Decentralized Blockchain-based Robust Multi-agent Multi-armed Bandit

Anonymous authors
Paper under double-blind review

Abstract

We study a robust multi-agent multi-armed bandit problem where multiple clients or participants are distributed on a fully decentralized blockchain, with the possibility of some being malicious. The rewards of arms are homogeneous among the clients, following time-invariant stochastic distributions that are revealed to the participants only when the system is secure enough. The system's objective is to efficiently ensure the cumulative rewards gained by the honest participants. To this end and to the best of our knowledge, we are the first to incorporate advanced techniques from blockchains, as well as novel mechanisms, into the system to design optimal strategies for honest participants. This allows various malicious behaviors and the maintenance of participant privacy. More specifically, we select a pool of validators who have access to all participants, design a brand-new consensus mechanism based on digital signatures for these validators, invent a UCB-based strategy that requires less information from participants through secure multi-party computation, and design the chain-participant interaction and an incentive mechanism to encourage participants' participation. Notably, we are the first to prove the theoretical guarantee of the proposed algorithms by regret analyses in the context of optimality in blockchains. Unlike existing work that integrates blockchains with learning problems such as federated learning which mainly focuses on numerical optimality, we demonstrate that the regret of honest participants is upper bounded by $\log T$. This is consistent with the multi-agent multi-armed bandit problem without malicious participants and the robust multi-agent multi-armed bandit problem with purely Byzantine attacks.

1 Introduction

Multi-armed Bandit (MAB) (Auer et al., 2002a;b) models the classical sequential decision making process that dynamically balances between exploration and exploitation in an online context. Specifically, in this paradigm, a player engages in a game, from which the player selects precisely one arm and observes the corresponding reward at each time step, and aims to maximize the cumulative reward throughout the game. This is also equivalent to the so-called regret minimization problem navigating the trade-off between exploration (e.g., exploring unknown arms) and exploitation (e.g., favoring the currently known optimal arm). The recent emerging advancement of federated learning, wherein multiple clients jointly train a shared model, has spurred a surge of interest in the domain of multi-agent multi-armed bandit (multi-agent MAB). In this context, multiple clients concurrently interact with multiple MABs, with the objective being the optimization of the cumulative averaged reward across all the clients through communications. Significantly, in addition to the exploration-exploitation trade-off, these clients engage in communication constrained by the underlying graph structure, which necessitates the exploration of the information of other clients and developing strategies accordingly.

Numerous research has been working on the multi-agent MAB problem, including both centralized settings as in (Bistritz and Leshem, 2018; Zhu et al., 3–4, 2021; Huang et al., 2021; Mitra et al., 2021; Réda et al., 2022; Yan et al., 2022), and decentralized settings as in (Landgren et al., 2016a;b; 2021; Zhu et al., 2020; Martínez-Rubio et al., 2019; Agarwal et al., 2022), where it is assumed that reward distributions are uniform among clients, namely homogeneous. Recent attention has shifted towards addressing decentralized, heterogeneous variants, including (Tao et al., 1546–1574, 2022; Wang et al., 1531–1539, 2021; Jiang and Cheng, 1–33, 2023; Zhu et al., 2020; 2021; 3–4, 2021; Zhu and Liu, 2023; Xu and Klabjan, 2023b), which are more general and bring additional complexities. In these scenarios, the shared assumption is that all clients exhibit honesty, refraining from any malicious behaviors, and diligently adhere to both the shared objective and the designed strategies. However, real-world scenarios often deviate from this ideal, featuring inherently malicious clients. Examples include failed machines in parallel computing or the existence of hackers in the email system. Consequently, recent research, such as (Vial et al., 2021), has delved into the multi-agent MAB problem in the context of malicious clients, which is formulated as a robust multi-agent MAB problem. This line of work yields algorithms that perform optimally, provided that the number of malicious clients remains reasonably limited, effectively capturing more general and practical settings. More recently, the work of (Zhu et al., 2023) propose a byzantine-resilient framework and show that collaboration in a setting with malicious clients upgrades the performance if at every time step, the neighbor set of each client contains at least $\frac{2}{3}$ ratio of honest clients and downgrades the performance otherwise.

It is important to note three major concerns with the robust multi-agent MAB framework. First, despite improved regret bounds by (Vial et al., 2021), the possibility of malicious clients compromising estimators cannot be ignored, particularly when accurate estimators are crucial, such as in IoT-driven smart homes (Zhao et al., 2020). This undermines the applicability of the framework in scenarios requiring reliable ground truth knowledge. Second, malicious clients may engage in various disruptive behaviors, not just through estimator manipulation. For instance, they could cause channel congestion, which affects the system's stability and significantly degrades the performance of honest clients, a scenario not adequately addressed in current studies (Vial et al., 2021; Zhu et al., 2023). Third, the existing literature assumes clients are open to sharing detailed interaction data with bandits, imposing significant privacy concerns. These concerns have not been adequately explored and thus pose a major motivation for our work.

Blockchains are fully decentralized structures allowing multiple clients to interact without a central authority, proving highly effective in enhancing security and accuracy across various domains (Feng et al., 2023). Originally developed for peer-to-peer networking and cryptography as discussed by (Nakamoto, 2008), a blockchain consists of a data storage system, a consensus mechanism for secure updates, and a verification process for assessing these updates, often referred to as block operations (Niranjanamurthy et al., 2019). This structure addresses key concerns: first, the verification process ensures the accuracy of information before it is added to the chain, checking the validity of new blocks. Second, the consensus mechanism allows honest clients to reach agreement without prior knowledge of each other's identities, enhancing trust and security, thus mitigating systematic attacks. Despite its potential for increasing privacy through cryptography and decentralization, there is still little understanding of how to integrate blockchain technology within online sequential cooperative decision-making, marking a significant research gap addressed by this paper.

