

# Contextual Multi-armed Bandit Algorithm for Semiparametric Reward Model

---

Gi-Soo Kim, Myunghee Cho Paik

Seoul National University

June 13, 2019

# Introduction

- We propose a new contextual multi-armed bandit (MAB) algorithm for the **nonstationary semiparametric reward model**.
- The proposed method is **less restrictive, easier to implement and computationally faster** than previous works.
- The high-probability upper bound of the regret for the proposed method is of the same order as the Thompson Sampling algorithm for linear reward models.
- We propose a **new estimator for the regression parameter** without requiring an extra tuning parameter and prove that it converges to the true parameter faster than existing estimators.

# Motivation: News article recommendation

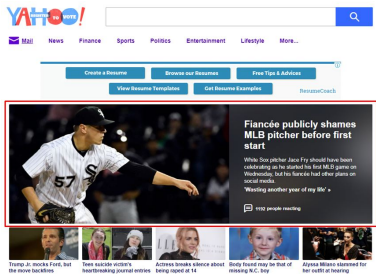


Figure 1: Yahoo! front page snapshot

- 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.
- 5 The web system repeats steps 1-4.

## Remark

This problem can be framed as a multi-armed bandit (MAB) problem [Robbins, 1952, Lai and Robbins, 1985].

# Contextual MAB problem

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

$$\mathbb{E}(r_i(t)|b_i(t), \mathcal{H}_{t-1}) = \theta_t(b_i(t)), \quad i = 1, \dots, N,$$

where

$b_i(t) : \in \mathbb{R}^d$ , context vector of arm  $i$  at time  $t$ ,

$\mathcal{H}_{t-1}$ : observed data until time  $t-1$ ,

$\theta_t(\cdot)$ : unknown function.

- At time  $t$ , the learner pulls arm  $a(t)$ , and observes the reward  $r_{a(t)}(t)$ .
- The optimal arm at time  $t$  is  $a^*(t) := \operatorname{argmax}_{1 \leq i \leq N} \{\theta_t(b_i(t))\}$ .
- **Goal** is to minimize sum of regrets,

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

# Contextual MAB problem

- **Linear** contextual MABs assume a **stationary** reward model,

$$\theta_t(b_i(t)) = b_i(t)^T \mu.$$

- We consider a **nonstationary, semiparametric** reward model,

$$\theta_t(b_i(t)) = \nu(t) + b_i(t)^T \mu.$$

## Remarks

- The nonparametric  $\nu(t)$  represents the **baseline tendency** of the user visiting at time  $t$  to click any article on the Featured tab.
- $\nu(t)$  can depend on history,  $\mathcal{H}_{t-1}$
- The optimal arm is solely determined by  $\mu$ :  $a^*(t) = \underset{1 \leq i \leq N}{\operatorname{argmax}} \{b_i(t)^T \mu\}$ .  
⇒ We don't need to estimate  $\nu(t)$ ! We only need to estimate  $\mu$ !

- Additional assumption:  $\eta_i(t) := r_i(t) - \theta_t(b_i(t))$  is  $R$ -sub-Gaussian.

# Proposed Method

We propose,

- **Thompson sampling** framework [Agrawal and Goyal, 2013]:

$$a(t) = \operatorname{argmax}_{1 \leq i \leq N} \{b_i(t)^T \tilde{\mu}(t)\}, \text{ where } \tilde{\mu}(t) \sim \mathcal{N}(\hat{\mu}(t), v^2 B(t)^{-1}).$$

$\Rightarrow \pi_i(t) := \mathbb{P}(a(t) = i | \mathcal{H}_{t-1}, b(t))$  needs not to be solved. It is determined by Gaussian distribution of  $\tilde{\mu}(t)$ .

- **New estimator for  $\mu$**  based on a centering trick on  $b_{a(t)}(t)$ :

$$\hat{\mu}(t) = \left( I_d + \sum_{\tau=1}^{t-1} \{X_\tau X_\tau^T + \mathbb{E}(X_\tau X_\tau^T | \mathcal{H}_{\tau-1}, b(\tau))\} \right)^{-1} \sum_{\tau=1}^{t-1} 2X_\tau r_{a(\tau)}(\tau),$$

where  $X_\tau = b_{a(\tau)}(\tau) - \bar{b}(\tau)$  and

$$\bar{b}(\tau) = \mathbb{E}(b_{a(\tau)}(\tau) | \mathcal{H}_{\tau-1}, b(\tau)) = \sum_{i=1}^N \pi_i(\tau) b_i(\tau).$$

