Model selection in contextual Bandits

Gopinath Balamurugan Sahasrajit Sarmasarkar

160010053 & 160070010
Fifth Year Dual Degree
Dept. of Electrical Engineering
IIT Bombay

September 3, 2024

Overview

- Contextual Bandits (2-3 slides max)
- Model Selection (2-3 slides max)
- Linear Contextual Bandit (1-2 slides max)
- Problem 1 : Pure Exploration
  - Problem statement (1-2 slides max)
  - Experiments (2-3 slides max)
- Problem 2 : Arms with different costs
  - Problem statement (1-2 slides max)
  - Experiments (2-3 slides max)
- References(1-2 slides max)

Contextual Bandits

For $t = 1, 2 \ldots T$:

Contextual Bandits

For $t = 1, 2 \ldots T$:

- Observe $x_t \in \mathcal{X}$.
- Take action $a_t \in \mathcal{A}$.

Contextual Bandits

For $t = 1, 2 \ldots T$:

- Observe $x_t \in \mathcal{X}$.
- Take action $a_t \in \mathcal{A}$.
- Incur loss $l_t(a_t) \in [0, 1]$.
- $\{x_t, l_t(.)\}$ are drawn i.i.d. from a fixed distribution D where $l_t : \mathcal{A} \to [0, 1]$

Contextual Bandits

For $t = 1, 2 \ldots T$:

- Observe $x_t \in \mathcal{X}$.
- Take action $a_t \in \mathcal{A}$.
- Incur loss $l_t(a_t) \in [0, 1]$.
- $\{x_t, l_t(.)\}$ are drawn i.i.d. from a fixed distribution D where $l_t : \mathcal{A} \to [0, 1]$

$$Regret(T) = \sum_{t=1}^{T} l_t(a_t) - \sum_{t=1}^{T} l_t(\pi^*(x_t))$$

Contextual Bandits

For $t = 1, 2 \ldots T$:

- Observe $x_t \in \mathcal{X}$.
- Take action $a_t \in \mathcal{A}$.
- Incur loss $l_t(a_t) \in [0, 1]$.
- $\{x_t, l_t(.)\}$ are drawn i.i.d. from a fixed distribution D where $l_t : \mathcal{A} \to [0, 1]$

$$Regret(T) = \sum_{t=1}^{T} l_t(a_t) - \sum_{t=1}^{T} l_t(\pi^*(x_t))$$

Where,

$$\pi^*(x) = \arg\min_{a \in \mathcal{A}} f^*(x, a)$$

And the goal is to minimize the regret.

Contextual Bandits

Key assumptions:

Contextual Bandits

Key assumptions:

- $\{x_t, l_t\}$ are drawn i.i.d. from a fixed distribution D.

Contextual Bandits

Key assumptions:

- $\{x_t, l_t\}$ are drawn i.i.d. from a fixed distribution D.
- $l_t \sim \mathcal{P}_{l|x}(.|x_t)$ independently given x_t; $l_t \in [0, 1]$

Contextual Bandits

Key assumptions:

- $\{x_t, l_t\}$ are drawn i.i.d. from a fixed distribution D.
- $l_t \sim \mathcal{P}_{l|x}(.|x_t)$ independently given x_t; $l_t \in [0, 1]$
- Given a class of $\mathcal{F}$ of value reward functions there exists $f^* \in \mathcal{F}$ such that $\mathbb{E}[l(a)|x] = f^*(x, a)$

Contextual Bandits

Key assumptions:

- $\{x_t, l_t\}$ are drawn i.i.d. from a fixed distribution D.
- $l_t \sim \mathcal{P}_{l|x}(.|x_t)$ independently given x_t; $l_t \in [0, 1]$
- Given a class of $\mathcal{F}$ of value reward functions there exists $f^* \in \mathcal{F}$ such that $\mathbb{E}[l(a)|x] = f^*(x, a)$

Related work:

