Doubly-Robust Lasso Bandit

Gi-Soo Kim, Myunghee Cho Paik

Seoul National University

What is Multi-Armed Bandit ?

Consider a **sequential decision problem.**

Suppose..

- A learner is sequentially faced with N actions.
- At each time, the learner can choose only one action.
- The chosen action yields a reward.
- The learner repeats the process and accumulates the rewards.

The **goal** of the learner is to **maximize the sum of rewards**.

What is Multi-Armed Bandit ?

Multi-Armed Bandits (MAB) [Robbins, 1952, Lai and Robbins, 1985] frame the sequential decision problem.

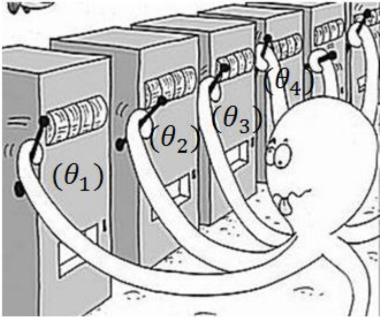


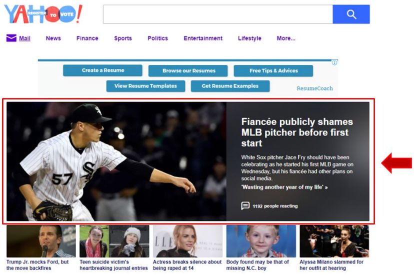
image source: Microsoft Research

- Arms=Actions (# of arms: N)
- At time t, the i-th action yields a random reward $r_i(t)$, such that

$$\mathbb{E}(r_i(t)) = \theta_i(t), \quad i = 1, \cdots, N,$$

  where $\theta_i(t)$'s are unknown.
- At time t, the learner chooses one action $a(t)$, and observes the reward $r_{a(t)}(t)$.
- Goal is to maximize $\sum_{t=1}^{T} \theta_{a(t)}(t)$.

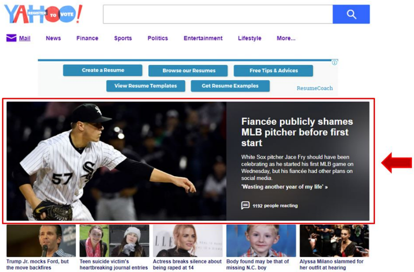
Application 1: News article recommendation



Yahoo! front page snapshot

1. At each user visit, the web system selects one article from a large pool of articles.
2. The system displays it on the Featured tab.
3. The user clicks the article if he/she is interested in the contents.
4. Based on user click feedback, the system updates its article selection strategy.

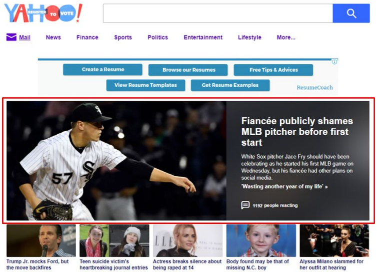
Application 1: News article recommendation



Yahoo! front page snapshot

1. At each user visit, the web system selects one article from a large pool of articles. Arms=Articles
2. The system displays it on the Featured tab.
3. The user clicks the article if he/she is interested in the contents.
4. Based on user click feedback, the system updates its article selection strategy.

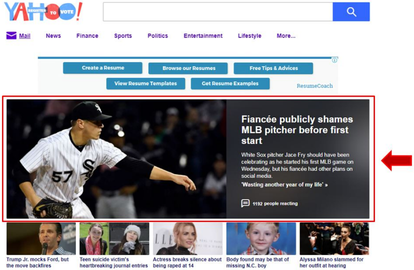
Application 1: News article recommendation



Yahoo! front page snapshot

1. At each user visit, the web system selects one article from a large pool of articles. Arms=Articles
2. The system displays it on the Featured tab. $a(t)$: index of chosen article
3. The user clicks the article if he/she is interested in the contents.
4. Based on user click feedback, the system updates its article selection strategy.

