

# 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  $s_0 (\ll d)$  variables related to the reward.

## SETTINGS

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

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

where  $\mathcal{P}_b$  is some distribution over  $\mathbb{R}^{N \times d}$ .

- Reward of pulling  $i$ -th arm at time  $t$ :

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

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  $0 < R < O(\sqrt{\log T/T})$ .

- The optimal arm at time  $t$  is  $a^*(t) := \operatorname{argmax}_{1 \leq i \leq N} \{b_i(t)^T \beta\}$ .
- At time  $t$ , the learner chooses and pulls one arm  $a(t)$  according to probability vector  $[\pi_1(t), \dots, \pi_N(t)]$  ( $\sum_{i=1}^N \pi_i(t) = 1$ ), and observes  $r_{a(t)}(t)$ .
- Goal:** minimize sum of regrets,

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

## MAIN RESULTS

- We propose a new linear contextual MAB 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  $d$ .

## 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 correlated**. Under minor conditions, compatibility holds for i.i.d. data.

- In contextual MABs however, contexts of the chosen arms  $b_{a(t)}(t)$ 's tend to be highly correlated as the learner adapts his (her) decision rule.

## PROPOSED METHOD

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

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

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

and  $\hat{\beta}_{\tau-1}$  is the  $\beta$  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 unbiased, **doubly-robust** estimate (Bang and Robins, 2005) of the reward corresponding to the average context  $\bar{b}(\tau)$ .

---

**Algorithm 1** Doubly-Robust Lasso Bandit algorithm

---

Inputs:  $\lambda_1, \lambda_2, z_T, \hat{\beta}_0 = 0_d, \mathbb{S} = \{\}$ .

**for**  $t = 1, 2, \dots, T$  **do**

Observe  $\{b_1(t), b_2(t), \dots, b_N(t)\} \sim \mathcal{P}_b$

**if**  $t \leq z_T$  **then**

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

$\pi_{a(t)}(t) \leftarrow 1/N$

**else**

$\lambda_{1t} \leftarrow \lambda_1 \sqrt{(\log t + \log d)/t}$ , sample  $m_t \sim \text{Ber}(\lambda_{1t})$

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

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

**else**

Pull arm  $a(t) = \operatorname{argmax}_{1 \leq i \leq N} \{b_i(t)^T \hat{\beta}_{t-1}\}$

**end if**

$\pi_{a(t)}(t) \leftarrow \lambda_{1t}/N + (1 - \lambda_{1t}) I\left\{a(t) = \operatorname{argmax}_{1 \leq i \leq N} \{b_i(t)^T \hat{\beta}_{t-1}\}\right\}$

**end if**

Observe  $r_{a(t)}(t)$

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

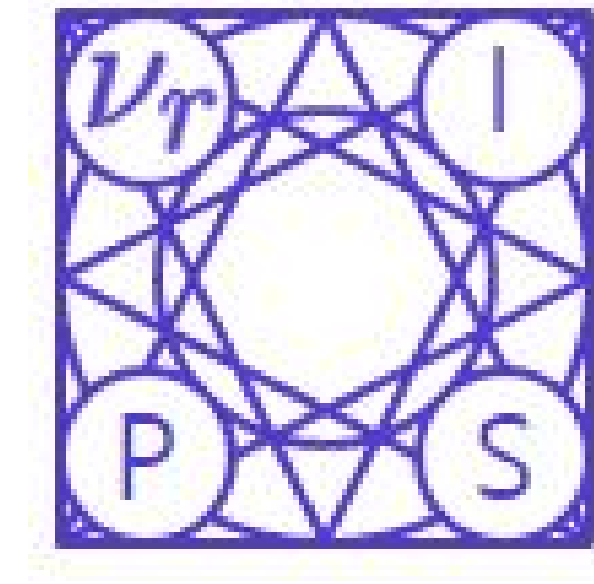
$\mathbb{S} \leftarrow \mathbb{S} \cup \{(\bar{b}(t), \hat{r}(t))\}$

$\lambda_{2t} \leftarrow \lambda_2 \sqrt{(\log t + \log d)/t}$

$\hat{\beta}_t \leftarrow \operatorname{argmin}_{\beta} \left\{ \frac{1}{t} \sum_{(\bar{b}, \hat{r}) \in \mathbb{S}} (\hat{r} - \bar{b}^T \beta)^2 + \lambda_{2t} \|\beta\|_1 \right\}$

**end for**

---



## PROPERTIES OF $\hat{r}(\tau)$

- UNBIASEDNESS**  $\hat{r}(\tau)$  is **unbiased** for  $\bar{b}(\tau)^T \beta$  given filtration  $\mathcal{F}_{\tau-1}$ :

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

where  $\mathbb{E}_{\tau}[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_{\tau-1}]$ ,  $\mathcal{F}_{\tau-1} = \{\mathcal{H}_{\tau-1}, \{b_i(\tau)\}_{i=1}^N\}$ ,  $\mathcal{H}_{\tau-1}$  is history until time  $\tau - 1$ .

