Proceedings of CONF-APMM 2025 Symposium: Multi-Qubit Quantum Communication for Image Transmission over Error Prone Channels
DOI: 10.54254/2753-8818/2025.GL25881

MHNIW: A Formulated Multi-Agent Strategy to Solve High-Dimensional Multi-Arm Bandit Problem

Yixuan Hua

School of Mathematics and Statistics of Shandong University, Weihai, China
202200820055@mail.sdu.edu.cn

Abstract: This paper has designed an innovative method to solve the high-dimensional multi-arm bandit problems. Basic overview, this paper combines the Metropolis-Sampling method together with the Thompson Sampling methods and the algorithm is shown to be an anytime algorithm. Moreover, the unit of the size of the agents is the element-wise. In each training process, the agent will firstly communicate about the best result then do the choices after consideration which is indicated by the Metropolis-Hasting Sampling. Practically, the agents will be more willing to migrate to an arm with higher attraction so that all the agents will form a consensus of the best arm after some rounds. As a result, the cumulative regret of the mentioned algorithm is actually bounded by a logarithmic increment with respect to the gap between optimal arm and sub-optimal ones. Worth to mention, the upper bound of the cumulative regret will not present a significant increment relatively, if the aggregate number of the arms increases. As for the implementation, this paper has applied this paper's algorithm to the dataset provided by the Microsoft research which contains different sizes of data. The github source can be found at *https://github.com/whyulookingat/MHNIW*.

Keywords: Thompson Sampling, Metropolis-Hasting Sampling, High-Dimensional, Multi-Agents, Multi-Arm Bandit Problem

1. Introduction

In many real-world systems, the High-Dimensional Multi-Arm Bandit problems(HDMAB) which derive from the traditional Multi-Arm Bandit problems(MAB) and can be applied in assorted facets like recommendation platforms, sensor networks, and multi-robot coordination etc[1-2]. Inevitably, due to the sparsity of the related researches on the HDMAB and its derived version(i.e. Multi-Agent High-Dimensional Multi-Arm Bandit problems (HDMAMAB)) and the challenge to learn the reward distribution in multivariate situation, this paper have proposed a new effective model named as Metropolis-Hasting Normal-Inverse-Wishart abbreviated as MHNIW. The advantages of implementation of multi-agents strategy can be summarized as follow: 1. It can boost the exploration and exploitation conspicuously. 2. In large reward distribution, especially under high-dimension, it will alleviate the cases when single agent encounters the bottleneck 3.the multi-agent strategy will be more robust when the noises are intensive.

DOI: 10.54254/2753-8818/2025.GL25881

2. Related works

The Thompson Sampling(TS) is a basic algorithm with versatile variants[3-5]. But most of the problems are confined under the domain of single dimension.

While there are already some researches diving into the high-dimensional scenario by using structure based on TS combined with LASSO, the cumulative regret bound is still not satisfying enough and some of them devise some methods to deal with the high-dimensional contextual bandits[6-7].

Or some papers have suggested to use the group division trick to improve the TS and explore under the case that the agents have limited bit of communication[8-9].

Other than the TS methods to solve **HDMAB** problems. Many paper also attempt to use the variants of UCB and some innovative algorithms including OFUL under supervise learning and adaptive algorithm[10-12].

Whereas, this paper restarts from the vanilla TS method to both improve the efficiency and the performance so build this paper's model.

3. Problem formulation

To continue introduce this paper's **MHNIW** models, this paper should first establish some dozens of **Assumptions** and **Definitions.**

Assumption 1. *In this paper, the distribution always locates in somewhat hyper-rectangle(R) where*

$$R \triangleq \{[a_1, b_1], \ldots, [a_n, b_n]\} \tag{1}$$

Assumption 2. *In this paper, all the entries in the reward vectors are* ***Positive***

Assumption 3. *In this paper, this paper assume all the Symmetric Positive Definite(**SPD**) matrix satisfy that :*

$$\lambda_{max}(\Sigma) \leq \sigma^2 \exists \sigma \in \mathbb{R} \tag{2}$$

Definition 1. *The n-dimensional sub-gaussian distribution is well-defined by* $E[e^{u^\top X}] \leq e^{\frac{1}{2}u^\top \Sigma u}$

Definition 2. *If any entries in* $\Xi_{n\times n}$ *are smaller than corresponding elements in* $\Sigma_{n\times n}$*:* $\forall i, j \in n, \sigma_{ij} \in \Sigma_{n\times n} and \xi_{ij} \in \Xi_{n\times n}$ *if* $\xi_{ij} \leq \sigma_{ij}$ *then this paper says* Ξ *is dominated by* Σ *and denote as :*

$$\Xi_{n\times n} \lesssim \Sigma_{n\times n} \tag{3}$$

Definition 3. *The* $\frac{1}{E_i}$ *will be called of the* ***Attraction*** *if it satisfies:* $E_i = \| \mu\hat{}_{a_i} - \mu\hat{}_{a_{best}} \|$ *where* $\mu\hat{}_{a_i}$ *and* $\mu\hat{}_{a_{best}}$ *are the estimated mean reward one of which is played by agent i and one of which is the best arm in the current round respectively. The probability will be:*

$$\pi(E_i) = \frac{\frac{1}{E_i^k}}{\sum_{j=1}^{m} \frac{1}{E_j^k}} \propto \frac{1}{E_i^k} \; \forall k \geq 1 \tag{4}$$

Definition 4. *This paper measures the arm with best reward(i.e.* ***the best arm****) by*

$$\| \mu\hat{} \| \tag{5}$$

DOI: 10.54254/2753-8818/2025.GL25881

4. Methodology

4.1. Reward approximation

4.1.1. Overview

In this section, the current study adopts a case study approach. and inspired by the Thompson Sampling Method in 1-dimensional Multi-Arm Bandit Problems(**MAB**), this paper has formulated a hybrid structure containing Bayesian Methods. The method can perform effectively even the aggregate of arms is much more enormous.