A line of research has successfully adapted blockchains to learning paradigms, notably in blockchain-based federated learning as discussed in (Li et al., 2022; Zhao et al., 2020; Lu et al., 2019; Wang et al., 2022). In these systems, multiple clients on a blockchain aim to optimize model weights of a target model, despite interference from malicious clients. The scale of these models necessitated the introduction of the Interplanetary File System (IPFS), an off-chain storage solution that enhances the stability and efficiency of blockchain operations. However, the decision-making processes in multi-agent MABs differ fundamentally from those in federated learning, making current approaches inapplicable. This gap motivates the development of a new, secure, and reliable framework for multi-agent MAB challenges. Furthermore, while existing studies often focus on deployment performance, the theoretical effectiveness of blockchain-based federated learning, crucial for cybersecurity, remains underexplored. This paper aims to bridge these gaps, analyzing theoretical properties and introducing frameworks with provable optimality for multi-agent MAB in cybersecurity.

Moreover, existing blockchain frameworks have limitations that may not be suitable for multi-agent MAB problems due to their online sequential decision-making nature. For instance, the consensus protocol often assumes the existence of a leader, introducing authority risks. Additionally, the general rule is to secure more than $\frac{2}{3}$ of the votes, which can be impractical in real-world scenarios. Furthermore, the online decision-making problem necessitates the deployment of strategies for real-time interaction with an exogenous environment, a feature not present in traditional blockchain frameworks. Consequently, the adaptations presented in this paper require careful modifications to blockchains and the introduction of new mechanisms. Moreover, little attention has been given to understanding the theoretical properties of blockchains, creating a gap between existing learning theory and blockchains. This holds as the current literature on blockchains and related topics has not yet explored or addressed the theoretical guarantees, even though empirical examination and validation have been conducted across a wide range of domains.

To this end, in this paper, we propose a novel formulation of robust Multi-agent Multi-armed Bandit (multi-agent MAB) within the framework of Blockchains. We are the first to study the robust multi-agent MAB problem where clients are distributed and operate on Blockchains. In this context, clients can only receive rewards when a block is approved to ensure security at each time step, which differs largely from the existing MAB framework. Here, clients are allowed to be malicious and can take various disruptive actions during the game. Blockchains keep track of everything and guarantee functionality through chain operations. This introduces additional complexities, as clients not only design strategies for selecting arms but also interact with both the blockchain and the exogenous environment. Moreover, the presence of blockchain also complicates the traditional bandit feedback, as disapproved blocks introduce new challenges in this online and partial information setting.

We also develop an algorithmic framework for the new formulation with Blockchains, drawing from existing literature while introducing novel techniques given the limitations. This framework includes the design of a validator selection mechanism that eliminates the need for an authorized leader, a departure from existing literature. We also incorporate the arm selection strategy into the framework to perform online sequential decision making. Furthermore, we modify the consensus protocol without relying on majority voting; we use a digital signature scheme (Goldwasser et al., 1988). Moreover, we introduce the role of a smart contract (Hu et al., 2020) and surprisingly enable interaction with the environment through this smart contract. To incentivize the participation of malicious clients in the game, we are the first to design a cost mechanism inspired by the area of mechanism design (Murhekar et al., 2023). It is worth noting that the existence of this smart contract and cost mechanism also guarantees the correctness of the information transmitted on the chain.

On top of this breakthrough in terms of the framework, we also perform theoretical analyses of the proposed algorithms. This involves analyzing the regret, which aims at fundamentally understanding the impact and mechanisms of blockchains within this multi-agent MAB setting given the existence of malicious behaviors. Precisely, we show that under mild assumptions, the regret of honest clients is upper bounded by $O(\log T)$, which is consistent with the existing algorithms for robust multi-agent MAB problems (Zhu et al., 2023; Vial et al., 2021). This is the very first theoretical result on leveraging Blockchains for online learning problems, to the best knowledge of us. Furthermore, this regret bound coincides with the existing regret lower bounds in multi-agent MAB when assuming no clients are malicious (Xu and Klabjan, 2023a), implying its optimality.

The paper is structured as follows. We start by introducing the notations that are used throughout and presenting the problem formulation. Following that, we propose an algorithmic framework for solving the proposed problem. Subsequently, we provide a detailed analysis of the theoretical guarantee regarding the regret associated with the proposed algorithms. Lastly, we present a summary of the paper.

2 Problem Formulation

We start by introducing the notations used throughout the paper. Consistent with the traditional MAB setting, we consider K arms, labeled as $1, 2, \ldots, K$. The time horizon of the game is denoted as T, and it implies the time step $1 \leq t \leq T$. Additionally as in the standard Multi-agent MAB setting, we denote the number of clients as M, and the clients are labeled from 1 to M. It is worth noting that we use the terminologies "client" and "participant" interchangeably for the rest of the paper.

Meanwhile, in our newly proposed blockchain framework, we denote the total number of blocks as $B = T$ and each block at time step t is denoted as b_t. Let us denote the reward of arm i at client m in block b at time step t as $\{r_i^m(b,t)\}_{i,m,b,t}$, which follows a stochastic distribution with a time-invariant mean value $\{\mu_i^b\}_{i,b}$. We denote the set of honest participants and malicious participants as M_H and M_A, respectively. Note that they are time-invariant. We denote the estimators maintained at participant m as $\bar{\mu}_{m,i}(t), \tilde{\mu}_{m,i}(t)$, and the validators estimators as $\tilde{\mu}_i(t)$.