Application 1: News article recommendation



Yahoo! front page snapshot

1. At each user visit, the web system selects one article from a large pool of articles. Arms=Articles
2. The system displays it on the Featured tab. $a(t)$: index of chosen article
3. The user clicks the article if he/she is interested in the contents. $r_{a(t)}(t) = 1$ or 0.
4. Based on user click feedback, the system updates its article selection strategy.

Application 2: Mobile Health

Example of mHealth app

1. At specific times in a day, the mHealth system selects one type of message among various types of messages.
2. The system pushes the message to the app user.
3. The user reacts to the message.
4. Based on user reaction, the system updates its message selection strategy.

Application 2: Mobile Health

Example of mHealth app

1. At specific times in a day, the mHealth system selects one type of message among various types of messages. Arms=Messages
2. The system pushes the message to the app user.
3. The user reacts to the message.
4. Based on user reaction, the system updates its message selection strategy.

Application 2: Mobile Health

Example of mHealth app

1. At specific times in a day, the mHealth system selects one type of message among various types of messages. Arms=Messages
2. The system pushes the message to the app user. $a(t)$: index of chosen message
3. The user reacts to the message.
4. Based on user reaction, the system updates its message selection strategy.

Application 2: Mobile Health

Example of mHealth app

1. At specific times in a day, the mHealth system selects one type of message among various types of messages. Arms=Messages
2. The system pushes the message to the app user. $a(t)$: index of chosen message
3. The user reacts to the message. $r_{a(t)}(t)$ =step counts.
4. Based on user reaction, the system updates its message selection strategy.

Other Applications

Applications include..

- mobile healthcare system [Tewari and Murphy, 2017]
- news article recommendation algorithms [Li et al., 2010]
- web page ad placement algorithms [Langford et al., 2008]
- revenue management [Ferreira et al., 2017]
- marketing [Schwartz et al., 2017]
- recommendation systems [Kawale et al., 2015]

Multi-armed bandit (MAB)

- Let $a^*(t) = \underset{1 \le i \le N}{\operatorname{argmax}} \; \theta_i(t)$ and $regret(t) := \theta_{a^*(t)}(t) - \theta_{a(t)}(t)$.
  Goal is to minimize the sum of regrets,

$$R(T) := \sum_{t=1}^{T} regret(t) = \sum_{t=1}^{T} \{\theta_{a^*(t)}(t) - \theta_{a(t)}(t)\}.$$

- Since $\theta_i(t)$'s are unknown, the learner has to **learn** which actions yield maximum mean reward by trying them out.
- Only $r_{a(t)}(t)$ is observed among the whole reward vector $r(t) = [r_1(t), \cdots, r_N(t)]^T$.
  $\Rightarrow$ Trade-off between **exploitation** and **exploration**.

Contextual MAB

- Contextual MABs assume there is context vector $b_i(t) \in \mathbb{R}^d$ associated with each arm i at time t.
- Reward $r_i(t)$ is assumed to depend on $b_i(t)$:

$$\mathbb{E}(r_i(t)|b_i(t)) = \theta(b_i(t)), \quad i = 1, \cdots, N.$$

Linear Contextual MAB

- $\mathbb{E}(r_i(t)|b_i(t))$ is linear in $b_i(t)$, i.e.,

$$\theta(b_i(t)) = b_i(t)^T \mu, \quad i = 1, \cdots, N,$$

  where $\mu \in \mathbb{R}^d$ is unknown.

Linear Contextual MAB

- $\mathbb{E}(r_i(t)|b_i(t))$ is linear in $b_i(t)$, i.e.,

$$\theta(b_i(t)) = b_i(t)^T \mu, \quad i = 1, \cdots, N,$$

  where $\mu \in \mathbb{R}^d$ is unknown.
- Error $\eta_i(t) := r_i(t) - \mathbb{E}(r_i(t)|b_i(t))$ is R-sub-Gaussian, i.e., for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[\exp(\lambda \eta_i(t))\big] \leq \exp\Big(\frac{\lambda^2 R^2}{2}\Big).$$

- WLOG, $||b_i(t)|| \leq 1$, $||\mu|| \leq 1$.