4.1.2. Basis

Lemma 1. *Under the definition Def. 1, the sub-gaussian distribution is bounded by the Gaussian distribution w.r.t* Σ *(i.e.*$\mathcal{N}(0,\Sigma)$*)*

Proof. expand the term on both side of the inequality:

$$\mathbb{E}\left(\sum_{m=0}^{\infty}\frac{(u_1x_1+\cdots+u_nx_n)^m}{m!}\right)\leq\sum_{m=0}^{\infty}\frac{\sum_{i,j}\sigma_{i,j}u_iu_j}{2^m m!} \tag{6}$$

for any terms inside the addition:

$$\frac{(u_1x_1+\cdots+u_nx_n)^m}{m!} \tag{7}$$

and

$$\frac{\sum_{i,j}\sigma_{i,j}u_iu_j}{2^m m!} \tag{8}$$

expand the two terms and balance the terms respectively:

$$\sum_{i_1,\ldots,i_{2m}}u_{i_!}\ldots u_{i_{2m}}\mathbb{E}\left(X_{i_1}\ldots X_{i_{2m}}\right) \tag{9}$$

and

$$\frac{(2m)!}{2^m m!}\sum_{i_1,j_1,\ldots,i_m,j_m}\left(\prod_{k=1}^{m}\sigma_{i_kj_k}\right)u_{i_1}u_{j_1}\ldots u_{i_m}u_{j_m} \tag{10}$$

Interestingly, there is a trivial fact that the term $\frac{(2m)!}{2^m m!}=(2m-1)!!$ just stand for all the possible combination:$\left(u_{k_1},u_{k_2}\right),\ldots,\left(u_{k_{2m-1}},u_{k_{2m}}\right)$ and $k_1,\ldots,k_{2m}$ is all possible permutations of $u_1,\ldots,u_{2m}$. And compare the coefficient u_k on both two terms respectively, one can conclude that each different combination of the tuple will map to a same term. In other words:

$$\mathbb{E}\left(X_{i_1}\ldots X_{i_{2m}}\right)=\sum_{\pi\in\mathcal{P}_{2m}}\prod_{(a,b)\in\pi}\sigma_{i_ai_b} \tag{11}$$

where $\mathcal{P}_{2m}$ is

$$\mathcal{P}_{2m}\triangleq\left\{\left(u_{k_1},u_{k_2}\right),\ldots,\left(u_{k_{2m-1}},u_{k_{2m}}\right)\right\} \tag{12}$$

DOI: 10.54254/2753-8818/2025.GL25881

from Isserlis' Theorem it is obvious that $\mathbb{E}(X_{i_1}\dots X_{i_{2m}})$ is dominated by $\mathcal{M}^{2m}_{i_1\dots i_{2m}}(\mathcal{N}(0,\Sigma))$ where the M stands for the moments[13]. □

Lemma 2. *In a n-dimensional rectangle, if a certain probability measure the law of X:* $\mathbb{P}$ *is defined inside of that rectangle. In other words,* $\mathbb{P}(X \cap R) =^{a.s.} 1$ *where* $R \triangleq \{[a_1, b_1], \dots, [a_n, b_n]\}$. *Then, the* $max\mathbb{V}(X)$ *will follow a distribution* $\mathbb{P} \triangleq \{\mathcal{C}_i = \frac{1}{2^n} | i = 1,2, \dots, 2^n\}$ *where* $\mathcal{C}$ *is the vertices set of the hyper-rectangle*

Proof. **Start with the 1-dimensional case:**

For

$$\mathbb{P}(X \cap [a,b]) =^{a.s.} 1 \tag{13}$$

so $a \le E(X) \le b$ any fixed $E(X) = \mu(\mu \in [a,b))$ so all possible definition inside the $[a,b]$ can be expressed by the following class:

$$\mathcal{E} \triangleq \{\mathbb{P}: \mathbb{E}(X) = \mu, \forall \mu \in [a,b], \mathbb{P}(X \cap [a,b]) =^{a.s.} 1) \tag{14}$$

By the default definition of variance(i.e. $V(X)$), one can get:

$$\mathbb{V}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \int_{x\in[a,b]} (x-\mu)^2 d\mathbb{P} \tag{15}$$

$$= \int_{x\in[a,\mu)} (x-\mu)^2 d\mathbb{P} + \int_{x\in[\mu,b]} (x-\mu)^2 d\mathbb{P} \tag{16}$$

$$\le (a-\mu)^2 \int_{x\in[a,\mu)} d\mathbb{P} + \int_{x\in[\mu,b]} (b-\mu)^2 d\mathbb{P} \tag{17}$$

$$= (a-\mu)^2 \mathbb{P}([a,\mu]) + (b-\mu)^2 \mathbb{P}([\mu,b]) \tag{18}$$

$$= (a-\mu)^2 \omega + (b-\mu)^2 (1-\omega) \tag{19}$$

Thus, the distribution with the greatest variance must follow a two-endpoint distribution. Moreover, one can easily prove by take the $\frac{\partial \mathbb{V}(X,\omega)}{\partial \omega}$that with $\mathbb{P}(a) = \frac{1}{2}, \mathbb{P}(b) = \frac{1}{2}$, the variance will reach the maxima $\left(\frac{a-b}{2}\right)^2$.

As for n-dimensional case : this paper need to bound the $\mathbb{V}(X_i)$ $where$ $i = 1,2, \dots, n$, as for the entires other than the ones on the diagonal, this paper first supposes they are 0, since they do not have any linear relationship with each other. As for the update of these entries, this paper will discuss more specifically in later chapter.

Without losing generality, let's begin with:

$$\mathbb{V}(X_1) = \int_{X_1} \dots \int (X_1 - \mathbb{E}X_1)^2 d\mathbb{P} = \int_{X_1} (X_1 - \mathbb{E}X_1)^2 d\mathbb{P}_{X_1} \le \left(\frac{a_1 - b_1}{2}\right)^2 \tag{20}$$

where $\mathbb{P}_{X_1}$ is the marginal distribution w.r.t. X_1

Next,

$$\mathbb{V}(X_2 | \mathbb{P}_{max\mathbb{V}(X_1)}) = \int_{X_2} \dots \int (X_2 - \mathbb{E}X_2)^2 d\mathbb{P}_{max\mathbb{V}(X_1)} = \int_{X_2} (X_2 - EX_2)^2 d\mathbb{P}_{max\mathbb{V}(X_1),X_2} \tag{21}$$

$$\le \left(\frac{a_2 - b_2}{2}\right)^2 \tag{22}$$

DOI: 10.54254/2753-8818/2025.GL25881

where $\mathbb{P}_{max\mathbb{V}(X_1),X_2}$ is the marginal distribution w.r.t. X_2 Recursively, one will finally get that:

$$\mathbb{P}\left(X_n|\mathbb{P}_{\{max\mathbb{V}(X_t)\}_{t=1}^{n-1}}\right) = \int_{X_n}(X_n - \mathbb{E}X_n)^2 d\mathbb{P}_{\{max\mathbb{V}(X_t)\}_{t=1}^{n-1},X_n} \leq \left(\frac{a_n - b_n}{2}\right)^2 \tag{23}$$

In other words, one can conclude easily that:

$$\mathbb{P}(t_1, t_2, \dots, t_n) = \mathbb{P}(t_1)\mathbb{P}(t_2|t_1) \dots \mathbb{P}(t_n|t_1, \dots, t_{n-1}) = \left(\frac{1}{2}\right)^n \tag{24}$$

Inspired by the proof in Bandit Algorithm by Tor Lattimore and Csaba Szepesvári this paper have the following lemma[14].

