Doubly-Robust Lasso Bandit

Gi-Soo Kim & Myunghee Cho Paik, *Department of Statistics, Seoul National University*

gisoo1989@snu.ac.kr & myungheechopaik@snu.ac.kr

MOTIVATION

- Contextual multi-armed bandit (MAB) algorithms are widely used in sequential decision tasks such as news article recommendation systems, web page ad placement algorithms, and mobile health.
- Many algorithms require the dimension of the context (d) not be too large. The cumulative regrets are proportional to a polynomial function of d.
- In modern applications however, web or mobile-based contextual variables are often high-dimensional, with only a sparse subset of s_0 variables related to the reward.

MAIN RESULTS

- We propose a new linear contextual multi-armed bandit algorithm for **high-dimensional, sparse** reward models.
- We construct a new estimator for the linear regression parameter using **Lasso** along with the **context information of all arms** through techniques from missing data literature.
- The high-probability regret upper bound is tight, does not depend on number of arms, and scales with $\log d$ instead of a polynomial function of d.

SETTINGS

- Set of arms at time t ($t = 1, \cdots, T$):

$$\{b_1(t), \cdots, b_N(t)\} \overset{i.i.d.}{\sim} \mathcal{P}_b,$$

  where $||b_i(t)||_2 \leq 1$ and $\mathcal{P}_b$ is some distribution over $\mathbb{R}^{N\times d}$.
- Reward of i-th arm at time t:

$$r_i(t) = b_i(t)^T\beta + \eta_i(t), \quad i = 1, \cdots, N,$$

  where $\beta \in \mathbb{R}^d$ is **sparse** with $||\beta||_0 = s_0 (\ll d)$ and $\eta_i(t)$ is R-sub-Gaussian for some $0 < R < O(\sqrt[4]{\log T/T})$.
- The optimal arm at time t is $a^*(t) := \underset{1\leq i\leq N}{\text{argmax}}\{b_i(t)^T\beta\}$.
- At time t, the learner pulls one arm $a(t)$ with probability $\pi_{a(t)}(t)$, and observes $r_{a(t)}(t)$.
- **Goal:** minimize sum of regrets,

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

CHALLENGES

- Lasso is a good tool for estimating a high-dimensional, sparse regression parameter.
- Lasso estimate has fast convergence under assumption that covariates are **compatible**, i.e., **not too correleated**. Under minor conditions, compatibility holds for i.i.d. data.
- In the contextual MAB setting however, the contexts of the chosens arms, $b_{a(t)}(t)'s$, tend to be highly correlated as the learner adapts his (her) decision rule.

PROPOSED METHOD

Algorithm 1 Doubly-Robust Lasso Bandit algorithm

```
Input parameters: λ_1, λ_2, z_T
Set β̂(0) = 0_d, 𝕊 = { }.
for t = 1, 2, ··· , T do
    Observe {b_1(t), b_2(t), ··· , b_N(t)} ~ P_b
    if t ≤ z_T then
        Pull arm a(t) = i with probability 1/N (i = 1, ··· , N)
        π_{a(t)}(t) ← 1/N
    else
        λ_{1t} ← λ_1 sqrt((log t + log d)/t), sample m_t ~ Ber(λ_{1t})
        if m_t = 1 then
            Pull arm a(t) = i with probability 1/N (i = 1, ··· , N)
        else
            Pull arm a(t) = argmax_{1≤i≤N} {b_i(t)^T β̂(t − 1)}
        end if
        π_{a(t)}(t) ← λ_{1t}/N + (1 − λ_{1t}) I{ a(t) = argmax_{1≤i≤N} {b_i(t)^T β̂(t − 1)} }
    end if
    Observe r_{a(t)}(t)
    b̄(t) ← (1/N) Σ_{i=1}^N b_i(t),   r̂(t) ← b̄(t)^T β̂(t − 1) + (1/N) (r_{a(t)}(t) − b_{a(t)}(t)^T β̂(t−1)) / π_{a(t)}(t)
    𝕊 ← 𝕊 ∪ { (b̄(t), r̂(t)) }
    λ_{2t} ← λ_2 sqrt((log t + log d)/t)
    β̂(t) ← argmin_β { (1/t) Σ_{(b̄,r̂)∈𝕊} (r̂ − b̄^T β)^2 + λ_{2t} ||β||_1 }
end for
```

Instead of applying Lasso on the pairs $(b_{a(\tau)}(\tau), r_{a(\tau)}(\tau))$, $\tau = 1, \cdots, t$, we **apply Lasso on the pairs** $(\bar{b}(\tau), \hat{r}(\tau))$, $\tau = 1, \cdots, t$, where

$$\bar{b}(\tau) = \frac{1}{N}\sum_{i=1}^{N} b_i(\tau)$$

$$\hat{r}(\tau) = \bar{b}(\tau)^T\hat{\beta}(\tau-1) + \frac{1}{N}\frac{r_{a(\tau)}(\tau) - b_{a(\tau)}(\tau)^T\hat{\beta}_{\tau-1}}{\pi_{a(\tau)}(\tau)},$$

where $\hat{\beta}_{\tau-1}$ is the β estimate of the previous step.