Linear Contextual MAB

Remarks

- $a^*(t) = \underset{1 \le i \le N}{\operatorname{argmax}} \{b_i(t)^T \mu\}$ and,

$$regret(t) = b_{a^*(t)}(t)^T \mu - b_{a(t)}(t)^T \mu.$$

Lower Bounds

- When N is infinite, [Dani et al., 2008] proved that no algorithm can achieve lower bound than $O(d\sqrt{T})$.
- When N is finite, [Chu et al., 2011] proved a lower bound of $O(\sqrt{dT})$ when $d^2 < T$.

Linear Contextual MAB

Remarks

- $a^*(t) = \underset{1\leq i\leq N}{\operatorname{argmax}}\{b_i(t)^T\mu\}$ and,

$$regret(t) = b_{a^*(t)}(t)^T\mu - b_{a(t)}(t)^T\mu.$$

Lower Bounds

- When N is infinite, [Dani et al., 2008] proved that no algorithm can achieve lower bound than $O(d\sqrt{T})$.
- When N is finite, [Chu et al., 2011] proved a lower bound of $O(\sqrt{dT})$ when $d^2 < T$.

But, what if $d \gg T$?

High-dimensional linear contextual MAB

In modern applications, contexts are often high-dimensional with only a sparse subset correlated with the reward.

We consider,

- High-dimensional reward model:

$$\mathbb{E}\big(r_i(t)|b_i(t)\big) = b_i(t)^T\mu, \quad i = 1, \cdots, N,$$

where $\mu \in \mathbb{R}^d$ and $||\mu||_0 = s_0 \ll d$.

High-dimensional linear contextual MAB

- In low dimension,

$$(\text{Gram matrix}) = \sum_{\tau=1}^{t} b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T$$

  is positive definite after $O(d)$ rounds.

- But if $d \gg T$, Gram matrix is singular until the end.

High-dimensional linear contextual MAB

A weaker condition is ..

- **Compatibility Condition [van de Geer and Bühlmann, 2009].** Define $\hat{\Sigma}_t := \frac{1}{t}\sum_{\tau=1}^{t} b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T$ and let $I = supp(\mu)$, the set of indices of non-zero components of μ. Then $\exists \phi_1 > 0$ such that

$$\text{for } \forall v \in \mathbb{R}^d \text{ such that } ||v_{I^c}||_1 \leq 3||v_I||_1, \quad ||v_I||_1^2 \leq \frac{|I|(v^T \hat{\Sigma}_t v)}{\phi_1^2}.$$

High-dimensional linear contextual MAB

Lemma (Lemma 11.2 of [van de Geer and Bühlmann, 2009])

Let $x_\tau \in \mathbb{R}^d$ and $y_\tau \in \mathbb{R}$ be random variables with $y_\tau = x_\tau^T \beta + \varepsilon_\tau,\ \tau = 1, 2, \cdots, t$, where $\beta \in \mathbb{R}^d$, $||\beta||_0 = s_0$, and ε_τ's are i.i.d. gaussian with mean zero and variance R^2. Let $\lambda_t = R\sqrt{\frac{2\log(ed/\delta)}{t}}$ and

$$\hat{\beta}(t) = \operatorname*{argmin}_{\beta}\{\frac{1}{t}\sum_{\tau=1}^{t}(y_\tau - x_\tau^T\beta)^2 + \lambda_t||\beta||_1\}.$$

If $\hat{\Sigma}_t := \frac{1}{t}\sum_{\tau=1}^{t} x_\tau x_\tau^T$ satisfies compatibility condition for some $\phi > 0$, then for $\forall \delta \in (0, 1)$, with probability at least $(1 - \delta/t^2)$,

$$||\hat{\beta}(t) - \beta||_1 \leq \frac{4\lambda_t s_0}{\phi^2} = \frac{4s_0 R}{\phi^2}\sqrt{\frac{4\log(edt/\delta)}{t}}.$$