Lemma 3. *For any sub-gaussian distribution well-defined in Def. 1and Assum. 1. The Σ_X is dominated by somewhat Σ*

Def. 2 $\Sigma_X \lesssim \Sigma$

Proof. By the **Taylor Series** one knows,

$$f(x + \Delta x) = f(x) + \frac{(\nabla^T \Delta x)^1}{1!} f'(x) + \frac{(\nabla^T \Delta x)^2}{2!} f''(x) + \cdots + \frac{(\nabla^T \Delta x)^k}{k!} f^{(k)}(x)$$

$$+ \frac{(\nabla^T \Delta x)^{k+1}}{(k+1)!} f^{(k+1)}(x + \theta x) \exists \theta \in (0,1) \tag{25}$$

now let $k = 1$, this paper will compare the entries in that Hessian matrix:

firstly, this paper denotes,

$$\psi_X(u^T) \triangleq ln\mathbb{E}e^{u^T X} \tag{26}$$

Then,

$$\frac{\partial^2 \psi_X(u^T)}{\partial^{k_1} u_1 \dots \partial^{k_n} u_n} = \begin{cases} \frac{\mathbb{E}x_i x_j e^{u^T X}}{\mathbb{E}e^{u^T X}} - \frac{\mathbb{E}x_i e^{u^T X}}{\mathbb{E}e^{u^T X}} \frac{\mathbb{E}x_j e^{u^T X}}{\mathbb{E}e^{u^T X}} (i \neq j) \\ \frac{\mathbb{E}x_i{}^2 e^{u^T X}}{\mathbb{E}e^{u^T X}} - \left(\frac{\mathbb{E}x_i e^{u^T X}}{\mathbb{E}e^{u^T X}}\right)^2 (i = j) \end{cases}$$

$$\left(\sum_{s=1}^n k_s = 2, k_t \in N, u = (u_1, u_2, \dots, u_n)^T\right) \tag{27}$$

Now, this paper introduces a brand new probability measure $\mathbb{P}_\lambda$, defined as:

$$\mathbb{P}_\lambda(dz) = e^{-\psi_Z(u^T)} e^{u^T Z} \mathbb{P}(dz) \tag{28}$$

so for any $\mathbb{E}_\lambda\left(z_1^{k_1} \dots z_n^{k_n}\right)$ $\left(\sum_{s=1}^n k_s = 2, k_t \in N, z = (z_1, z_2, \dots, z_n)^T\right)$, one can conclude as following:

$$\mathbb{E}_\lambda\left(z_i^{k_1} \dots z_n^{k_n}\right) = \int z_i^{k_1} \dots z_n^{k_n} \frac{1}{\mathbb{E}e^{u^T Z}} e^{u^T Z} \mathbb{P}(dz) \tag{29}$$

and

DOI: 10.54254/2753-8818/2025.GL25881

$$\mathbb{E}_\lambda((z_1 - \mathbb{E}z_1)^{k_1} \dots (z_n - \mathbb{E}z_n)^{k_n})$$
$$= \{\frac{\mathbb{E}z_i z_j e^{u^T X}}{\mathbb{E}e^{u^T Z}} - \frac{\mathbb{E}z_i e^{u^T Z}}{\mathbb{E}e^{u^T Z}} \frac{\mathbb{E}z_j e^{u^T Z}}{\mathbb{E}e^{u^T Z}} = Cov_\lambda(z_i, z_j)(i \neq j)$$

$$\frac{\mathbb{E}{z_i}^2 e^{u^T Z}}{\mathbb{E}e^{u^T Z}} - \left(\frac{\mathbb{E}z_i e^{u^T Z}}{\mathbb{E}e^{u^T Z}}\right)^2 = \mathbb{V}_\lambda(z_i)(i = j) \tag{30}$$

So

$$\frac{\partial^2 \psi_X(u^T)}{\partial^{k_1} u_1 \dots \partial^{k_n} u_n} = \mathbb{E}_\lambda((z_1 - \mathbb{E}z_1)^{k_1} \dots (z_n - \mathbb{E}z_n)^{k_n}) \tag{31}$$

It indicates one can use $\mathbb{P}_\lambda$'s covariance matrix to bound the Σ_X

And since this paper has proved that the most sparse distribution from Lem. 2. Moreover, one can affirm that for

$$\mathbb{P} \triangleq \{\mathcal{C}_i = \frac{1}{2^n} | i = 1,2, \dots, 2^n\} \tag{32}$$

and $\forall i, j \in n, Cov(x_i, x_j) = \mathbb{E}x_i x_j - \mathbb{E}x_i \mathbb{E}x_j = 0$ and $\mathbb{P}_{X_t}(a_t) = \mathbb{P}_{X_t}(b_t) = \frac{1}{2}$ $(t \in \{1,2, \dots, n\})$. From Lem. 1 and Def. 1,It is bounded by a certain Gaussian Distribution. So now this paper can get the following Theorem.