- **SquareCB** [2] algorithm where,

$$Regret(T) \leq C.\sqrt{(KT)^{\frac{3}{2}}\sqrt{d}}$$

Model selection in Statistical learning theory

- $\{x_i, y_i\}_{i=1}^n$ are drawn i.i.d. from a fixed distribution D.
- Nested function class $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 ... \subseteq \mathcal{F}_M = \mathcal{F}$ where our goal is to find a $f \in \mathcal{F}$ that minimizes the $\mathcal{L}(f(x_i), y_i)$ with respect to a loss function $\mathcal{L}$.
- If f^*, the Bayes optimal predictor lies in $\mathcal{F}_{m^*}$(where $\mathcal{F}_{m^*}$ is the smallest class containing f^*) then we can find $\hat{f}_n$ such that,

$$\mathcal{R}(\hat{f}_n) \leq \mathcal{R}(f^*) + \sqrt{\frac{comp(\mathcal{F}_{m^*})}{n}.log(\frac{m^*}{\delta})}$$

w.p. 1 - δ.
and $\mathcal{R} : f \to [0, 1]$ is the function that computes the probability of misclassification ($\mathbb{E}_{\{x_i,y_i\}\sim\mathcal{D}}[\mathbb{1}_{(f(x_i)\neq y_i)}]$ by the function f .
Note 1: The complexity of the bound scales with m^* and not on M.
Note 2: Bayes optimal predictor(f^*) is the classifier that minimizes the probability of misclassification ($\mathbb{E}_{\{x_i,y_i\}\sim D}[\mathbb{1}_{(f(x_i)\neq y_i)}]$) [5]

Model selection

- $\{x_t, l_t\}$ are drawn i.i.d. from a fixed distribution D.
- Nested function class $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 ... \subseteq \mathcal{F}_M$.
- A set of policy classes nested as a sequence $\Pi_1 \subseteq \Pi_2 \subseteq \Pi_3 ... \subseteq \Pi_n = \Pi$.
- Each class $\Pi_m = \{\pi_f | f \in \mathcal{F}_m\}$, contains a set of policies π_f where $\pi_f(x) = \arg\min_{a \in A} f(x, a)$.
- The problem is realizable/well-specified in the sense that there exists index m^* (where $\mathcal{F}_{m^*}$ is the smallest class containing f^*). such that $\mathbb{E}[l(a)|x] = f^*(x, a)$, for all x, a.

The problem of model selection has been addressed in [1] in the case of linear contextual bandits where their algorithm scales with some function of the smallest class (m^*) containing the optimal function f^*.

Linear Contextual Bandit

For $t = 1, 2, ..., T$:

- Observe $x_t \in \mathcal{X}$.
- Take action $a_t \in \mathcal{A}$.
- Incur loss $l_t(a_t) \in [0, 1]$.

- Feature Maps: $\{\phi_m\}_{m \in [M]}$, $\phi_m : (\mathcal{X} x \mathcal{A}) \rightarrow \mathcal{R}^{d_m}$.
- Regression function: Note that $\mathcal{F}_m = \{(x, a) \rightarrow \langle \beta, \phi_m(x, a) \rangle | \beta \in \mathcal{R}_{d_m}\}$
- Realizability: $\exists \beta^* \in \mathcal{R}^{d_{m^*}}$, s.t. $E[l(a)|x] = \langle \beta^*, \phi_{m^*}(x, a) \rangle$

Note: $\forall\ m_1, m_2 \in [M]$ satisfying $m_1 < m_2$, the top m_1 elements in ϕ_{m_2} are precisely the elements of ϕ_{m_1}.

Some key assumptions and result

- $\phi_m(x, a) \sim \text{subG}(\tau_m^2)$ under $x \sim D$ $\forall a \in \mathcal{A}$.
- $l(a) - \mathbb{E}[l(a)|x] \sim \text{subG}(\sigma^2)$ $\forall a \in \mathcal{A}$ and $x \in \mathcal{X}$.
- $\lambda_{\min}\big(\sum_{a \in \mathcal{A}} \mathbb{E}_{x \in D}[\phi_m(x, a).\phi_m(x; a)^T]\big) > \gamma^2$ where $\lambda_{\min}(.)$ denotes the smallest eigen value.

Key result of the theorem

The mod-CB algorithm in [4] guarentees the following regret with probability at least $(1 - \delta)$

$$\text{Reg} \leq \tilde{O}\Big(\frac{\tau^4}{\gamma^3}(T.m^*)^{\frac{2}{3}}(K.d_{m^*})^{\frac{1}{3}}(\log\big(\frac{2}{\delta}\big)\Big)$$

Thus, the algorithm provides a regret which scales sublinearly with T and depends on the size of the "optimal" model class m^* instead of m.