High-dimensional linear contextual MAB

Lemma (Lemma EC.6 of [Bastani and Bayati, 2015])

Let $x_1, x_2, \cdots, x_t$ *be i.i.d. random vectors in* $\mathbb{R}^d$ *with* $||x_\tau||_\infty \leq 1$ *for all* τ*. Let* $\Sigma = \mathbb{E}[x_\tau x_\tau^T]$ *and* $\hat{\Sigma}_t = \frac{1}{t}\sum_{\tau=1}^t x_\tau x_\tau^T$*. Suppose* Σ *satisfies compatibility for some* $\phi > 0$*. Then if* $c = \min(0.5, \frac{\phi^2}{256 s_0})$ *and* $t \geq \frac{3}{c^2}\log d$*, with probability at least* $1 - \exp\big(-c^2 t\big)$*,* $\hat{\Sigma}_t$ *satisfies compatiblity as well for* $\phi/\sqrt{2}$*.*

High-dimensional linear contextual MAB

Lemma (Lemma EC.6 of [Bastani and Bayati, 2015])

Let $x_1, x_2, \cdots, x_t$ be i.i.d. random vectors in $\mathbb{R}^d$ with $||x_\tau||_\infty \leq 1$ for all τ. Let $\Sigma = \mathbb{E}[x_\tau x_\tau^T]$ and $\hat{\Sigma}_t = \frac{1}{t}\sum_{\tau=1}^{t} x_\tau x_\tau^T$. Suppose Σ satisfies compatibility for some $\phi > 0$. Then if $c = \min(0.5, \frac{\phi^2}{256 s_0})$ and $t \geq \frac{3}{c^2}\log d$, with probability at least $1 - \exp\big(-c^2 t\big)$, $\hat{\Sigma}_t$ satisfies compatiblity as well for $\phi/\sqrt{2}$.

$\Rightarrow$ But $b_{a(t)}(t)$'s are not i.i.d. !

High-dimensional linear contextual MAB

Recall...

- $\{b_{a(1)}(1), r_{a(1)}(1)\}, \{b_{a(2)}(2), r_{a(2)}(2)\}, \cdots, \{b_{a(t)}(t), r_{a(t)}(t)\}$ are highly correlated!

$b_1(1)$	$b_1(2)$	$b_1(3)$	$\cdots$	$b_1(t)$
$b_2(1)$	$b_2(2)$	$b_2(3)$	$\cdots$	$b_2(t)$
$b_3(1)$	$b_3(2)$	$b_3(3)$	$\cdots$	$b_3(t)$
$b_4(1)$	$b_4(2)$	$b_4(3)$	$\cdots$	$b_4(t)$
$b_5(1)$	$b_5(2)$	$b_5(3)$	$\cdots$	$b_5(t)$
$b_6(1)$	$b_6(2)$	$b_6(3)$	$\cdots$	$b_6(t)$
$b_7(1)$	$b_7(2)$	$b_7(3)$	$\cdots$	$b_7(t)$
$b_8(1)$	$b_8(2)$	$b_8(3)$	$\cdots$	$b_8(t)$

High-dimensional linear contextual MAB

Recall...

- $\{b_{a(1)}(1), r_{a(1)}(1)\}, \{b_{a(2)}(2), r_{a(2)}(2)\}, \cdots, \{b_{a(t)}(t), r_{a(t)}(t)\}$ are highly correlated!

$b_1(1)$		$b_1(2)$	$b_1(3)$	$\cdots$	$b_1(t)$
$b_2(1)$		$b_2(2)$	$b_2(3)$	$\cdots$	$b_2(t)$
$b_3(1)$	$r_3(1)$	$b_3(2)$	$b_3(3)$	$\cdots$	$b_3(t)$
$b_4(1)$		$b_4(2)$	$b_4(3)$	$\cdots$	$b_4(t)$
$b_5(1)$		$b_5(2)$	$b_5(3)$	$\cdots$	$b_5(t)$
$b_6(1)$		$b_6(2)$	$b_6(3)$	$\cdots$	$b_6(t)$
$b_7(1)$		$b_7(2)$	$b_7(3)$	$\cdots$	$b_7(t)$
$b_8(1)$		$b_8(2)$	$b_8(3)$	$\cdots$	$b_8(t)$

a(1)=3

High-dimensional linear contextual MAB

Recall...