- As opposed to $b_{a(\tau)}(\tau)$'s, the average contexts $\bar{b}(\tau)$'s are i.i.d. $\Rightarrow$ the average contexts **satisfy compatibility condition.**
- The pseudo-reward $\hat{r}(\tau)$ is the **doubly-robust** estimate of the reward corresponding to the average context $\bar{b}(\tau)$.
  - $\hat{r}(\tau)$ is **unbiased** for $\bar{b}(\tau)^T\beta$ given filtration $\mathcal{F}_{\tau-1}$:

$$\begin{aligned}\mathbb{E}_\tau[\hat{r}(\tau)] &= \mathbb{E}_\tau\Big[\frac{1}{N}\sum_{i=1}^{N}\Big(1 - \frac{I(a(\tau)=i)}{\pi_i(\tau)}\Big)b_i(\tau)^T\hat{\beta}_{\tau-1} + \frac{1}{N}\sum_{i=1}^{N}\frac{I(a(\tau)=i)}{\pi_i(\tau)}r_i(\tau)\Big] \\ &= \mathbb{E}_\tau\Big[\frac{1}{N}\sum_{i=1}^{N} r_i(\tau)\Big] = \bar{b}(\tau)^T\beta,\end{aligned}$$

    where $\mathbb{E}_\tau[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{\tau-1}]$.
  - $\hat{r}(\tau)$ has **constant variance** under restriction $\pi_{a(\tau)}(\tau) \geq O\Big(\frac{1}{N}\sqrt{(\log d + \log\tau)/\tau}\Big)$ if $\hat{\beta}_{\tau-1}$ fasly converges, i.e., $||\hat{\beta}(\tau-1) - \beta||_1 \leq O(\sqrt{(\log d + \log\tau)/\tau})$.

$\Rightarrow$ The resulting estimate $\hat{\beta}_t$ fastly converges. By induction, the next pseudo-reward $\hat{r}(t+1)$ is unbiased and has constant variance, so $\hat{\beta}_{t+1}$ fastly converges, and so on....

Theorem 1. *With high probability proposed algorithm achieves,*

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

EXPERIMENTS

- We compare the Doubly-Robust Lasso Bandit with Lasso Bandit (Bastani and Bayati, 2015), which assumes a different reward model, $r_i(t) = b(t)^T\beta_i$ with $||\beta_i|| = s_0$ and imposes compatibility through forced-sampling of each arm.
- We set $d = 100$ and $s_0 = 5$.
- We conduct 10 replications for each case. (For each algorithm, we use the value of the tuning parameter which incurs minimum median regret, which is found by grid search.)

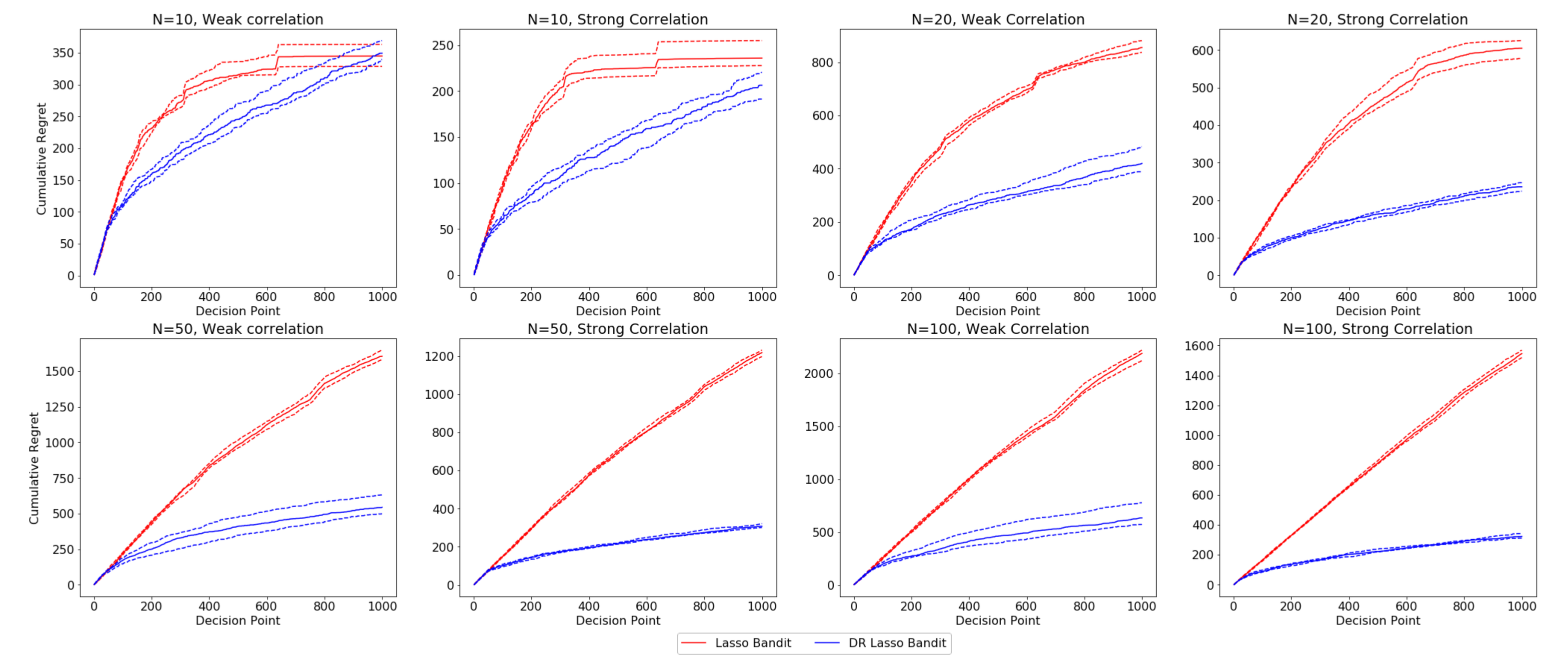

Figure 1: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret.

Acknowledgements

This work was supported by the National Research Foundation of Korea under grant NRF-2017R1A2B4008956.