Meanwhile, we introduce some terminologies relevant to this paper.

Existential Forgery Following the definition in (Goldwasser et al., 1988), malicious participants successfully perform an existential forgery if there exists a pair consisting of a message and a signature, such that the signature is produced by an honest participant.

Adaptive Chosen Message Attack Consistent with (Goldwasser et al., 1988), we consider the most general form of message attack, namely the adaptive chosen message attack. In this context, a malicious participant not only has access to the signatures of honest participants but also can determine the message list after seeing these signatures. This grants the malicious participant a high degree of freedom, thereby making the attack more severe.

Universal Composability Framework For homomorphic encryption, we follow the standard framework as in (Canetti, 2001). Specifically, an exogenous environment, also known as an environment machine, interacts sequentially with a protocol. The process runs as follows: the environment sends some inputs to the protocol and receives outputs from the protocol that may contain malicious components. If there exists an ideal adversary such that the environment machine cannot distinguish the difference between interacting with this protocol or the ideal adversary, the protocol is deemed universally composable secure.

Additionally, we propose a novel cost mechanism that penalizes honest clients if the chain is approved with malicious information, inspired by real-world scenarios. The precise description is as follows.

Cost Mechanism We assume that if the estimators from the malicious participants are used in the validated estimators, i.e. $\frac{\partial \tilde{\mu}_i(t)}{\partial \bar{\mu}_{m,i}(t)} \neq 0$, then the honest participants incur a cost of c_t, which they are not aware of until the end of the game.

Subsequently, we define the regret as follows. Formally, the goal is to maximize the total cumulative (expected) reward of honest participants, defined as $\sum_{m \in M_H} \sum_{t=1}^{T} r_{a_m^t}^{m,b}(t)$, $\sum_{m \in M_H} \sum_{t=1}^{T} \mu_{a_m^t}^{b}$, or equivalently, to minimize the regret $R_T = \max_i \sum_{m \in M_H} \sum_{t=1}^{T} r_i^{m,b}(t) - \sum_{m \in M_H} \sum_{t=1}^{T} r_{a_m^t}^{m,b}(t) + \sum_{t=1}^{T} c_t$ and pseudo regret $\bar{R}_T = \max_i \sum_{m \in M_H} \sum_{t=1}^{T} \mu_i^b - \sum_{m \in M_H} \sum_{t=1}^{T} \mu_{a_m^t}^{b} + \sum_{t=1}^{T} c_t$.

We show the rationality of this regret definition as follows. It holds true that these two regret measures are well-defined, considering that M_H is fixed and does not change with time. Furthermore, these definitions align with those used in the context of Blockchain-based federated learning. In these frameworks, the objective is to optimize the model maintained by honest participants, regardless of the intermediate performance and without involving online decision making. Additionally, this definition is consistent with the existing multi-agent MAB problem, except that the cost mechanism is introduced given the existence of malicious participants. In the latter case, the regret is averaged over all participants, which is equivalent to honest participants in both our context and blockchain-based federated learning.

3 Methodologies

In this section, we present the proposed methodologies in this new setting, as outlined in Algorithms 1 and 2, which represent different stages. Notably, we develop the first algorithmic framework at the interface of Blockchains in cybersecurity and multi-agent MAB, addressing the joint challenges of security, privacy, and optimality in online sequential decision making. We leverage the Blockchain structure while introducing new advancements to the existing ones, to theoretically guarantee the functionality of the chain with a new consensus mechanism and cost mechanism. Compared to existing work on Byzantine-resilient multi-agent MAB with malicious participants, our methodology operates on a blockchain with an added layer of reward approval and incorporates secure multi-party computation into the communication process, which largely improves the security and privacy of honest participants with minimal risk of being compromised. We also introduce a cost mechanism to incentivize the participation of malicious participants. This mechanism is driven by real-world applications and is consistent with blockchain-based federated learning. It has been customized for this online decision-making regime to guarantee the correctness and optimality of the framework.

More specifically, the algorithmic framework consists of 3 algorithms. Algorithms 1 and 2 constitute the core of the methodology, including the sequential strategies executed by the honest participants, black-box operations by the malicious participants, and the chain executions. Moreover, we use a sub-algorithm (see Appendix) integrated into the validation selection procedure of Algorithm 1 to ensure the proper execution of the sampling process based on the desired criteria.

The main algorithm includes several stages, as indicated in the following order.

Validator selection At each time step, the entire system first selects a sub-pool of participants allowed to act on the chain. Specifically, the system samples the set of validators based on the trust coefficients of participants, which are initialized as 1 and updated sequentially. The detailed pseudo code is in Appendix. The purpose of this step is to guarantee security and efficiency, as the chain relies only on a proportion of participants whcih have access to the system's information.

Arm selection This step is common in the MAB framework, where participants decide which arm to pull sequentially. The strategies depend on whether participants are honest or malicious. For honest participants, the strategy follows a UCB-like approach. More specifically, for each honest participant m, it assigns a decision criterion to each arm i and selects the arm with the highest criterion, which can be formally written as $a_m^t = argmax \tilde{\mu}_{m,i}(t) + F(m,i,t)$ where $\tilde{\mu}_{m,i}(t)$ is the maintained estimator at participant m. Here $F(m,i,t) = \sqrt{\frac{C_1 \log t}{n_{m,i}(t)}}$ with C_1 being specified in Theorem 1. For malicious participant j, however, it selects arms based on arbitrary strategies, which is also known as Byzantine's attack and written as $a_j^t = f_j^t(i)$ where f_j^t is any mapping from a space R^K to a scalar space R^1.