Estimation of "gaps" between model classes

Loss function gap $\Delta_{i,j}$

- The loss function corresponding to a policy π is defined as the expected loss on choosing policy π i.e. $L(\pi) = \mathbb{E}_{(I,x)\sim D}[I(\pi(x))]$
- The loss function gap between model classes Π_i and Π_j is $\Delta_{i,j} = L_i^* - L_j^*$ where L_i^* denotes the optimal loss function in model class Π_i namely $L_i^* = \min_{\pi\in\Pi_i} L(\pi)$.

- Also note that $i, j \geq m^*$ would imply that $\Delta(i, j) = 0$
- The function *Estimate Residual(i,j)* in [4] estimates this gap between model classes Π_i and Π_j with high probability using the idea of square loss gap predictors..

As, we shall discuss later that this gap is of paramount importance in the design of the algorithm described below.

Idea of the algorithm proposed in [4]

- The algorithm maintains an active index class starting from index $\hat{m} = 1$.
- At each round, the algorithm runs a variant of the EXP algorithm namely EXP-IX algorithm on model class $\Pi_{\hat{m}}$ to decide to pull which arm. This was originally described in [3] which was used for obtaining high probability regret bounds for contextual bandits with a finite policy class.
- At each round the algorithm computes *Estimate Residual*$(\hat{m}, m)$ $\forall m > \hat{m}$.
- We update $\hat{m}$ to model class m only when the above evaluation crosses some pre-defined threshold i.e. we are sure that the optimal policy does not belong in $\Pi_{\hat{m}}$.

Problem 1 : Pure Exploration under a given horizon T

We now modify the algorithm described in [4] to the pure exploration framework where we aim to identify the optimal model class m^* with the lowest probability of error. We make two major changes in the algorithm.

- Instead of running EXP-IX as described above in every round, we pull the arms uniformly at random.
- We now evaluate the function *Estimate Residual(i,j)* at the end of all T rounds of the algorithm without changing the criterion for shifting from one model class to another.

Intuition

- We know typically that pure exploration strategies have much lower probability of error in identifying the best arm than the regret minimisation algorithms like EXP/UCB for stochastic non-contextual bandits.
- Also, we run the estimate residual algorithm at the end as we have much more samples to estimate the gap (between model classes) more accurately.

Simulation results

We simulate our modified and the original algorithm where the dimension of the model classes are given by 2,4,8,16,32,64,128,256,512 and 1000. Also for the optimal regression function f^*, only the top- s coefficients of β^* can be non-zero.
Note that the optimal model class m^* in this case is given by $\lceil \log_2(s) \rceil$

k	s	σ	Prob. mod. alg.(rounds)	Prob. orig. algorithm (rounds)
2	15	0	1(6000)	0(8000)
5	15	0	1(6000)	0(10000)
5	15	0.4	0.5(7000)	0(8000)
2	24	0	1(6000)	0.05 (8000)
5	83	0.15	1(15000)	0(15000)
5	100	0.25	0.75(20000)	0(20000)
5	100	0	1(10000)	0.2(20000)

Table: Empirical probability of correctness under original and modified algorithm

Simulation description and conclusion

- For estimating the probability of the algorithm detecting the model class m^* correctly, we perform multiple simulations. We divide the number of correct predictions with the total number of simulations to get the empirical estimate.
- Note that the last two columns in Tab. 1 denote the empirical probability under our modified algorithm and the original proposed algorithm respectively. Rounds basically denote the number of rounds for which each algorithm was run for.
- Clearly our modified algorithm estimates the correct model class with a far lower probability of error.

How does our algorithm perform in regret minimisation?

- Traditionally, algorithms which minimise the probability of error in identifying the best arm perform pretty badly in regret minimisation as it explores the sub-optimal arms too often.
- Interestingly, a similar thing can be said in the case of contextual bandits as well where the algorithm which identifies the best model class accurately performs poorly in regret minimisation (gives almost linear regret).
- However, the algorithm proposed in [4] performs much better in regret minimisation with sub-linear regret as discussed earlier.

