Proposed Method: LinGapE

LinGapE — Linear Gap-based Exploration

Stopping condition:
- Based on the confidence bound in (2).
- Valid for adaptive strategies as well.

Arm selection strategy:
- At each round $t$, repeat the following:
  - Nominate two arms $i, j$:
    - Pull the arm $i_0$ that discriminates $i, j$, the most.

The algorithm for selecting $i_t, j_t$

Arm $i_t$ is the estimated best arm, and arm $j_t$ is the arm that is the most likely to surpass $i_t$.

Algorithm 2: Select-direction

Procedure Select-direction($i_t$):
- $i_t$ = arg max$_i$ $\hat{A}_i - \hat{A}_j + \frac{1}{2} \lambda \left( \frac{1}{n_t} \right)$
- $j_t = i_t$, $\hat{A}_i = \hat{A}_j$
- $\hat{A}_j(t) = (x_{j_t} - x_{j_{t-1}})^T \hat{d}_{j_t}$
- $\hat{A}_i = (x_{i_t} - x_{i_{t-1}})^T \hat{d}_{i_t} + \frac{1}{\sqrt{t}} \sum_{t'=1}^{t-1} x_{i_{t'}}^T \hat{d}_{i_t}$

The algorithm for selecting $\alpha_{t+1}$

- Compute the optimal arm selection ratio $(\hat{p}(i, j))_{i,j}$ for discriminating arms $i$ and $j$ by

\[
(\hat{p}(i, j))_{i,j} = \arg \min_{i,j} \left( x_{i_{t+1}} - x_{j_{t+1}} \right), \text{ s.t. } \sum_{i=1}^{N} \alpha_{i} \geq 0, \text{ and } \sum_{i=1}^{N} \alpha_{i} x_{i}^{2} \leq \delta.
\]

which can be solved by the linear program.
- $\alpha_{i+1} = \arg \min_{\alpha \in \mathbb{R}^N} T(t+1)/p_{i,t+1}(i, j)$.

Theoretical Analysis

Theorem 1: Sample Complexity of LinGapE

If $\lambda \leq \frac{\sqrt{n_t}}{2 \log \left( \frac{n_t}{\delta} \right)}$, then the number of samples $\tau$ of LinGapE satisfies

\[
P[\tau \leq 8R^2 \log \frac{K^2}{\lambda} + C(H, \delta)] \geq 1 - \delta,
\]

where $C(H, \delta) = O \left( \frac{1}{\delta} \log \left( \frac{2eK}{\delta} \right) \right)$ and $p(i, j)$ is the optimal value of (4).

Experiments

Synthetic setting used in [3]:
- The number of arms is $K = d + 1$, where features are $x_1 = e_1, x_2 = e_2, \ldots, x_d = e_d, x_{d+1} = (\cos(0.01), \sin(0.01), 0, 0, \ldots, 0)^T$.
- Set $\theta = (2, 0, \ldots, 0)^T$.
- $x_{i}^T \theta = 2 \times x_{j}^T \theta = 2(0.01) = 0.0038$.
- Arm 2 can discriminate arms 1 and $d + 1$.
- LinGapE mostly select arm 2.
- $\hat{A}_2(t)$ makes (2) tight.
- LinGapE stops faster than $\lambda^2$-oracle.

Setting based on Yahoo! Webscope Dataset R6A [4]:
- Consists of pairs of user-attribute feature $x$ and target $y$ (if $y = 1$ if seen, and $y = 0$ otherwise).
- Relatively high-dimensional data (36-dimensional).
- Estimate $\theta$ by linear ridge regression in $x^T \theta$.
- Run simulations based on the estimated $\theta$.
- 5 times less observations compared to $\lambda^2$-static.
- Less dependent on $K$ compared to $\lambda^2$-static.
- $\lambda$-liveness of (2) does not harm empirical performances.

Our Contributions
- Use a stopping condition based on (2), which allows employing adaptive strategies.
- Provide that $O(\sqrt{n})$ looseness in (2) does not appear in the main term of the sample complexity.