Consensus The consensus protocol is central to the execution of the chain and guaranteeing that the chain is growing as expected. More specifically, we incorporate the digital signature scheme (Goldwasser et al., 1988) into the protocol and use the solution to the Byzantine General Problem (Lamport et al., 2019) under any number of malicious validators. To expand, malicious participants broadcast their estimators $\bar{\mu}_{m,i}(t)$ to the validators using possibly Byzantine's attack or a backdoor attack. On the contrary, honest participants broadcast their estimators $\bar{\mu}_{m,i}(t)$ to the validators.

For each honest validator h, it determines the set, A_t, B_t as follows. The set A_t is $m \in A_t \Leftrightarrow k_i n_{m,i}(t) > n_{j,i}(t) \Leftrightarrow m \in M_H$ that can be constructed through the secure multi-party computation protocol as in (Asharov et al., 2012) and the set B_t is as follows. If $|A_t| > 2f$, then $B_t = \{m : \bar{\mu}_{m,i}(t) \text{ is smaller than the top } f \text{ values and larger than the below } f \text{ values}\}$, and otherwise, $B_t = \{t \bmod K\}$. Alternatively, we use the notations $top(C, -f)$ and $top(C, f)$ to denote values below f and the top f values in set C, respectively. Once again, the malicious participants choose the sets A_t and B_t in a black-box manner.

And then, validators broadcast B_t to other validators after attaching their signatures, repeating this process at least M times, based on the algorithm in (Lamport et al., 2019). The consensus is successful if at least one estimator is present in B_t. Otherwise, the consensus step fails, resulting in an empty set of estimators.

Global Update The set B_t is sent to the smart contract, which then computes the average of the estimators within B_t, known as the global update. More precisely, for each arm i, the estimator is computed as $\tilde{\mu}_i(t) = \frac{\sum_{m \in B_t} \bar{\mu}_{m,i}(t)}{|B_t|}$ if B_t is not empty, and $\tilde{\mu}_i(t) = \infty$ otherwise.

Algorithm 1: BC-UCB

Initialization: For participants $1, 2, \ldots, M$, arms $1, \ldots, K$, at time step 0, in block $1 \leq b \leq B$ we have $\tilde{\mu}_i^{m,b}(1)$, $\hat{\tilde{\mu}}_i^b(1)$, $w_i^m(1)$, $N_{m,i}(1) = n_{m,i}(1)$; cumulative reward $V_1^m = 1$ for participant $1 \leq m \leq M$; parameter δ; the number of honest participants $M_H \geq \frac{2}{3}M$; the expected number of validators M_V; The decision on whether participant j is honest at validator m is initialized as $D_{m,j}(0) = 1$; random seed r; for any set S_0, $majority(S_0)$ refers to the majority or the median of S_0;

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

  Participants compete to be the validator of block b using Proof-of-Work: [**Input:** $w_m(t)$, M_V; **Output:** $S_V(t)$]

  `// Validators`

    Participant m's probability of being a validator is proportional to the trust coefficient $w_m(t)$, i.e. $p_m(t) = \frac{w_m(t)}{\sum_m w_m(t)}$;

    for *each participant* j **do**

      Sample whether it belongs to the set of validators $S_V(t)$ based on $p_j(t)$ in a decentralized manner

    end

  for *each participant* $m \in M_H$ **do** `// [Input:` $M_H, n_{m,i}(t), N_{m,i}(t), K, F(m,i,t), S_V(t)$`; Output:` a_m^t`] UCB`

    if *there is no arm* i *such that* $n_{m,i}(t) \leq N_{m,i}(t) - K$ **then**

      $a_m^t = argmax_i \tilde{\mu}_{m,i}(t) + F(m,i,t)$

    else

      Randomly sample an arm a_m^t.

    end

    Pull arm a_m^t;

    Broadcast its estimators $\tilde{\mu}_i^{m,b}(t)$ to validators $S_V(t)$ that are determined in the *Validators step* regarding the select arm a_m^t;

  end

  for *each participant* $m \notin M_H$ **do** `// [Input:` $M_H, f, S_V(t)$`; Output:` a_m^t`] Byzantine's attack`

    Broadcast its estimators to validators $S_V(t)$ that are determined in the *Validator step* using possibly Byzantine's attack regarding an arm a_m^t, i.e. $\tilde{\mu}_i^{m,b}(t) = f(m,b,t)$ where f is an arbitrary mapping from the space of $R^{M \times B \times T}$ to R;

  end

  Validators use majority voting to achieve consensus on the estimators $\tilde{\mu}_i^{m,b}(t)$ as follows: `// Consensus` ;

    for *each validator* $k \in S_V(t) \cap M_H$ **do** `// [Input:` $\{\tilde{\mu}_i^{m,b}(t)\}_{m \leq M}, S_V(t), M_H, \delta$`; Output:` $D_{k,j}(t)$`]`

      for *each participant* j **do**

        $D_{k,j}(t) = 1 \cdot 1_{j \in A_k(t)}$

      end

    end

    for *each validator* $k \in S_V(t)$ *and* $k \notin M_H$ **do** // [Input: $\{\tilde{\mu}_i^{m,b}(t)\}_{m \leq M}, S_V(t), M_H, \delta, \bar{f}$; Output: $D_{k,j}(t)$]

      for *each participant* j **do**

        $D_{k,j}(t) = \bar{f}(\tilde{\mu}_i^{j,b}(t), \tilde{\mu}_i^{k,b}(t), \delta)$

      end

    end

    for *each validator* k **do** `// [Input:` $S_V(t), \{D_{k,j}(t)\}_{j \leq M}$`; Output:` $\{D_{k,j}(t)\}_{k \in S_V(t), j \leq M}$`]`