# Proposed Method

---

## Algorithm 1 Proposed algorithm

---

- 1: Set  $B(1) = I_d$ ,  $y = 0_d$ ,  $v = (2R + 6)\sqrt{6d\log(T/\delta)}$ .
  - 2: **for**  $t = 1, 2, \dots, T$  **do**
  - 3:     Compute  $\hat{\mu}(t) = B(t)^{-1}y$ .
  - 4:     Sample  $\tilde{\mu}(t)$  from distribution  $\mathcal{N}(\hat{\mu}(t), v^2 B(t)^{-1})$ .
  - 5:     Pull arm  $a(t) := \operatorname{argmax}_{i \in \{1, \dots, N\}} b_i(t)^T \tilde{\mu}(t)$ .
  - 6:     Compute probabilities  $\pi_i(t) = \mathbb{P}(a(t) = i | \mathcal{F}_{t-1})$  for  $i = 1, \dots, N$ .
  - 7:     Observe reward  $r_{a(t)}(t)$  and update:  
       
$$B(t+1) = B(t) + (b_{a(t)}(t) - \bar{b}(t))(b_{a(t)}(t) - \bar{b}(t))^T + \{\sum_i \pi_i(t) b_i(t) b_i(t)^T - \bar{b}(t) \bar{b}(t)^T\},$$
  
       
$$y = y + 2(b_{a(t)}(t) - \bar{b}(t))r_{a(t)}(t).$$
  - 8: **end for**
-

# Proposed Method

## Remarks

- In [Krishnamurthy et al., 2018],  $\pi_i(t)$  should be solved out from a convex program with  $N$  quadratic conditions. The authors only showed the existence of such solution when  $N > 2$ .
- [Greenewald et al., 2017] proposed to center the reward instead of the context. The regret of their algorithm depends on  $M = 1/\min\{\pi_1(t)(1 - \pi_1(t))\}$ . Hence, [Greenewald et al., 2017] considers restricted policy,  $p_{min} < \pi_1(t) < p_{max}$ , where  $p_{min} > 0$  and  $p_{max} < 1$ .
- [Krishnamurthy et al., 2018] proposed

$$\hat{\mu}(t) = \left(\gamma I_d + \sum_{\tau=1}^{t-1} X_{\tau} X_{\tau}^T\right)^{-1} \sum_{\tau=1}^{t-1} X_{\tau} r_{a(\tau)}(\tau),$$

but a tight regret bound is valid under  $\gamma \geq 4d\log(9T) + 8\log(4T/\delta)$  when  $N > 2$ , which can overwhelm the denominator when  $t$  is small.



# Proposed Method

## Theorem

With probability at least  $1 - \delta$ , the proposed algorithm achieves,

$$R(T) \leq O\left(d^{3/2}\sqrt{T}\sqrt{\log(Td)\log(T/\delta)}\left(\sqrt{\log(1 + T/d)} + \sqrt{\log(1/\delta)}\right)\right).$$

## Remarks

- Same order (in  $T$ ) as original Thompson sampling for linear model.
- There is no big constant  $M$  multiplied!

# Proposed Method

Table 1: Comparison of the 3 semiparametric contextual MAB algorithms.

Properties	ACTS*	BOSE**	Proposed TS
Restriction on $\pi(t)$	$\pi_1(t) \in [p_{min}, p_{max}]$	None	None
Derivation of $\pi(t)$	from $\tilde{\mu}(t)$	not specified when $N > 2$	from $\tilde{\mu}(t)$
# of Computations per step	$O(1)$	$O(N^2)$	$O(N)$
Tuning parameters	1	2	1
$R(T)$	$O(Md^{\frac{3}{2}}\sqrt{T}\sqrt{\log(T/\delta)^3})$	$O(d\sqrt{T}\log(T/\delta))$	$O(d^{\frac{3}{2}}\sqrt{T}\sqrt{\log(T/\delta)^3})$

\*: [Greenewald et al., 2017]

\*\* : [Krishnamurthy et al., 2018]

# Simulation

## Simulation settings

- Number of arms:  $N = 2$  or  $N = 6$ .
- Dimension of context vector  $b_i(t)$ :  $d = 10$ .
- Distribution of the reward:

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

where  $\eta_i(t) \sim \mathcal{N}(0, 0.1^2)$ , and

$\mu = [-0.55, 0.666, -0.09, -0.232, 0.244, 0.55, -0.666, 0.09, 0.232, -0.244]^T$ .