4.1.3. Main part

Theorem 1. *As for the Muli-Agent Multi-Arm Bandit Problem(**MAMAB**).We can use the conjugate Normal-Inverse-Wishart distribution(NIW) to update the parameter and reach to the arm with best mean reward. Specifically,*

Prior Distribution

$$\mu | \Sigma \sim \mathcal{N}\left(\mu_0, \frac{1}{\kappa_0}\Sigma\right) \tag{33}$$

$$\Sigma \sim \mathcal{IW}(\Psi_0, \nu_0) \tag{34}$$

Posterior Distribution

$$\mu | \Sigma, D \sim \mathcal{N}\left(\mu_n, \frac{1}{\kappa_n}\Sigma\right) \tag{35}$$

$$\Sigma | D \sim \mathcal{IW}(\Psi_n, \nu_n) \tag{36}$$

where,

$$D \triangleq \{X = (X_1, X_2, \dots, X_n), X_i \sim^{i.i.d.} \mathcal{N}(\mu, \Sigma)) \tag{37}$$

$$\mu_n = \frac{nX' + \kappa_0 \mu_0}{\kappa_0 + n} \tag{38}$$

$$\kappa_n = \kappa_0 + n \tag{39}$$

DOI: 10.54254/2753-8818/2025.GL25881

$$\Psi_n = \Psi_0 + S + \frac{\kappa_0 n}{\kappa_n}(X' - \mu_0)(X' - \mu_0)^T \tag{40}$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(X_i - X')(X_i - X')^T \tag{41}$$

$$\Psi_0 = \begin{pmatrix} \left(\frac{a_1 - b_1}{2}\right)^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \left(\frac{a_n - b_n}{2}\right)^2 \end{pmatrix} \tag{42}$$

Proof. In brief, this paper only proves why to set the scale matrix as shown above. From Lem. 1 and Lem. 3.As one does not actually acquire the linear relationship inside the covariance matrix, this paper assumes they are all **independent**(i.e. 0) at first. Consequently, from Lem. 2 one can summarize that $\Sigma_X \lesssim \Sigma where N(\mu, \Sigma)$.We can conclude very simply that Σ_X is well-bounded by the Ψ_0. Thus, this paper takes the Ψ_0 as the initial scale matrix of Inverse-Wishart distribution.Additionally, by drawing the new observations continuously from the real distribution, elements other than the ones on the diagonal will get updated, meaning the model will start to learn the relationship between different entries. □

The reason why to choose the **NIW** as the update model because this paper can not conclude the linear relationship between different entries(i.e. $Conv(x_i, x_j)$). Consequently, this paper need to activate the Σ as a random variable.

4.2. Agent strategy

4.2.1. Overview

In this part, this paper need to discuss the strategy used and shared by different agents, unlike the strategy used by which focus on the theme with limited communication between agents.[9] Inspired by the Metropolis-Hasting Sampling methods(i.e. **M-H Sampling**), this paper has proposed a division to separate them agent-wise which will guarantee they will congregate into the appropriate(best) arms Def. 4 even when the data is very enormous[15]. And as a result, the efficiency reaches a pretty good level compared with REF.

4.2.2. Main part

As for the main part, The Metropolis-Hasting Sampling(**M-H Sampling**) is frequently used to proximate the real distribution of somewhat random variables by calculating the entries inside the transition matrices[15]. But this paper has applied some modification on the original definition so that this paper could congregate the agents to the best arms after some rounds.

Theorem 2. this paper accepts the same notations in the Def. 3: for any agents i and j, they will follow the migration below:

1.Compute E_i and E_j respectively

2.Let Δ be the $\frac{\frac{1}{E_j^k}}{\frac{1}{E_i^k}} \forall k \geq 1$

3.Draw $v \sim U[0,1]$

4.If $v \leq min(\Delta, 1)$, agent i will play the arm which agent j is playing

5.If $v > min(\Delta, 1)$, agent i will reject the migration towards agent j

DOI: 10.54254/2753-8818/2025.GL25881

Analysis 1. *Because one knows the less the E_i for some agent i is, the more likely that arm will be the best arm. Thus, the Δ will be greater, the more likely the agent i will migrate to the arm which agent j is playing. In the end, most of the agents will play the arm with the best mean reward. Figure 1 is the embodiment of the strategy.*

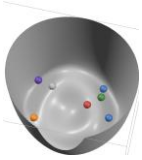

Figure 1: The agents willing to scroll down to a place with lower altitude. Picture credit: Original

4.3. Algorithm

4.3.1. Overview

Based on the two $Theorems$ this paper has mentioned in Chapter A and Chapter B (Theo. 1, Theo. 2). We have the following algorithhm. It combines the **M-H Sampling** and **Thompson Sampling** altogether.

4.3.2. Main part

Algorithm **MHNIW** proceeds as :
Algorithms 1.

Begin
Dataset D, *number of agents* n, *dimension* d, *horizon* T, *variance* σ^2
Averaged cumulative regret, arm selection matrix
Initialize $\mu_0 \leftarrow 0_d$, $\kappa_0 \leftarrow 1$, $\nu_0 \leftarrow d+2$, $k > 1$ Compute Ψ_0 as diagonal range variance of dataset Initialize buffers $B_a \leftarrow \emptyset$ for each arm a Initialize estimated means $\mu\hat{}_a \leftarrow 0_d$ Compute true means $\mu_a^\star \leftarrow mean(D_a)$
Agent 1 pulls all arms once and stores samples in corresponding B_a
Randomly assign agent i an arm a_i Sample one reward from a_i and store in B_{a_i}
For i to n do
  Randomly assign agent i an arm a_i
  Sample one reward from a_i and store in B_{a_i}
  Update posterior estimates $\mu\hat{}_a$ for all a using NIW posterior sampling
For t=1 to T do
  Identify best arm $a\hat{}^\star = argmax_a \| \mu\hat{}_a \|$
  Count popularity of current arms, pick $a\hat{}^\star$ as a_{best} Randomly sample arm a_j
  For s=1 to S=⌈log(n)⌉
    For agent i in random order do
      Randomly sample arm a_j
      Compute $E_i = \| \mu\hat{}_{a_i} - \mu\hat{}_{a_{best}} \|$
      Compute $E_j = \| \mu\hat{}_{a_j} - \mu\hat{}_{a_{best}} \|$

DOI: 10.54254/2753-8818/2025.GL25881

$$\Delta \leftarrow \frac{\frac{1}{E_j^k}}{\frac{1}{E_i^k}}$$

If Uniform(0, 1) < min(1, Δ)

$a_i \leftarrow a_j$

End if

End for

End for

▷ Proceed to Algorithm 1 extra for reward sampling and regret evaluation Averaged cumulative regret

End for

End.