Abstract

A fully adaptive algorithm for pure exploration in linear bandits

Liyuan Xu$^{2,3}$, Junya Honda$^{1,3}$, Masashi Sugiyama$^{2,3}$

1The University of Tokyo 2RIKEN

Linear Bandits
- The set of arms $|\mathcal{K}| = 1, 2, \ldots, K$ and the features of arms $x_1, x_2, \ldots, x_K \in \mathbb{R}^d$.
- At each round $t$, an agent pulls one arm $a_t \in \mathcal{K}$ and observes reward $r_t$.
- Rewards $r_t$ is determined as $r_t = a_t^T \theta + \epsilon$.
- $\epsilon$: i.i.d. Gaussian noise.
- $\theta$: unknown parameter with $\ell_2$-norm at most $\Sigma$.
- The best arm $a^{*} = \arg \max_a a^T \theta$.

Problem Settings

Recent works
- Proposed a new adaptive algorithm for best arm identification in linear bandits, LinGapE, which matches the sample complexity of an oracle algorithm up to a constant in some limit.
- Showed superiority of LinGapE through experiments based on synthetic and realistic settings.

Confidence Bounds

There are two types of the confidence bounds on $\theta$ for sequence of arm selection $x_1, x_2, \ldots, x_n$ and $a_n = \sum_{i=1}^{n} x_i x_i^T b_i = \sum_{i=1}^{n} x_i x_i^T$.

Confidence Bound for Static Strategies [3]

For any fixed sequence $x_{n, \cdot}$, if noise variance $\epsilon$ is bounded $\epsilon \in [-R, R]$, which is $R$-sub-Gaussian

\[
|\epsilon^T \theta - x_i^T \theta| \leq 2R|x_i| \sqrt{2 \log \left( \frac{n}{\delta} \right)} \epsilon
\]

for all $i \in \mathcal{K}$ and all $x_i \in \mathcal{X}_i$, with probability at least $1 - \delta$ for $\|x_i\| = \sqrt{\lambda K}$.

Confidence Bound for Adaptive Strategies [1]

For any adaptive sequence $x_{n, \cdot}$, and $a_n^* = \lambda + \Delta$, for $\lambda > 0$,

\[
|\epsilon^T \theta - x_i^T \theta| \leq R \sqrt{\frac{2K}{n} \log \left( \frac{2K}{\delta} \right)} + \frac{K}{n} \lambda + 2 \Delta.
\]

for all $i \in \mathcal{K}$ and all $x_i \in \mathcal{X}_i$, with probability at least $1 - \delta$.

Work by Soare et al. [3]

- Constructs a stopping condition based on (1) to avoid $O(\sqrt{n})$ looseness of (2).
- Proposes static and semi-adaptive arm selection strategies which make (1) valid.
- Derives the lower bound of sample complexity for static strategies.

Arm selection strategies:
- $\lambda^2$-static: Fix all arm selection before observing any samples.
- $\lambda^2$-adaptive: Arm selection strategy based on the literature of transductive experimental design.
- Cannot change arm adaptively based on rewards.
- $\lambda^2$-adaptive: Semi-adaptive algorithm that adaptively changes static arm allocations.
- Divide rounds into multiple phases, employ different arm allocations in different phases.
- Must discard all samples collected in previous phases for the validity of (1).

Lower bound of static strategies:
- The lower bound is $O(H^{\text{oracle}} \log(1/\delta))$, where $H^{\text{oracle}}$ is defined as

\[
H^{\text{oracle}} = \min_{\theta} \frac{1}{H} \sum_{t=1}^{H} \max_{i,j} \Theta^T \theta \|\epsilon^T \theta - x_{i,j}^T \theta\|_{\mathcal{X}_{i,j}}
\]

for

\[
\Delta = \Lambda x_{1}^T \theta - x_{2}^T \theta.
\]

Lower bound is derived by considering an oracle algorithm, $\lambda^2$-oracle.

$\lambda^2$-oracle computes the optimal arm selection ratio using true $\theta$, which is unknown in reality.

Prior Methods

Use a stopping condition based on (2), which allows employing adaptive strategies.
- Prove that $O(\sqrt{n})$ looseness in (2) does not appear in the main term of the sample complexity.