- Algorithms: Thompson Sampling, Action-Centered TS, BOSE, Proposed TS

# Simulation: $N = 2$

- Case (1):  $\nu(t) = 0$

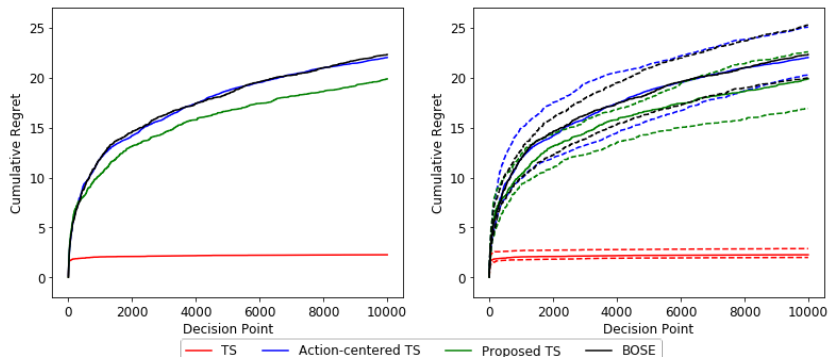


Figure 2: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations for case (1)

## Simulation: $N = 2$

- Case (2):  $\nu(t) = -b_{a^*(t)}(t)^T \mu$

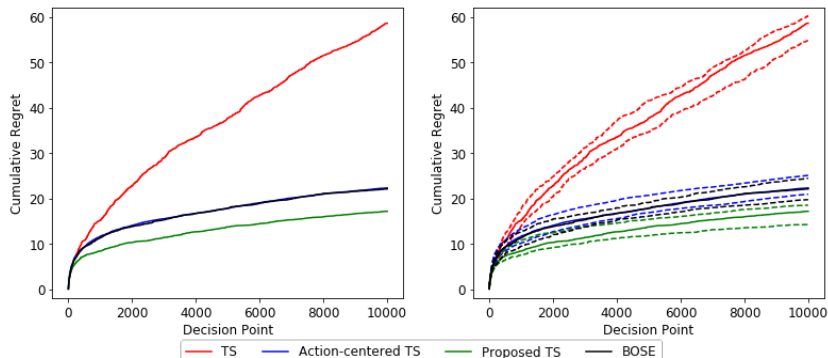


Figure 3: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations for case (2)

## Simulation: $N = 2$

- Case (3):  $\nu(t) = \log(t + 1)$

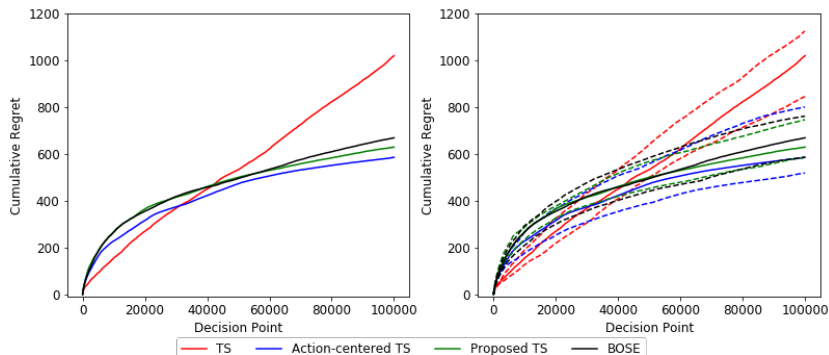


Figure 4: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations for case (4)

## Simulation: $N = 6$

- Case (1):  $\nu(t) = 0$

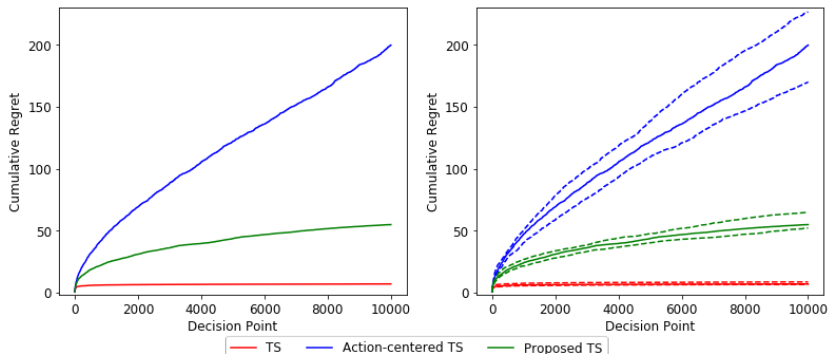


Figure 5: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations for case (1)