- $\{b_{a(1)}(1), r_{a(1)}(1)\}, \{b_{a(2)}(2), r_{a(2)}(2)\}, \cdots, \{b_{a(t)}(t), r_{a(t)}(t)\}$ are highly correlated!

$b_1(1)$		$b_1(2)$		$b_1(3)$	$\cdots$	$b_1(t)$
$b_2(1)$		$b_2(2)$		$b_2(3)$	$\cdots$	$b_2(t)$
$b_3(1)$	$r_3(1)$	$b_3(2)$		$b_3(3)$	$\cdots$	$b_3(t)$
$b_4(1)$		$b_4(2)$		$b_4(3)$	$\cdots$	$b_4(t)$
$b_5(1)$		$b_5(2)$		$b_5(3)$	$\cdots$	$b_5(t)$
$b_6(1)$		$b_6(2)$		$b_6(3)$	$\cdots$	$b_6(t)$
$b_7(1)$		$b_7(2)$	$r_7(2)$	$b_7(3)$	$\cdots$	$b_7(t)$
$b_8(1)$		$b_8(2)$		$b_8(3)$	$\cdots$	$b_8(t)$

a(1)=3   a(2)=7

High-dimensional linear contextual MAB

Recall...

- $\{b_{a(1)}(1), r_{a(1)}(1)\}, \{b_{a(2)}(2), r_{a(2)}(2)\}, \cdots, \{b_{a(t)}(t), r_{a(t)}(t)\}$ are highly correlated!

$$
\begin{array}{cccccccc}
b_1(1) & & b_1(2) & & b_1(3) & & \cdots & b_1(t) \\
b_2(1) & & b_2(2) & & b_2(3) & r_2(3) & \cdots & b_2(t) \\
b_3(1) & r_3(1) & b_3(2) & & b_3(3) & & \cdots & b_3(t) \\
b_4(1) & & b_4(2) & & b_4(3) & & \cdots & b_4(t) \\
b_5(1) & & b_5(2) & & b_5(3) & & \cdots & b_5(t) \\
b_6(1) & & b_6(2) & & b_6(3) & & \cdots & b_6(t) \\
b_7(1) & & b_7(2) & r_7(2) & b_7(3) & & \cdots & b_7(t) \\
b_8(1) & & b_8(2) & & b_8(3) & & \cdots & b_8(t)
\end{array}
$$

a(1)=3   a(2)=7   a(3)=2

High-dimensional linear contextual MAB

Recall...

- $\{b_{a(1)}(1), r_{a(1)}(1)\}, \{b_{a(2)}(2), r_{a(2)}(2)\}, \cdots, \{b_{a(t)}(t), r_{a(t)}(t)\}$ are highly correlated!

$b_1(1)$		$b_1(2)$		$b_1(3)$		$\cdots$	$b_1(t)$	
$b_2(1)$		$b_2(2)$		$b_2(3)$	$r_2(3)$	$\cdots$	$b_2(t)$	
$b_3(1)$	$r_3(1)$	$b_3(2)$		$b_3(3)$		$\cdots$	$b_3(t)$	
$b_4(1)$		$b_4(2)$		$b_4(3)$		$\cdots$	$b_4(t)$	$r_4(t)$
$b_5(1)$		$b_5(2)$		$b_5(3)$		$\cdots$	$b_5(t)$	
$b_6(1)$		$b_6(2)$		$b_6(3)$		$\cdots$	$b_6(t)$	
$b_7(1)$		$b_7(2)$	$r_7(2)$	$b_7(3)$		$\cdots$	$b_7(t)$	
$b_8(1)$		$b_8(2)$		$b_8(3)$		$\cdots$	$b_8(t)$	
a(1)=3		a(2)=7		a(3)=2			a(t)=4	

High-dimensional linear contextual MAB

- We can assume $\{\bar{b}(t) = \frac{1}{N}\sum_{i=1}^{N} b_i(t)\}$ is i.i.d.
- Bandit data is missing data!