      Broadcast the estimators $\{D_{k,j}(t)\}_{j \leq M}$ to validatorrs $S_V(t) \backslash \{j\}$

    end

    for *each validator* $k \in S_V(t)$ *and* $k \notin M_H$ **do**

      for *each participant* j **do** `// [Input:` $S_V(t), \{D_{l,j}(t)\}_{l \in S_V(t)}$`; Output:` $\tilde{M}_H^{k,t}$`]`

        Arbitrarily determine whether j is malicious, denoted as $j \notin \tilde{M}_H^{k,t}$ or honest, denoted as $j \in \tilde{M}_H^{k,t}$

      end

      Broadcast the estimators $B_k(t) = \tilde{M}_H^{k,t}$ to validators $S_V(t) \backslash \{k\}$

    end

    for *each validator* $k \in S_V(t) \cap M_H$ **do**

      for *each participant* j **do** `// [Input:` $S_V(t), \{D_{l,j}(t)\}_{l \in S_V(t)}$`; Output:` $\tilde{M}_H^{k,t}$`]`

        if $j \in A_k(t)$ *and* $|A_k(t)| > 2f$ **then**

          participant j belongs to $B_k(t)$ if $top(\{\tilde{\mu}_{j,i}(t)\}_j, -f) < \tilde{\mu}_{j,i}(t) < top(\{\tilde{\mu}_{j,i}(t)\}_j, f)$

        else

          the estimator from participant j is ignored

        end

      end

      Broadcast the estimators $B_k(t) = \tilde{M}_H^{k,t}$ to validators $S_V(t) \backslash \{k\}$

    end

    Denote the estimator $\tilde{M}_H^t = \cup_k \tilde{M}_H^{k,t}$;

    for *each validator* $k \in S_V(t)$ **do**

      for *round* $1 \leq h \leq M$ **do** `// [Input:` $S_V(t), \{\tilde{M}_H^{l,t}\}_{l \in S_V(t)}$`; Output:` $\tilde{M}_H^t$`]`

        generate the digital signature as in (Goldwasser et al., 1988);

        execute Algorithm $SM(M)$ in (Lamport et al., 2019);

        derive the set $\tilde{M}_H^t$

      end

    end

end

Algorithm 2: BC-UCB

for $t = 1, 2, \ldots, T$ **do**
  The validators obtain the corresponding subset of estimators $S = \{\tilde{\mu}_i^{j,b}\}_i^j$ for $j \in \tilde{M}_H^t$; **[Input:** $\tilde{M}_H^t, \tilde{\mu}_i^{m,b}(t)$**; Output:** S]
  The verified estimators/arms are immutable and used for updating global estimators $\hat{\tilde{\mu}}_i^b(t)$ by the following:`// Global update` ;
  if *set S is not empty* **then**
    for each arm i, the global estimator in block b at time step t is the averaged local estimators in set S written as $\hat{\tilde{\mu}}_i^b(t) = \frac{1}{|\tilde{M}_H^t|} \sum_{m \in \tilde{M}_H^t} \tilde{\mu}_i^{m,b}(t)$
  else
    for each arm i, the global estimator is $\hat{\tilde{\mu}}_i^b(t) = -\infty$
  end
  A smart-contract validates block b: `// [Input:` $\hat{\tilde{\mu}}_i^b(t)$`; Output: approve or disapprove] Verification`
    Approve the block if the global estimators $\hat{\tilde{\mu}}_i^b(t)$ is not $-\infty$ for arm i;
    Disapprove the block otherwise;
  if *the block is approved* **then** `// Block operation`
    Input: $V_t, \{\tilde{\mu}_i^{m,b}(t),\ N_{m,i}(t), n_{m,i}(t)\}_{m \leq M}$**; Output: the corresponding estimators at** $t+1$
    for *each participant m* **do**
      Append the validated block to the chains;
      Collect the reward $r^m_{a^t_m(b)}(b,t)$ of the pulled arm in that block and add it to the cumulative reward by $V_{t+1} = V_t + r^m_{a^t_m(b)}(b,t)$;
      Update estimators $\tilde{\mu}_i^{m,b}(t)$, $N_{m,i}(t)$, and $n_{m,i}(t)$ based on the rewards;
      $\tilde{\mu}_i^{m,b}(t+1) = \frac{\tilde{\mu}_i^{m,b}(t) \cdot n_{m,i}(t) + r^m_{a^t_m(b)}(b,t) \cdot 1_{a^t_m(b)=i}}{n_{m,i}(t) + 1_{a^t_m(b)=i}}$; $n_{m,i}(t+1) = n_{m,i}(t) + 1_{a^t_m(b)=i}$;
      $N_{m,i}(t+1) = N_{m,i}(t) + 1_{\{\exists m, a^t_m(b)=i\}} \cdot 1_{m \in S_V(t)}$
    end
  else
    Skip the current time step by not collecting the rewards and not updating the estimators;
  end
  for *each participant m* **do** `// [Input:` $w_m(t), \delta, \tilde{\mu}^{m,b}(t), \hat{\tilde{\mu}}_i^b(t), b$`; Output:` $w_m(t+1), b$`]Participants' update`
    The smart contract computes $Dist_m(t) = Distance(\tilde{\mu}^{m,b}(t), \hat{\tilde{\mu}}^b(t)) \leq \delta$;
    Receive reward $w_t = \delta - Dist_m(t)$;
    Update $w_m(t+1) = w_m(t) \exp\{w_t\}$ and $b = b + 1$;
  end
end

Block Verification The smart contract validates the block and assigns $b_t = 1$ if the estimator satisfies the condition $\tilde{\mu}_i(t) \leq 2$. It disapproves the block otherwise, which is denoted as $b_t = 0$.