## Simulation: $N = 6$

- Case (2):  $\nu(t) = -b_{a^*(t)}(t)^T \mu$

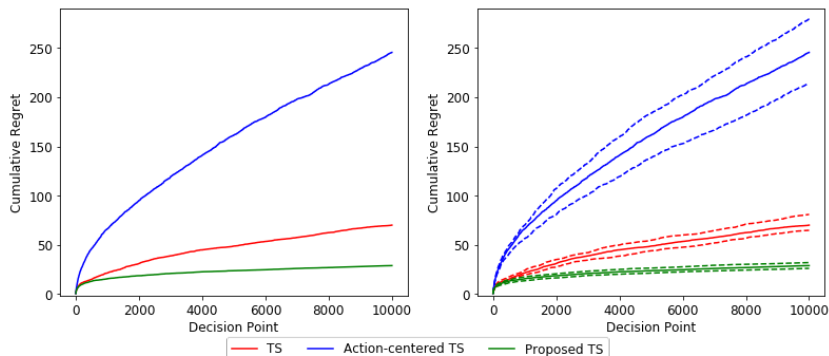


Figure 6: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations for case (2)



## Simulation: $N = 6$

- Case (3):  $\nu(t) = \log(t + 1)$

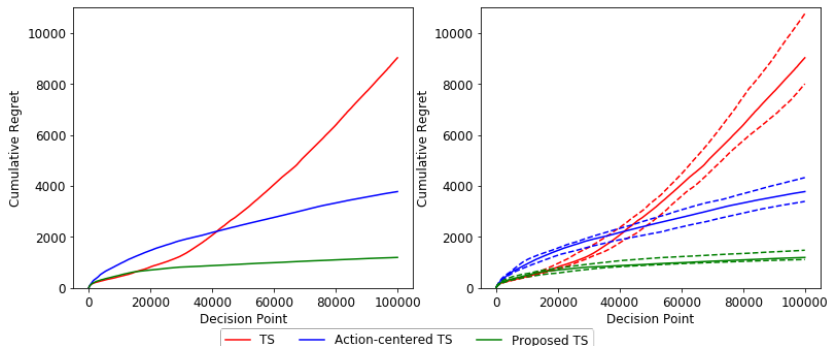
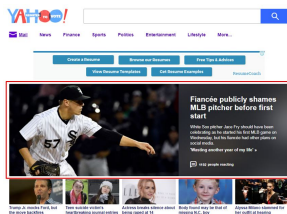


Figure 7: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations for case (4)

# Real data application



- Log data of user clicks from May 1st, 2009 to May 10th, 2009. (45,811,883 visits!)
- At every visit, one article was chosen uniformly at random from 20 articles ( $N=20$ ), and displayed in the Featured tab.
- $r_i(t) = 1$  if user clicked,  $r_i(t) = 0$  otherwise.
- $b_i(t) \in \mathbb{R}^{35}$ ,  $i = 1, \dots, 20$ .








- We applied the method of [Li et al., 2011] for **offline policy evaluation**.

Table 2: User clicks achieved by each algorithm over 10 runs

Policies	Mean	1st Q.	3rd Q.
Uniform policy	66696.7	66515.0	66832.8
TS algorithm	86907.0	85992.8	88551.3
Proposed TS	90689.7	90177.3	91166.3

Thank you !

# References I

-  Agrawal, S. and Goyal, N. (2013), “Thompson sampling for contextual bandits with linear payoffs,” *Proceedings of the 30th International Conference on Machine Learning*, 127–135.
-  Greenewald, K., Tewari, A., Murphy, S. and Klasnja, P. (2017), “Action centered contextual bandits,” *Advances in Neural Information Processing Systems*, 5977–5985.
-  Krishnamurthy, A., Wu, Z. S. and Syrgkanis, V. (2018), “Semiparametric contextual bandits,” *Proceedings of the 35th International Conference on Machine Learning*.
-  Lai, T.L. and Robbins, H. (1985), “Asymptotically efficient adaptive allocation rules,” *Advances in Applied Mathematics*, **6**(1), 4–22.
-  Li, L., Chu, W., Langford, J. and Wang, X. (2011), “Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms,” *Proceedings of the 4th ACM International Conference on Web search and data mining*, 297–306.
-  Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
-  Yahoo! Webscope. Yahoo! Front Page Today Module User Click Log Dataset, version 1.0. <http://webscope.sandbox.yahoo.com>. Accessed: 09/01/2019.