$b_1(1)$		$b_1(2)$		$b_1(3)$		$\cdots$	$b_1(t)$	
$b_2(1)$		$b_2(2)$		$b_2(3)$	$r_2(3)$	$\cdots$	$b_2(t)$	
$b_3(1)$	$r_3(1)$	$b_3(2)$		$b_3(3)$		$\cdots$	$b_3(t)$	
$b_4(1)$		$b_4(2)$		$b_4(3)$		$\cdots$	$b_4(t)$	$r_4(t)$
$b_5(1)$		$b_5(2)$		$b_5(3)$		$\cdots$	$b_5(t)$	
$b_6(1)$		$b_6(2)$		$b_6(3)$		$\cdots$	$b_6(t)$	
$b_7(1)$		$b_7(2)$	$r_7(2)$	$b_7(3)$		$\cdots$	$b_7(t)$	
$b_8(1)$		$b_8(2)$		$b_8(3)$		$\cdots$	$b_8(t)$	
a(1)=3		a(2)=7		a(3)=2			a(t)=4	

High-dimensional linear contextual MAB

- Let $\mathcal{H}_{t-1}$ be history until time $t-1$,

$$\mathcal{H}_{t-1} = \{a(\tau), r_{a(\tau)}(\tau), \{b_i(\tau)\}_{i=1}^N, \tau = 1, \cdots, t-1\}.$$

- Define filtration $\mathcal{F}_{t-1} = \{\mathcal{H}_{t-1}, \{b_i(t)\}_{i=1}^N\}$.

- Let $\pi_i(t)$ be action selection probability,

$$\pi_i(t) = \mathbb{P}(a(t) = i | \mathcal{F}_{t-1}).$$

Remark $\pi_i(t)$ is observation probability. In bandits, $\pi_i(t)$ is controlled by the learner, i.e., the value is known.

Doubly-Robust Lasso Bandit

We construct a doubly-robust pseudo reward.

$$\begin{aligned}\hat{r}(t) &= \frac{1}{N}\sum_{i=1}^{N}\hat{r}_i(t)\\ &= \frac{1}{N}\sum_{i=1}^{N}\Big\{\frac{I(a(t)=i)}{\pi_i(t)}r_i(t) + \Big(1-\frac{I(a(t)=i)}{\pi_i(t)}\Big)b_i(t)^T\hat{\mu}(t-1)\Big\},\end{aligned}$$

where $\hat{\mu}(t-1)$ is the Lasso estimate of μ obtained from last step. Whether or not $\hat{\mu}(t-1)$ is a valid estimate, this value has conditional expectation $\bar{b}(t)^T\mu$ given that $\pi_i(t) > 0$ for all i:

$$\mathbb{E}[\hat{r}(t)|\mathcal{F}_{t-1}] = \bar{b}(t)^T\mu.$$

$\Rightarrow$ Apply Lasso on $\{(\bar{b}(\tau), \hat{r}(\tau))\}_{\tau=1}^{t}$ instead of $\{(b_{a(\tau)}(\tau), r_{a(\tau)}(\tau))\}_{\tau=1}^{t}$.

