
x_4	2	5	29524	28015	1
x_5	1	2	29524	28015	1
x_6	0	2	1	1	1
Total	**41652**	**52988**	147620	140075	149964

Setting:

- Fixed confidence $\delta = 0.05$.
- Set of arms: $\mathcal{X} \in \mathbb{R}^d, |\mathcal{X}| = d+1$ and d $= 2, \ldots, 10$.
- Canonical basis $(x_1, \ldots, x_d)$ and additional arm x_{d+1} very close to x_1.
- $\theta^* = [2 \;\; 0 \;\; 0 \;\; \ldots \;\; 0]^\top \to \Delta_{\min} = (x_1 - x_{d+1})^\top \theta^*$ much smaller than the other gaps.
- Identifying the best arm $\to$ reducing uncertainty in the direction $\tilde{y} = (x_1 - x_{d+1})$.
- x_2 is almost aligned with $\tilde{y} \to$ **the most informative arm**.

The sample complexity **grows linearly with the dimension**:

- Fully-adaptive – despite pulling only the informative arms, the additional d term in the bound prevents a good performance.
- G and $\mathcal{XY}$ – always consider the complete set $\mathcal{Y}$.

The sample complexity remains **constant**:

- **$\mathcal{XY}$-Adaptive** and **$\mathcal{XY}$-Oracle** – exclusively pull the two most informative arms, independently of the number of dimensions.