Algorithms 1 extra part:

Begin

For agent i=1 to n do

Sample reward x from a_i and store in B_{a_i}

Record arm choice

Update all $\mu\hat{}_a$ using current buffers and NIW posterior sampling

For agent i=1 to n do

Compute regret as $\| \mu\hat{}_{a_i} - \hat{\mu_{a^*}} \|$

End for

End for

Accumulate average regret across agents

End.

5. Experiments

5.1. Overview

In this section, this paper has conducted some experiments to analyze and compare the different algorithms quantitatively, this paper compares this paper's algorithm with **MATS** and **MH-Empirical** algorithms. And it seems like this paper's algorithm seems to work pretty well in the domain of **MAMAB**(Multi-Agent Multi-Arm Bandit Problem).

5.2. Preparation

Based on the fact that there are few works that focus on the problems when the reward distribution is **High Dimensional**(HD). Thus, the paper has made some inevitable modifications to apply the existing algorithms onto the **HD-MAMAB** problems. The following algorithms are the adapted versions:

5.2.1. HD-MATS

Algorithms 2.

Begin

DOI: 10.54254/2753-8818/2025.GL25881

Prior NIW parameters $(\mu_0, \kappa_0, \Psi_0, \nu_0)$ for each agent e, data dimension d
$H_0 \leftarrow \{\}$
For t=1 to T do
 For e=1 to ρ do
  For $a^e \in \mathcal{A}^e$
   Sample $\mu_t^e(a^e)$ from posterior NIW $\sim NIW(\mu_0, \kappa_0, \Psi_0, \nu_0)$
  End for
End for
$a_t \leftarrow argmax_{a \in A} \sum_{e=1}^{\rho} \| \mu_t^e(a^e) \|$
Pull joint arm a_t, observe vector rewards $\{f_t^e(a_t^e)\}_{e=1}^{\rho}$ $H_t \leftarrow H_{t-1} \cup \{(a_t, \{f_t^e(a_t^e)\}_{e=1}^{\rho})\}$
Update NIW posterior for each e based on H_t
End for
End.

5.2.2. MH-empirical

Algorithms 3.

Begin
Continue the same procedures in Part 1 of Alg. 1
For agent i=1 to n do
 Sample reward x from a_i and store in B_{a_i}
Record arm choice
End for
For each arm a do
Compute empirical mean $\hat{\mu}_a = \frac{1}{|B_a|} \sum_{x \in B_a} x$
End for
For agent i=1 to n do
Compute regret as $\|\hat{\mu}_{a_i} - \hat{\mu}_{a^\star}\|$
End for
Accumulate average regret across agents
End.

5.3. Result

By applying this paper's algorithms onto the dataset provided by *Microsoft Research*, this paper has got the following graphs with respect to different algorithms[16]. In this experiment, this paper only use the first two training datasets. Specifically, the row data in that file has such format as follow:

0 qid:1 1:3 2:0 3:2 4:2 … 135:0 136:0

2 qid:1 1:3 2:3 3:0 4:0 … 135:0 136:0

This paper use the qid as the arm index and the rest of dictionary as the reward which is presented as the marginal index : marginal reward the following. The following part relate to MHNIW:

DOI: 10.54254/2753-8818/2025.GL25881

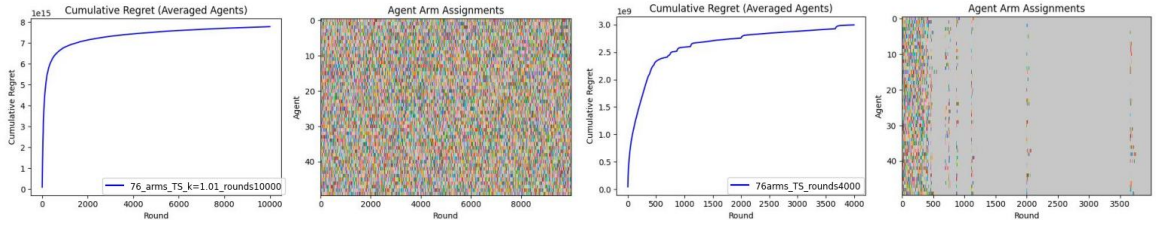

Figure 2: MHNIW on 76-Arms Picture credit: Original

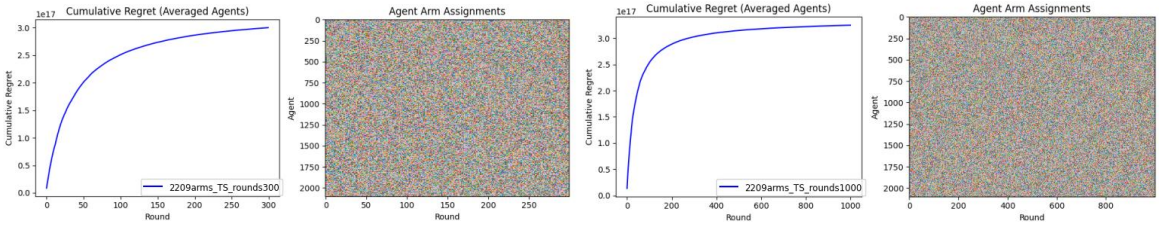

Figure 3: MHNIW on 2209-Arms Picture credit: Original

The following part relate to **MH-Empirical**

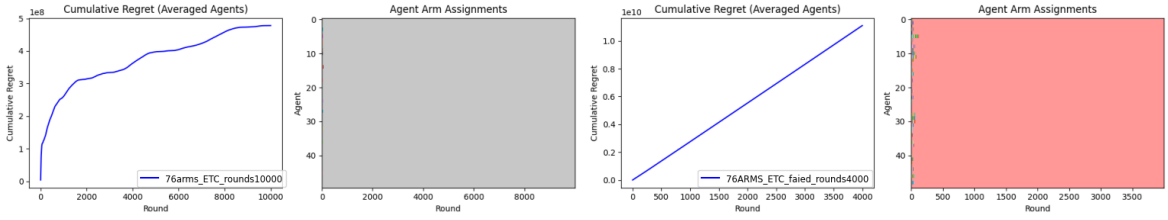

Figure 4: MH-Empirical on 76-Arms Picture credit: Original

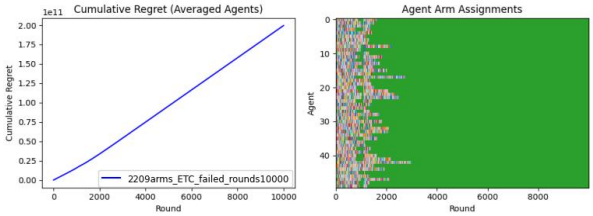

Figure 5: MH-Empirical on 2209-Arms Picture credit: Original

The following part relate to **HD-MATS**

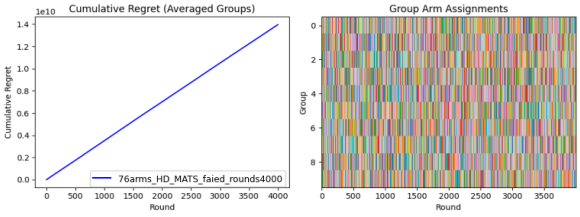

Figure 6: HD-MATS on 76-Arms Picture credit: Original

DOI: 10.54254/2753-8818/2025.GL25881

6. Results and discussion

6.1. Statement

Firstly, to facilitate this paper's comparison over different algorithms and datasets. this paper will refer the pictures by their titles and state that: Figure 2 and Figure 3 are assigned to the MHNIW algorithm and Figure 4 and Figure 5 are assigned to the MH-Empirical algorithm and Figure 6 is assigned to the HD-MATS algorithms.

6.2. Comparison

MHNIW performs well regardless the number of the arm. This is due to the fact that the **M-H Sampling** consistently arranges the **Exploration** and **Exploitation** by a moderate scale. As shown in Figure 2 and Figure 3, the regret still follows a logarithmic increment even if the number of arms has an exponential increment like the case in Figure 3, which is also proved by rigorous analysis shown in Theo. 3.

As for MH-Empirical, under the scenario with less arms(i.e. 76-Arms or Figure 4) the algorithm can find and confirm the best arm more effectively, but its accuracy is still under the MHNIW's. And as for the case with more arms (i.e. Figure 5), the algorithm cannot find the best arm after 10 experiments each containing 10,000 runs. Maybe some of the careful readers have noticed the values of the regret of MH-Empirical's are much smaller than the MHNIW's, but based on the poor accuracy fact and the fact that MH-Empirical will likely stick on the sub-optimal arms thus it will have a **Linear Regret**, the regret of MH-Empirical will exceed the one of MHNIW after **sufficient** runs.

As for HD-MATS(i.e. Figure 6), originally proposed by the Timothy Verstraeten et al, the paper firstly conducts this algorithm to the scenario of 76-Arms.Specifically, as recorded in the vanilla paper, the agents are always be divided into several **groups**[8]. Thus, this paper separates the 50 agents into 10 groups evenly. It performs poorly on the **HDMAB** task after this paper run the experiment up to 4,000 rounds. Moreover, the growth of the regret still be **Linear**. So, HD-MATS will undoubtedly perform less satisfying on 2209-Arms.Theoretically, this phenomenon can be explained by the bound of the regret that HD-MATS is bounded by $\mathcal{O}\left(\sqrt{Tlog(T)}\right)$ but A is bounded $\mathcal{O}\left(log(T)\right)$(Theo. 3) briefly[8].

7. Conclusion

7.1. Main part

Based on the work this paper has already discussed in all previous chapters and has successfully built the **MHNIW** model (alg: 1) and witness the efficiency regardless of the size of the data. The algorithm always settles at the best arm within 1,500 rounds which is a significant hop. By combining the idea of the multi-agents, the algorithm is not trapped by the obstacles same as the single-agent strategy like the bottleneck problems and efficiency problems. From the panorama, the **MHNIW** can be capable of the **HDMAB** tasks.

7.2. Future work

Some improvements still need to be clarified:

1.Optimizing the details of the **MHNIW**

DOI: 10.54254/2753-8818/2025.GL25881

2.In the real practice, whether one can take the power of the **Attraction** Def. 3 as some positive real value smaller than 1 (i.e. $0 < k < 1$) when this paper need to set $k \geq 1$ theoretically

Appendix

A. Overview

In this section, this paper is going to discuss the regret bound over the algorithm MHNIW(alg: 1), let us recall the **regret decomposition formula** at first, given as:

$$R(T) = \sum_{t=1}^{T}\left(\mu_{a^\star} - \mathbb{E}X_{a_t}\right) = \sum_{a=1}^{K} \Delta_a \cdot \mathbb{E}[N_a(T)] = \sum_{a=1}^{K} \Delta_a \cdot \mathbb{E}[N_a^{init}(T) + N_a^{alloc}(T)] \quad (43)$$

where,$N_a^{init}(T)$and$N_a^{alloc}(T)$are the times the agents exploring those arms at the initialization phase and the M-H Samplings phase respectively. And the regret shows that it will have logarithmic increment.

B. Lemma

Lemma 4. *With the previous NIW method, the estimated parameters will have the following bounding terms:*

$$\mathbb{P}\left(\| \hat{\mu}_a - \mu_a \| \leq c\sqrt{\frac{dlog\left(\frac{1}{\delta}\right)}{m}}\right) \geq 1 - \delta^{\frac{1}{T}} \quad (44)$$

and

$$\mathbb{E}[\| \hat{\mu}_a - \mu_a \|^2] \leq \frac{c\prime d}{m} \quad (45)$$

where, c, c' and d are the constants and m is the times one selects the arm A by T rounds.

Proof. Without losing generality, if a random variable $Z \sim \mathcal{N}(0, \Sigma)$. Then, for the given series $Z = (Z_1, Z_2, \dots, Z_n)$, one will get:

$$\hat{\mu}_Z = \frac{1}{n}\sum_{i=1}^{n} Z_i \sim \mathcal{N}\left(0, \frac{1}{n}\Sigma\right) \quad (46)$$

$$\mathbb{E} \| \hat{\mu}_Z \|^2 = \frac{1}{n} Tr(\Sigma) \quad (47)$$

Thus if $X \sim \mathcal{N}(\mu, \Sigma)$, then,

$$\mathbb{E} \| \hat{\mu}_X - \mu \|^2 = \frac{1}{n} Tr(\Sigma) \quad (48)$$

Now if this paper has $Z \sim \mathcal{N}(0, I_d)$ and $A \in \mathbb{R}^{d\times d}$ is **SPD**. Then the **Hanson–Wright inequality** states[17]:

$$\mathbb{P}(|Z^\top AZ - E[Z^\top AZ]| \geq t) \leq 2exp\left(-c \cdot min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\right)\right) \quad (49)$$