Doubly-Robust Lasso Bandit

Note that

$$
\begin{aligned}
\hat{r}(t) &= \frac{1}{N}\sum_{i=1}^{N}\left\{b_i(t)^T\hat{\mu}(t-1) + \frac{I(a(t)=i)}{\pi_i(t)}\left\{r_i(t) - b_i(t)^T\hat{\mu}(t-1)\right\}\right\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left\{b_i(t)^T\hat{\mu}(t-1) + \frac{I(a(t)=i)}{\pi_i(t)}\left\{\eta_i(t) + b_i(t)^T\left(\mu - \hat{\mu}(t-1)\right)\right\}\right\}
\end{aligned}
$$

If we have $||\hat{\mu}(t-1) - \mu||_1 \leq O\left(\sqrt{\frac{\log t}{t}}\right)$, and if we set $\pi_i(t) \geq O\left(\frac{1}{N}\sqrt{\frac{\log t}{t}}\right)$, the variance of $\hat{r}(t)$ is constant scale!

Doubly-Robust Lasso Bandit

To ensure $\pi_i(t) \geq O\left(\frac{1}{N}\sqrt{\frac{\log t}{t}}\right)$, we do:

Generate $m_t \sim Ber\left(O(\sqrt{\frac{\log t}{t}})\right)$.

- **if** $m_t = 1$ **then**

$$\text{Pull arm } a(t) = i \text{ with probability } \frac{1}{N} \ (i = 1, \cdots, N) \tag{1}$$

- **else**

$$\text{Pull arm } a(t) = \operatorname*{argmax}_{1 \leq i \leq N} \{b_i(t)^T \hat{\mu}(t-1)\}. \tag{2}$$

Remark: Step (1) is exploration, step (2) is exploitation.

Doubly-Robust Lasso Bandit

- Step (1) induces suboptimal choice of arms.
- Let $R(T,1)$ be sum of regrets due to step (1). Then $R(T,1) \leq \sum_{t=1}^{T} m_t$. Due to Hoeffding's inequality, with probability at least $1-\delta$,

$$R(T,1) \leq \sum_{t=1}^{T} m_t \leq \sum_{t=1}^{T} \mathbb{E}(m_t) + \sqrt{T \log(1/\delta)/2}.$$

where

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}(m_t) &= O\Big(\sum_{t=1}^{T} \sqrt{\frac{\log t}{t}}\Big) \leq O\Big(T \frac{1}{T} \sum_{t=1}^{T} \sqrt{\frac{\log T}{t}}\Big) \\
&\leq O\Big(T \sqrt{\frac{1}{T} \sum_{t=1}^{T} \frac{\log T}{t}}\Big) \quad (\because \text{Jensen's inequality}) \\
&= O(\sqrt{T} \log T).
\end{aligned}$$

Doubly-Robust Lasso Bandit

- In Step (2),

$$\begin{aligned} b_{a^*(t)}(t)^T\mu &\leq b_{a^*(t)}(t)^T\hat{\mu}(t-1) + ||\hat{\mu}(t-1)-\mu||_1 \\ &\leq b_{a(t)}(t)^T\hat{\mu}(t-1) + ||\hat{\mu}(t-1)-\mu||_1 \\ &\leq b_{a(t)}(t)^T\mu + ||\hat{\mu}(t-1)-\mu||_1 + ||\hat{\mu}(t-1)-\mu||_1 \end{aligned}$$

$$\Rightarrow regret(t) \leq 2||\hat{\mu}(t-1)-\mu||_1$$

Doubly-Robust Lasso Bandit

- Let $R(T,2)$ be sum of regrets from step (2). With probability at least $1-\delta$,

$$
\begin{aligned}
R(T,2) &\leq 2\sum_{t=1}^{T} ||\hat{\mu}(t-1)-\mu||_1 \\
&\leq 2\sum_{t=1}^{T} \frac{4s_0R}{\phi^2}\sqrt{\frac{4\log(\mathrm{e}dt/\delta)}{t}} \\
&\leq O\Big(s_0\sqrt{T}\log(dT)\Big)
\end{aligned}
$$