Simulation results of regret for our algorithm

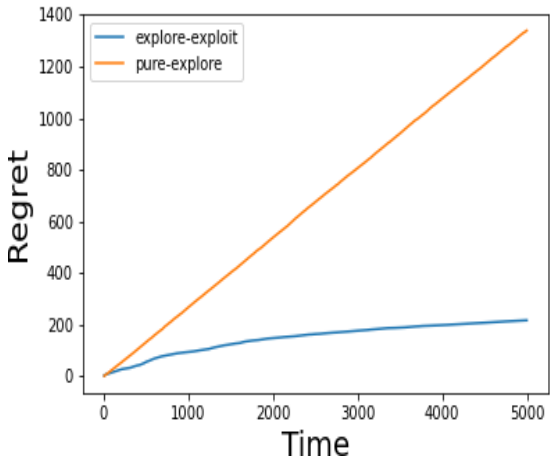

Figure: $T = 5000$,$s = 10$, $K = 2$ and $\sigma = 0$

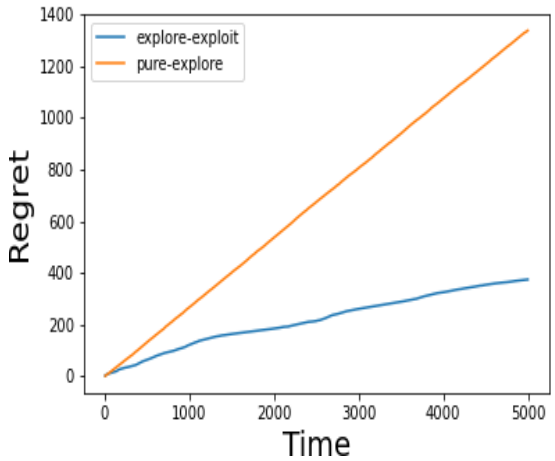

Figure: $T = 5000$,$s = 24$, $K = 2$ and $\sigma = 0.3$

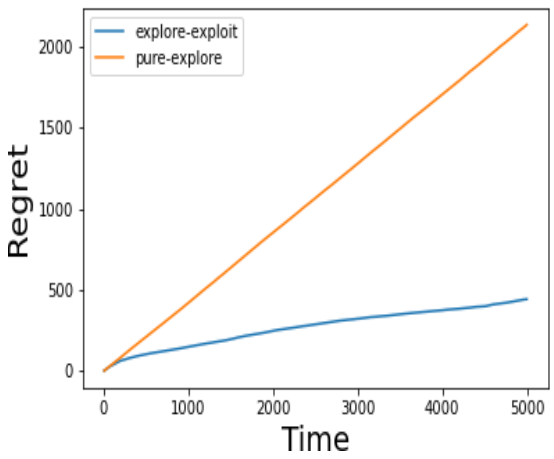

Figure: $T = 5000$,$s = 100$, $K = 5$ and $\sigma = 0$

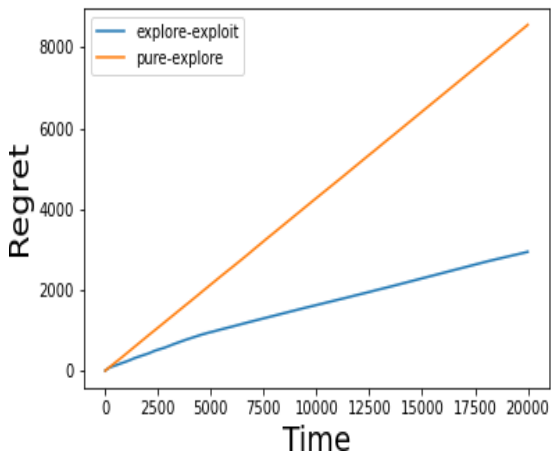

Figure: $T = 20000$,$s = 24$, $K = 5$ and $\sigma = 0.0$

Simulation results of regret for our algorithm

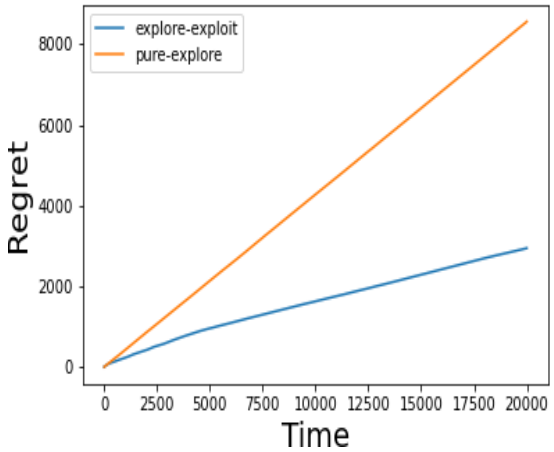

Figure: $T = 20000$,$s = 100$, $K = 5$ and $\sigma = 0.4$

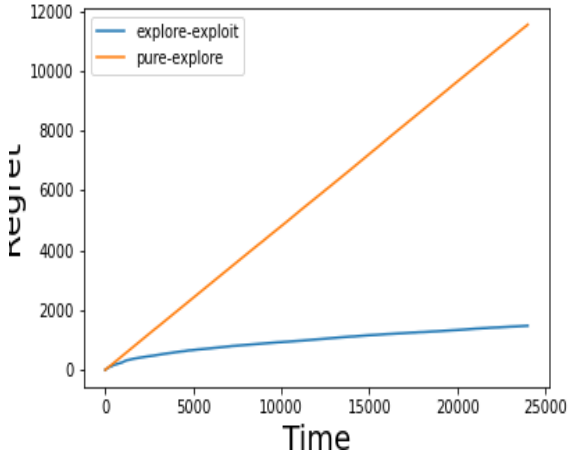

Figure: $T = 24000$,$s = 24$, $K = 10$ and $\sigma = 0.3$

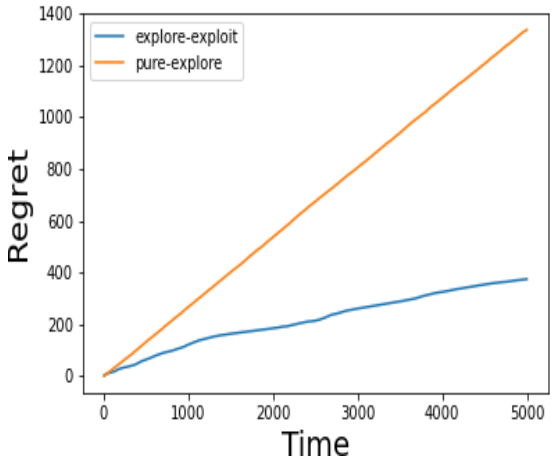

Figure: $T = 5000$,$s = 24$, $K = 2$ and $\sigma = 0.3$

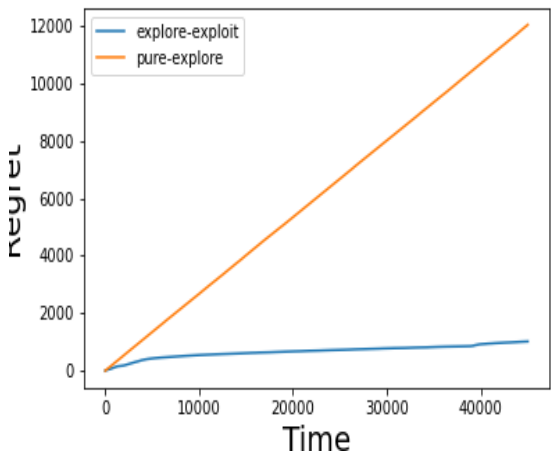

Figure: $T = 45000$,$s = 24$, $K = 2$ and $\sigma = 0.3$

Problem 2: Regret minimisation under a fixed budget and diff. arm cost

- Under this model, the cost of pulling each arm a may be different (denoted by c_a) and we are allocated a fixed cost budget S which we have to consume entirely.