DOI: 10.54254/2753-8818/2025.GL25881

Apply this with $A = \Sigma, X = \Sigma^{\frac{1}{2}}Z$ and one can easily get that:

$$Z^\top \Sigma Z = \| X \|^2, \mathbb{E}[Z^\top \Sigma Z] = Tr(\Sigma) \tag{50}$$

And given definition :

$$\| A \| = \| \Sigma \| = \lambda_{max}(\Sigma) \tag{51}$$

$$\| A \|_F^2 = \| \Sigma \|_F^2 = \sum_{i,j} \Sigma_{ij}^2 \tag{52}$$

Substitute the variable in the **Hanson–Wright inequality**, one gets:

$$\mathbb{P}(\| X \|^2 \geq Tr(\Sigma) + t) \leq 2exp\left(-c \cdot min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\right)\right) \tag{53}$$

Choose $t = \| \Sigma \|_F \sqrt{log\left(\frac{2}{\delta}\right)} + \| \Sigma \| log\left(\frac{2}{\delta}\right)$ this paper gets :

$$\mathbb{P}\left(\| X \|^2 \geq Tr(\Sigma) + \| \Sigma \|_F \sqrt{log\left(\frac{2}{\delta}\right)} + \| \Sigma \| log\left(\frac{2}{\delta}\right)\right) \leq \delta \tag{54}$$

Plus, this paper has these trivial facts based on Assum. 3 and denote d as the rank of the matrix, if $\Sigma \preccurlyeq \sigma^2 Ii.e.(\| \Sigma \| \leq \sigma^2)$:

$$\{Tr(\Sigma) \leq \sigma^2 d \ \ \| \Sigma \|_F^2 \leq \sigma^4 d \ \ \| \Sigma \| \leq \sigma^2 \tag{55}$$

Substitute the variables into (54), one can conclude :

$$\| X \|^2 \leq \sigma^2 d + c \cdot \left(\sigma^2 \sqrt{dlog\left(\frac{2}{\delta}\right)} + \sigma^2 log\left(\frac{2}{\delta}\right)\right) \to \sigma^2 \left(d + log\left(\frac{2}{\delta}\right)\right) \tag{56}$$

The second limit and following inequality will work if one takes the δ sufficiently small. Eventually,

$$\mathbb{P}(\| X \| \leq \sigma \sqrt{dlog\left(\frac{1}{\delta}\right)}) \geq 1 - \delta \tag{57}$$

And thus one can get,

$$\mathbb{P}(\| X \| \leq \sigma \sqrt{\frac{dlog\left(\frac{1}{\delta}\right)}{m}}) \geq \mathbb{P}(\| X \| \leq \sigma \sqrt{\frac{dlog\left(\frac{1}{\delta}\right)}{T}}) \geq 1 - \delta^{\frac{1}{T}} \tag{58}$$

The Lemma is proved by taking some constant c

Lemma 5. *The regret $\mathcal{R}_{alloc} \triangleq \sum_{a=1}^{K} \Delta_a \cdot N_a^{alloc}(T)$ will be bounded by the increment of*

$$\mathcal{O}\left(dlogT \sum_{a \neq a^\star} \frac{1}{\Delta_a}\right) \tag{59}$$

when $T \to \infty$

DOI: 10.54254/2753-8818/2025.GL25881

Proof. Firstly, one can summarize that : if

$$\| \mu\hat{}_a - \mu_a \| \leq \frac{\Delta_a}{h} \text{ W.P . } \geq 1 - \delta^{\frac{1}{T}} \ \forall h \geq \frac{2\Delta^*}{\|\mu_{a^*}\| - \|\mu_a\|} \tag{60}$$

denote $\Delta^* = max_a \Delta_a$

By the Lem. 4(44), and take $\delta = \frac{1}{T^{2T}}$,

$$\sqrt{\frac{2cdlog(T)}{m_a}} \leq \frac{\Delta_a}{h} \tag{61}$$

Then,

$$m_a \geq \frac{2h^2 cdlog(T)}{\Delta_a^2} \tag{62}$$

Reasonably, this paper could define the **Stopping Time** as follow:

$$\tau_a = min\{t: m_a(t) \geq \frac{2h^2 cdlogT}{\Delta_a^2} - 1\} \tag{63}$$

Where $m_a(t)$ is the times the arm a played by the round t . Thus, for any $t \leq \tau_a$, one has :

$$\mathcal{R}_{alloc}^{t \leq \tau_a} \leq \frac{2h^2 cdlog(T)}{\Delta_a} \tag{64}$$

For any $t > \tau_a$:

Since $T \to \infty$ and this paper need to bound the terms of the times the agents to choose the arms that $a \neq a^*$.Firstly, one gets:

$$\mathbb{P}_{A_t}(a \neq a^*) \leq \mathbb{P}\left(\| \mu\hat{}_a - \mu_a \| \geq \frac{\Delta_a}{h} \cup \| \mu\hat{}_{a^*} - \mu_{a^*} \| \geq \frac{\Delta_a}{h}\right) \tag{65}$$

$$\leq \mathbb{P}\left(\| \mu\hat{}_a - \mu_a \| \geq \frac{\Delta_a}{h}\right) + \mathbb{P}\left(\| \mu\hat{}_{a^*} - \mu_{a^*} \| \geq \frac{\Delta_a}{h}\right) \tag{66}$$

This is because the obvious fact that if $\| \mu\hat{}_a - \mu_a \| < \frac{\Delta_a}{h}$ and $\| \mu\hat{}_{a^*} - \mu_{a^*} \| < \frac{\Delta_a}{h}$ for any $h \geq \frac{2\Delta^*}{\|\mu_{a^*}\| - \|\mu_a\|}$ one can easily conclude

$$\| \mu\hat{}_a \| < \| \mu_a \| + \frac{\Delta_a}{h} \leq \| \mu_{a^*} \| - \frac{\Delta_a}{h} < \| \mu\hat{}_{a^*} \| \tag{67}$$

So, the algorithm will likely to take arm a^* as the best arm. So by taking the inverse-negate event this paper has the previous inequality.