Doubly-Robust Lasso Bandit

Theorem

For $\forall\delta \in (0,1)$, *with probability at least* $1-\delta$,

$$R(T) \leq O\big(s_0\sqrt{T}\log(dT)\big).$$

Related Works

- [Abbasi-Yadkori et al., 2012]: given any online prediction algorithm for the regression parameter, their algorithm constructs a high-probability confidence set for the true parameter using a bound on the prediction loss, and then pulls arms according to the *optimism in the face of uncertainty* rule. Their high-probability regret bound is proportional to $\sqrt{d}$ instead of $\log d$, so is not sublinear in T when d scales with T.
- [Carpentier and Munos, 2012]: used an explicit exploration phase to identify the support of the regression parameter using techniques from compressed sensing. Their regret bound is tight scaling with $\log d$, but the algorithm is specific to the case where the set of arms is the unit ball for the $||\cdot||_2$ norm and fixed over time.
- [Gilton and Willett, 2017]: leveraged ideas from linear Thompson Sampling and Relevance Vector Machines [Tipping, 2001]. The theoretical results are weak since they derived the regret bound under the assumption that a sufficiently small superset of the support for the regression parameter is known in advance.

Related Works

- [Bastani and Bayati, 2015]: proposed the Lasso Bandit under a different reward model,

$$\mathrm{E}\big(r_i(t)|b_i(t)\big) = b(t)^T\beta_i \tag{1}$$

  with $||\beta_i|| = s_0$, $i = 1, \cdots, N$. They imposed compatibility through forced-sampling of each arm. The upper bound of the *expected* regret is proportional to number of arms, N:

$$\mathbb{E}[R(T)] \leq O\big(Ns_0^2[\log T + \log d]^2\big).$$

  An application of the Hoeffding's inequality gives an additional term of order $O(\sqrt{T})$ for the high-probability bound.

- [Wang et al., 2018]: proposed the Minimax Concave Penalized (MCP) Bandit algorithm for the reward model (1), which uses forced-sampling along with the MCP estimator [Zhang, 2010]. They obtained,

$$\mathbb{E}[R(T)] \leq O\big(Ns_0^2[s_0 + \log d]\log T\big).$$

Experiments

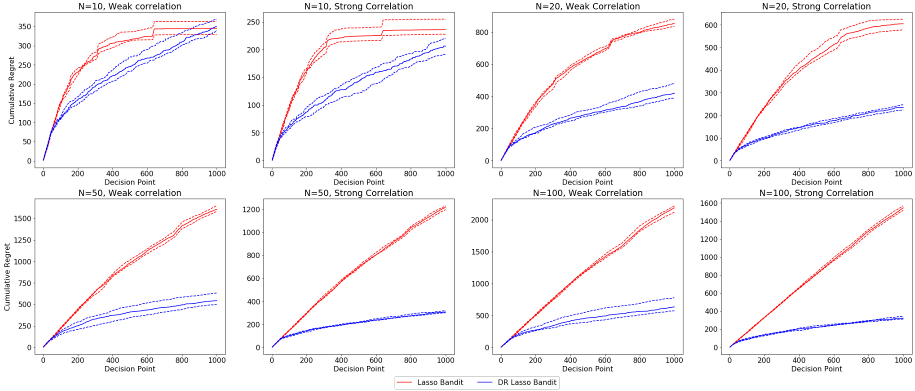
We compare the Doubly-Robust Lasso Bandit with Lasso Bandit [Bastani and Bayati, 2015].

- Number of arms: $N = 50$ or 100.
- Dimension of context vector: $d = 100$.
- Distribution of context vector: for fixed $j = 1, \cdots, d$, $[b_{1j}(t), \cdots, b_{Nj}(t)]^T \sim \mathcal{N}(0_N, V)$, where $V(i,i) = 1$ for every i and $V(i,k) = \rho^2$ for every $i \neq k$. $\rho^2 = 0.3$ (weak correlation) or $\rho^2 = 0.7$ (strong correlation).
- Regression parameter: $s_0 = 5$ and $\beta_{supp(\beta)} \sim U([0,1])^5$.
- Distribution of the reward:

$$r_i(t) = b_i(t)^T \beta + \eta_i(t),$$

  where $\eta_i(t) \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.05^2)$.

Experiments



Thank You !