- BOUNDED VARIANCE** Suppose  $\hat{\beta}_{\tau-1}$  enjoys fast convergence,  $\|\hat{\beta}_{\tau-1} - \beta\|_1 \leq O(\sqrt{\log(d\tau)/\tau})$ . Then  $\hat{r}(\tau)$  has **constant variance** under restriction  $\pi_{a(\tau)}(\tau) \geq O(\frac{1}{N} \sqrt{\log(d\tau)/\tau})$ .

$\Rightarrow$  The resulting estimate  $\hat{\beta}_t$  using  $\{\bar{b}(\tau), \hat{r}(\tau)\}_{\tau=1}^t$  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...

$\Rightarrow$  We ensure  $\pi_{a(\tau)}(\tau) \geq O(\frac{1}{N} \sqrt{\log(d\tau)/\tau})$  by randomizing the arm selection between uniform random selection and deterministic selection based on  $\hat{\beta}_{\tau-1}$ . Uniform random selection leads to suboptimal choice of arms. However, probability of doing uniform random selection decreases with time.

**Theorem.** With high probability,  $R(T) \leq O(s_0 \log(dT) \sqrt{T})$ .

## 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  $s_0 = 5$ .

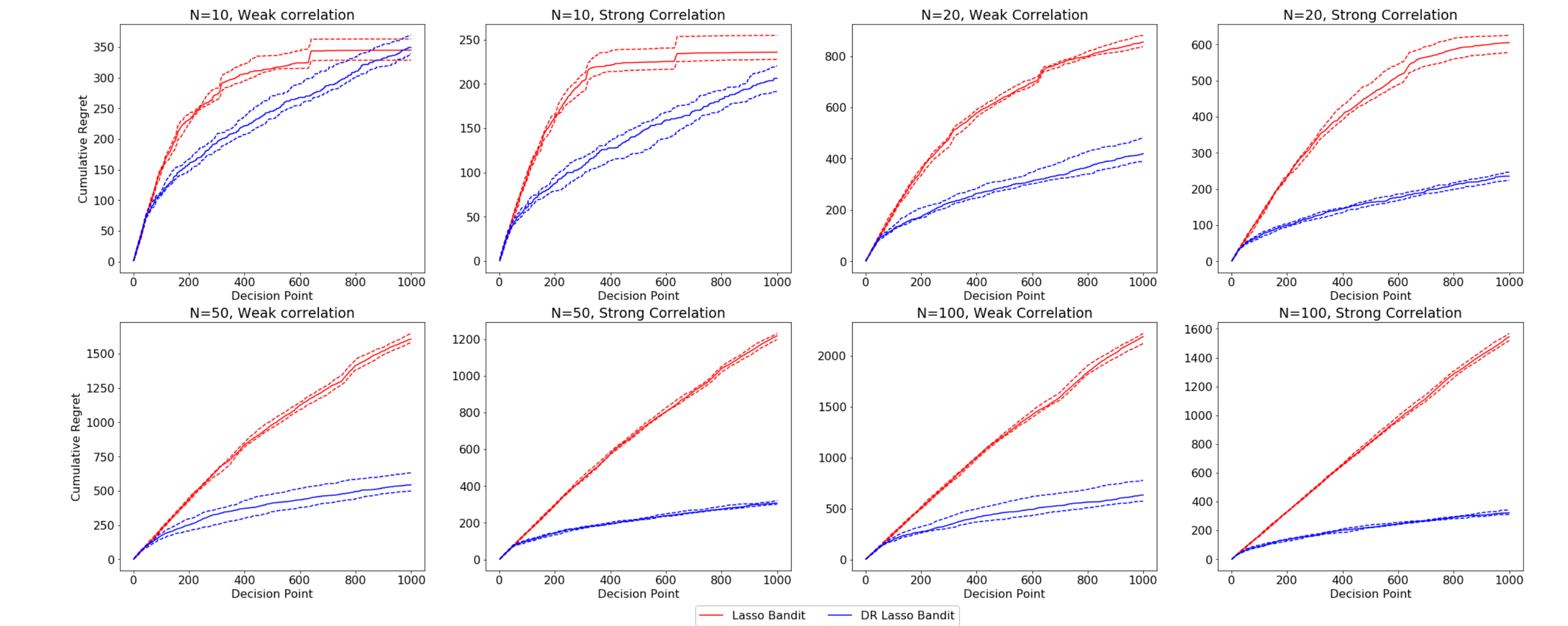


Figure 1: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 10 replications.

## Acknowledgements

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