By **Concentration Inequality** for $\forall \epsilon > 0$

$$\mathbb{P}\left(\| X \| \geq \sqrt{\mathbb{E} \| X \|^2} + \epsilon\right) \leq e^{-\frac{\epsilon^2}{2\|\Sigma\|}} \tag{68}$$

DOI: 10.54254/2753-8818/2025.GL25881

One can easily get the following result by take the Assum. 3 [17]

$$\mathbb{P}\left(\| \hat{\mu}_a - \mu_a \| \geq \sqrt{\mathbb{E}\| \hat{\mu}_a - \mu_a \|^2} + \epsilon\right) \leq e^{-\frac{\epsilon^2}{2\|\Sigma\|}} \tag{69}$$

Thus, if this paper denote d as the rank of the matrix , this paper will get,

$$\mathbb{P}\left(\| \hat{\mu}_a - \mu_a \| \geq \sqrt{\frac{d\sigma^2}{m_a}} + \epsilon\right) \leq e^{-\frac{m_a\epsilon^2}{2\sigma^2}} \tag{70}$$

Since, $\mathbb{E}\| X \|^2 = \frac{Tr(\Sigma)}{n} \leq \frac{d\sigma^2}{n}$ where d is the rank of the matrix and n is the sample size(Lem. 4(45))

And take the $\epsilon = \frac{\Delta_a}{h}$.Plus, $\frac{\Delta_a}{h} \gg \sqrt{\frac{d\sigma^2}{m_a}}$ because of $T \to \infty$, one can summarize the following undoubtedly by choosing somewhat constant c that is greater than h:

$$\mathbb{P}\left(\| \hat{\mu}_a - \mu_a \| \geq \frac{\Delta_a}{h}\right) \leq e^{-\frac{m_a\epsilon^2}{2\sigma^2}} \leq e^{-\frac{c^2 log(t^2)}{h^2}} \leq \frac{C}{t^2} \tag{71}$$

Similarly, one can also get $\mathbb{P}\left(\| \hat{\mu}_{a^*} - \mu_{a^*} \| \geq \frac{\Delta_a}{h}\right) \leq \frac{C}{t^2}$

Consequently, one can get the following upper bound:

$$\mathcal{R}_{alloc}^{t>\tau_a} = \sum_{a\neq a^*} \Delta_a \mathbb{E}[N_a(T)] = \sum_{a\neq a^*} \sum_{t>\tau_a} \Delta_a \mathbb{E}[\mathbb{I}\{A_t = a\}] \tag{72}$$

$$= \sum_{t>\tau_a} \Delta_a \mathbb{P}_{A_t}[a \neq a^*] \leq 2C\Delta^* \sum_{t=1}^{\infty} \frac{1}{t^2} \leq \frac{C\Delta^*\pi^2}{3} = \mathcal{O}(1) \tag{73}$$

Thus the $\mathcal{R}_{alloc} = \mathcal{O}\left(dlogT \sum_{a\neq a^*} \frac{1}{\Delta_a}\right)$

Lemma 6. *In the Initialization Phase, the regret is bounded by:*

$$\mathcal{R}_{init} \leq \Delta^*(A + K - 1) \tag{74}$$

where, A is the number of arms and K is the number of agents and $\Delta^* = max_a \Delta_a$

Proof. Trivial

C. Regret analysis

Based on the previous chapter, one can conclude that :

Theorem 3. *The **MHNIW** regret is bounded by :*

$$lim_{T\to\infty} \mathcal{R} \leq \mathcal{O}\left(dlogT \sum_{a\neq a^*} \frac{1}{\Delta_a}\right) \tag{75}$$

Proof. From Lem. 5 the R has already satisfied the requirement. When $T \to \infty$, the bound in Lem. 6 is negligible

DOI: 10.54254/2753-8818/2025.GL25881

References

[1] Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. Advances in Neural Information Processing Systems, 24.

[2] Liu, K., & Zhao, Q. (2010). Distributed learning in multi-armed bandit with multiple players. IEEE Transactions on Signal Processing, 58*(11), 5667–5681. https://doi.org/10.1109/TSP.2010.2050482.

[3] Riquelme, C., Tucker, G., & Snoek, J. (2018). Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks and other bandit algorithms. arXiv preprint arXiv:1802.09127.

[4] Osband, I., & Van Roy, B. (2015). Bootstrapped Thompson Sampling and Deep Exploration. arXiv preprint arXiv:1507.00300

[5] Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. arXiv preprint arXiv:1505.05424.

[6] Fan, J., Wang, Z., Yang, Z., & Yu, C. (2023). Provably efficient high-dimensional bandit learning with batched feedbacks. arXiv preprint arXiv:2311.13180v2.

[7] Chakraborty, S., Roy, S., & Tewari, A. (2022). Thompson sampling for high-dimensional sparse linear contextual bandits. arXiv preprint arXiv:2211.05964.

[8] Verstraeten, T., Bargiacchi, E., Libin, P. J. K., Helsen, J., Roijers, D. M., & Nowé, A. (2023). Multi-agent Thompson sampling for bandit applications with sparse neighbourhood structure. arXiv preprint arXiv:2303.04567.

[9] Hanna, O. A., Yang, L. F., & Fragouli, C. (2022). Solving multi-arm bandit using a few bits of communication.2022 IEEE International Symposium on Information Theory (ISIT)* (pp. 2574–2579). IEEE. https://doi.org/10.1109/ISIT50566.2022.9834677.

[10] Li, K., Yang, Y., & Narisetty, N. N. (2023). Regret lower bound and optimal algorithm for high-dimensional contextual linear bandit. arXiv preprint arXiv:2306.08446.

[11] He, J., Zhou, J., & Zhang, T. (2022). Nearly optimal algorithms for linear contextual bandits with corruption. arXiv preprint arXiv:2205.06811.

[12] Qian, W., Ing, C.-K., & Liu, J. (2024). Adaptive algorithm for multi-armed bandit problem with high-dimensional covariates. Journal of the American Statistical Association, 119(546), 970–982.

[13] Munthe-Kaas, H. Z., Verdier, O., & Vilmart, G. (2025). A short proof of Isserlis' theorem. arXiv preprint arXiv:2503.01588.

[14] Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press.

[15] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). Chapman & Hall/CRC.

[16] Qin, T., & Liu, T.-Y. (2013). Introducing LETOR 4.0 datasets. https://www.microsoft.com/en-us/research/project/mslr/

[17] Vershynin, R. (2018). High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press.