- The aim in this setup is to design an algorithm which minimises the cumulative loss incurred till the entire cost budget is utilised.
- Note that under this set up, there is no restriction on the total number of pulls Ton the arms that an algorithm can make.

However, note that the assumptions on the loss functions l_t and contexts x_t remain the same.

Reference arm chosen for regret

- We believe that an optimal algorithm which has aprioiri knowledge of all the loss functions would pull the arm a which minimises $\frac{\mathbb{E}[l(a)|x]}{c_a}$ for every context x and thus we choose it as the reference arm.
- Mathematically, we write it as

$$Regret(S) = \sum_{t=1}^{T'} (l_t(a_t) - l_t(\pi^*(x_t))) \tag{1}$$

where $\pi^*(x) = \arg\min_{a \in \mathcal{A}} \frac{f'(x,a)}{c_a}$ and $\sum_{t=1}^{T'} c_{a_t} = S$.

Change in the proposed algorithm in [4]

We follow the algorithm mod-CB [4] after replacing $\phi_m(x, a)$ by $\frac{\phi(x,a)}{c_a}$.

We simulate our algorithm and compare with the original algorithm in the above setting in the upcoming slides. Note that we choose the costs of arms as [0.3, 0.7,1.5,2.6,8.9,9.7,7.8,3.5,4.0,5.3].