Block Operation The smart contract sends the output containing the validated estimator $\tilde{\mu}_i(t)$, the set B_t, and the indicator b_t of whether the block is approved to the environment. Subsequently, the environment determines the rewards to be distributed to the participants as follows.

case 1: if the block is approved with $b_t = 1$ and $B_t \subset M_H$, i.e. $\frac{\partial \tilde{\mu}_i(t)}{\partial \bar{\mu}_{m,i}(t)} = 0$ where $m \notin M_H$ then the environment distributes $r^m_{a^t_m}(t)$ and $\tilde{\mu}_i(t)$ to participant m for any $1 \leq m \leq M$.

case 2: if the block is approved with $b_t = 1$ and $B_t \cap M_H < |B_t|$, i.e. $\frac{\partial \tilde{\mu}_i(t)}{\partial \bar{\mu}_{m,i}(t)} \neq 0$ where $m \notin M_H$ then the environment distributes $r^m_{a^t_m}(t) - c_t$ and $\tilde{\mu}_i(t)$ to any honest participant m, and $r^m_{a^t_m}(t) + c_t$ to any malicious participant.

case 3: if the block is disapproved with $b_t = 0$, the environment distributes nothing to the participants.

Participants' Update After receiving the information from the environment, the honest participants update their maintained estimators as follows. For the reward estimator $\tilde{\mu}m, i(t)$, they update it when they receive $\tilde{\mu}_i(t)$, i.e. $\tilde{\mu}_{m,i}(t) = \tilde{\mu}_{m,i}(t)$, and otherwise, $\tilde{\mu}_{m,i}(t) = \bar{\mu}_{m,i}(t)$. For the number of arm pulls, they update it as $n_{m,i}(t) = n_{m,i}(t-1) + 1_{b_t=1} \cdot 1_{a^t_m=i}$.

For the trust coefficients, every participant needs to update them based on their rewards, which can again be achieved through the secure multi-party computation protocol. More precisely, let the trust coefficient $w_m(t)$ be updated as follows $w_m(t) = w_m(t-1)\exp\{h^t_m\}$ where $h_m(t)$ can be any values.

4 Regret Analysis

In this section, we demonstrate the theoretical guarantee of our proposed framework, by analyzing the regret defined over the honest participants. The formal statement reads as follows.

Theorem 1. *Let us assume that the total number of honest participants is at least* $\frac{2}{3}M$ *and let us specify the quantities* $h_m(t) = 0$ *and* $c_t = c$*. Meanwhile, let us assume that the malicious participants perform existential forgery on the signatures of honest participants with an adaptive chosen message attack. Lastly, let us assume that the participants are in a standard universal composability framework when constructing* A*. Here the set* A *is defined as* $A = \{\forall 1 \leq t \leq T, b_t = 1\}$ *which satisfies that* $P(A) \geq \frac{1}{l^{T-1}}$*. Then we have that* $E[R_T|A] \leq (c+1) \cdot L + \sum_{m \in M_H} \sum_{k=1}^{K} \Delta_k([\frac{4C_1 \log T}{\Delta_i^2}] + \frac{\pi^2}{3}) + |M_H|Kl^{1-T}$ *where* $L = O(\log T)$ *is the length of the burn-in period,* c *is the cost,* C_1 *meets the condition that* $\frac{C_1}{2|M_H|k_i\sigma^2} \geq 1$*,* Δ_i *is the sub-optimality gap,* l *is the length of the signature of the participants, and* k_i *is the threshold parameter used in the construction of* A_t*.*

Proof sketch. The full proof is provided in Appendix; the main logic is as follows. We decompose the regret into three parts: 1) the length of the burn-in period, 2) the gap between the rewards of the optimal arm and the received rewards, and 3) the cost induced by selecting the estimators of the malicious participants. For the second part of the regret, we bound it in two aspects. First, we analyze the total number of times rewards are received, i.e., when the block is approved, which is of order $1 - l^{1-T}$. Then, we control the total number of times sub-optimal arms are pulled using our developed concentration inequality for the validated estimators sent for verification. Concerning the third part, we bound it by analyzing the construction of B_t, which depends on the presence of malicious clients in A_t. By demonstrating that A_t contains only a small number of malicious participants in comparison to the total number of honest participants, we show that B_t does not induce additional cost. Combining the analysis of these three parts, we derive the upper bound on regret. □

5 Conclusion

This paper addresses the robust multi-agent multi-armed bandit problem within the framework of blockchains, representing the first work to explore online sequential decision-making with clients distributed on a blockchain. Our approach focuses on ensuring security, privacy, and optimality from a system perspective in the context of distributed online sequential decision making, distinct from blockchain-based federated learning or byzantine-resilient multi-agent MAB. To tackle this challenge, we introduce a novel algorithmic framework. A group of participants forms a validator set responsible for achieving consensus on information transmitted by all participants, ensuring security and privacy. Validated information is then sent to a smart contract for verification, with rewards distributed only upon successful verification. As part of our contributions, we use trust coefficients to determine validator selection probabilities.Additionally, we incorporate a digital signature scheme into the consensus process, eliminating the traditional $\frac{1}{3}$ assumption of the Byzantine general problem. Furthermore, we introduce a cost mechanism to incentivize malicious participants by rewarding their contributions to the verification step. We provide a rigorous regret analysis demonstrating the optimality of our proposed algorithm under specific assumptions, marking a breakthrough in blockchain-related literature.