Simulation results

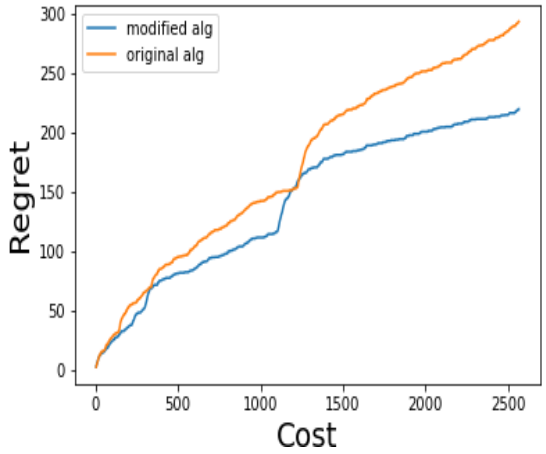

Figure: $T = 8000$,$s = 24$, $K = 2$ and $\sigma = 0.3$

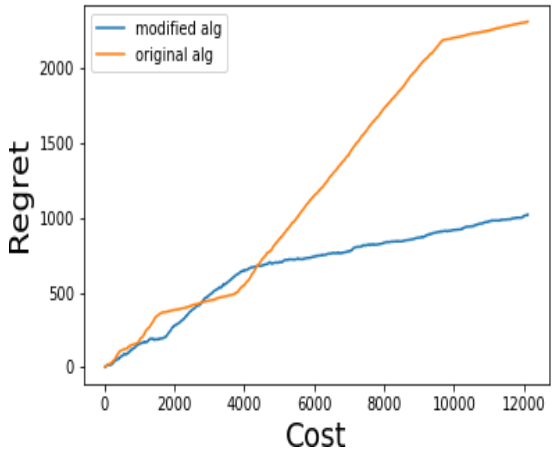

Figure: $T = 20000$,$s = 24$, $K = 5$ and $\sigma = 0.3$

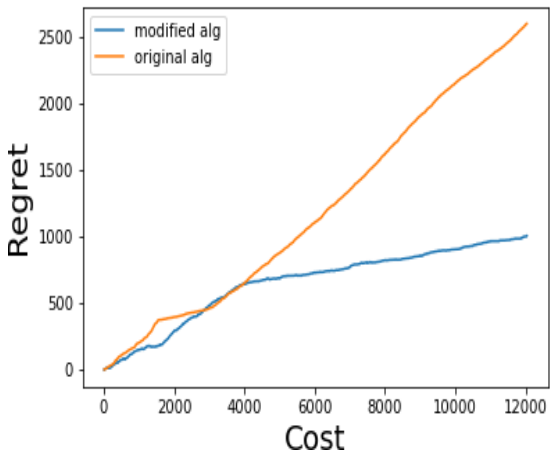

Figure: $T = 5000$,$s = 83$, $K = 5$ and $\sigma = 0.3$

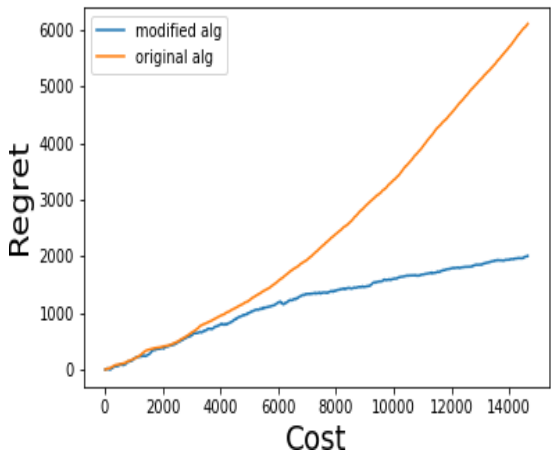

Figure: $T = 24000$,$s = 127$, $K = 5$ and $\sigma = 0.7$

References


Dylan J. Foster, Akshay Krishnamurthy, and Haipeng Luo.
Model selection for contextual bandits, 2019.


Dylan J. Foster and Alexander Rakhlin.
Beyond ucb: Optimal and efficient contextual bandits with regression oracles, 2020.


Gergely Neu.
Explore no more: Improved high-probability regret bounds for non-stochastic bandits.
In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.


Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari.
Model selection in contextual stochastic bandit problems, 2020.


Wikipedia.
Bayes optimal classifier, -.