References

M. Agarwal, V. Aggarwal, and K. Azizzadenesheli. Multi-agent multi-armed bandits with limited communication. *The Journal of Machine Learning Research*, 23(1):9529–9552, 2022.

G. Asharov, A. Jain, A. López-Alt, E. Tromer, V. Vaikuntanathan, and D. Wichs. Multiparty computation with low communication, computation and interaction via threshold fhe. In *Advances in Cryptology–EUROCRYPT 2012: 31st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Cambridge, UK, April 15-19, 2012. Proceedings 31*, pages 483–501. Springer, 2012.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

I. Bistritz and A. Leshem. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 31, 2018.

R. Canetti. Universally composable security: A new paradigm for cryptographic protocols. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 136–145. IEEE, 2001.

X. Feng, L. Li, T. Wang, W. Xu, J. Zhang, B. Wei, and C. Luo. Cobc: A blockchain-based collaborative inference system for the internet of things. *IEEE Internet of Things Journal*, 2023.

S. Goldwasser, S. Micali, and R. L. Rivest. A digital signature scheme secure against adaptive chosen-message attacks. *SIAM Journal on computing*, 17(2):281–308, 1988.

B. Hu, Z. Zhang, J. Liu, Y. Liu, J. Yin, R. Lu, and X. Lin. A comprehensive survey on smart contract construction and execution: Paradigms, tools, and systems. patterns, 2 (2), 100179, 2020.

R. Huang, W. Wu, J. Yang, and C. Shen. Federated linear contextual bandits. *Advances in Neural Information Processing Systems*, 34:27057–27068, 2021.

F. Jiang and H. Cheng. Multi-agent bandit with agent-dependent expected rewards. *Swarm Intelligence*, 1–33, 2023.

L. Lamport, R. Shostak, and M. Pease. The byzantine generals problem. In *Concurrency: the works of leslie lamport*, pages 203–226. 2019.

P. Landgren, V. Srivastava, and N. E. Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference*. 243–248. IEEE, 2016a.

P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control*. 167–172. IEEE, 2016b.

P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125:109445, 2021.

D. Li, Z. Luo, and B. Cao. Blockchain-based federated learning methodologies in smart environments. *Cluster Computing*, 25(4):2585–2599, 2022.

Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang. Blockchain and federated learning for privacy-preserved data sharing in industrial iot. *IEEE Transactions on Industrial Informatics*, 16 (6):4177–4186, 2019.

D. Martínez-Rubio, V. Kanade, and P. Rebeschini. Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Mitra, H. Hassani, and G. Pappas. Exploiting heterogeneity in robust federated best-arm identification. *arXiv preprint arXiv:2109.05700*, 2021.

A. Murhekar, Z. Yuan, B. R. Chaudhury, B. Li, and R. Mehta. Incentives in federated learning: Equilibria, dynamics, and mechanisms for welfare maximization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized business review*, 2008.

M. Niranjanamurthy, B. Nithya, and S. Jagannatha. Analysis of blockchain technology: pros, cons and swot. *Cluster Computing*, 22:14743–14757, 2019.

C. Réda, S. Vakili, and E. Kaufmann. Near-optimal collaborative learning in bandits. In *2022-36th Conference on Neural Information Processing System*, 2022.

Y. Tao, Y. Wu, P. Zhao, and D. Wang. Optimal rates of (locally) differentially private heavy-tailed multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, 1546–1574, 2022.

D. Vial, S. Shakkottai, and R. Srikant. Robust multi-agent multi-armed bandits. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 161–170, 2021.

J. Wang, H. Sun, and T. Xu. Blockchain-based secure and efficient federated learning with three-phase consensus and unknown device selection. In *International Conference on Wireless Algorithms, Systems, and Applications*, pages 453–465. Springer, 2022.

Z. Wang, C. Zhang, M. K. Singh, L. Riek, and K. Chaudhuri. Multitask bandit learning through heterogeneous feedback aggregation. In *International Conference on Artificial Intelligence and Statistics*, 1531–1539, 2021.

M. Xu and D. Klabjan. Regret lower bounds in multi-agent multi-armed bandit. *arXiv preprint arXiv:2308.08046*, 2023a.

M. Xu and D. Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. *Advances on Neural Information Processing Systems*, 2023b.

Z. Yan, Q. Xiao, T. Chen, and A. Tajer. Federated multi-armed bandit via uncoordinated exploration. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 5248–5252. IEEE, 2022.

Y. Zhao, J. Zhao, L. Jiang, R. Tan, D. Niyato, Z. Li, L. Lyu, and Y. Liu. Privacy-preserving blockchain-based federated learning for iot devices. *IEEE Internet of Things Journal*, 8(3): 1817–1829, 2020.

J. Zhu and J. Liu. Distributed multi-armed bandits. *IEEE Transactions on Automatic Control*, 2023.

J. Zhu, R. Sandhu, and J. Liu. A distributed algorithm for sequential decision making in multi-armed bandit with homogeneous rewards. In *59th IEEE Conference on Decision and Control*. 3078–3083. IEEE, 2020.

J. Zhu, E. Mulle, C. S. Smith, and J. Liu. Decentralized multi-armed bandit can outperform classic upper confidence bound. *arXiv preprint arXiv:2111.10933*, 2021.

J. Zhu, A. Koppel, A. Velasquez, and J. Liu. Byzantine-resilient decentralized multi-armed bandits. *arXiv preprint arXiv:2310.07320*, 2023.

Z. Zhu, J. Zhu, J. Liu, and Y. Liu. Federated bandit: A gossiping approach. In *Abstract Proceedings of the 2021 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, 3–4, 2021.

Appendix

Sampling procedure

The pseudo code for the sample step, based on the trust coefficients of clients, is presented in the following algorithm, namely Algorithm 3.

Algorithm 3: Sampling using sortition based on trust coefficients

Input: time step t; trust coefficients $w_m j(t)$; $VRF = (G, F)$ (verifiable random functions); the length of hash value is denoted as $hashlen$; For any integer $k < w$, the probability density function of selecting k samples out of w following binomial distribution with successful probability p, is denoted as $B(k, w, p)$;

for *time step t* **do**

  choose a public-private key pair (pk_j, sk_j), both of which are binary strings, by probabilistic function $G(j)$;

  the hash value $hash$ and the proof π, which are two binary strings, are given by $< hash, \pi >= F(r, sk_j)$;

  $p_j = \frac{M_V}{\sum_m w_m(t)}$;

  $z = 0$;

  while $\frac{hash}{2^{hashlen}} \notin [\sum_{k=0}^{z} B(k, w_j(t), p_j), \sum_{k=0}^{z+1} B(k, w_j(t), p_j))$ **do**

    $z = z + 1$

  end

end

Proof of Theorem 1

Proof. For regret, we have the following decomposition. Let us denote b_t as the indicator function of whether the block at time step t is approved. Likewise, for any time step t, we denote whether the estimators from the malicious clients are utilized in the integrated estimators as h_t. Let the length of the burn-in period be L.

Note that

$$
\begin{aligned}
R_T &= \max_i \sum_{m \in M_H} \sum_{t=1}^{T} \mu_i - \sum_{m \in M_H} \sum_{t=1}^{T} (\mu^b_{a^t_m} - c_t) \\
&= \max_i \sum_{m \in M_H} \sum_{t=1}^{T} \mu_i - \sum_{m \in M_H} \sum_{t=1}^{T} \mu^b_{a^t_m} + \sum_{m \in M_H} \sum_{t=1}^{T} c_t \\
&= \max_i \sum_{m \in M_H} \sum_{t=1}^{T} \mu_i - \sum_{m \in M_H} \sum_{t=1}^{T} \mu_{a^t_m} 1_{b_t=1} + \sum_{m \in M_H} \sum_{t=1}^{T} c_t \\
&= \max_i \sum_{m \in M_H} \sum_{t=1}^{T} \mu_i - \sum_{m \in M_H} \sum_{t=1}^{T} \mu_{a^t_m} 1_{b_t=1} + \sum_{m \in M_H} \sum_{t=1}^{T} c 1_{h_t=1}
\end{aligned}
$$

Meanwhile, the regret can be bounded as follows

$$
\begin{aligned}
R_T &\leq L + c \cdot L + \sum_{t=L+1}^{T} \sum_{m \in M_H} (\mu_{i^*} - \mu_{a^t_m} 1_{b_t=1}) + \sum_{m \in M_H} \sum_{t=L+1}^{T} c 1_{h_t=1} \\
&\doteq (c+1) \cdot L + T_1 + T_2
\end{aligned}
$$

We start with the second term T_2. Note that $h_t = 1$ is equivalent to $\{m : m \in B_t \cap m \notin M_H\} \neq \emptyset$.

By taking the expectation over T_2, we derive

$$
\begin{aligned}
E[T_2|A] &= \sum_{m \in M_H} \sum_{t=L+1}^{T} cE[1_{h_t=1}] \\
&= \sum_{m \in M_H} \sum_{t=L+1}^{T} cE[1_{\{m:m \in B_t \cap m \notin M_H\} \neq \emptyset}]
\end{aligned}
$$

Based on Lemma 2 in (Zhu et al., 2023), we obtain that

$$1_{\{m:m\in B_t\cap m\notin M_H\}\neq\emptyset} = 1_{|A_t|<2f}$$

which immediately implies that

$$\begin{aligned} E[T_2|A] &= \sum_{m\in M_H}\sum_{t=L+1}^{T} cE[1_{h_t=1}] \\ &= \sum_{m\in M_H}\sum_{t=L+1}^{T} cE[1_{\{m:m\in B_t\cap m\notin M_H\}\neq\emptyset}] \\ &= \sum_{m\in M_H}\sum_{t=L+1}^{T} cE[1_{|A_t|<2f}]. \end{aligned}$$

In the meantime, we note that for any honest validators, the choice of A_t guarantees that honest participants are included after the burn-in period. More specifically, the set of A_t satisfies that for any validator $j \in M_H$,

$$m \in A_t \Leftrightarrow k_i n_{m,i}(t) > n_{j,i}(t) \Leftrightarrow m \in M_H$$

where $1 < k_i < 2$. This condition holds at the end of burn-in period which is straightforward since each honest. After the burn-in period, the honest participants has the same decision rule

$$a_m^t = argmax_i \tilde{\mu}_{m,i}(t) + F(m,i,t)$$

where $\tilde{\mu}_{m,i}(t) = \tilde{\mu}_i^b(t)$. In other words, each honest client uses the validated estimator $\tilde{\mu}_i^b(t)$.Since both $n_{m,i}(t)$ and $n_{j,i}(t)$ are larger than $\frac{L}{K}$, then we have that there exists $k_i = \frac{n_{j,i}(t)K}{L}$, such as $k_i n_{m,i}(t) > n_{j,i}(t)$ for every $m \in M_H$.

This implies that

$$A_t > |M_H| \geq 2f$$

by the assumption that the number of honest participants is at least $\frac{2}{3}M$.

That is to say,

$$E[1_{|A_t|>2f}] = 1$$

and subsequently, we have

$$\begin{aligned} E[T_2|A] &= \sum_{m\in M_H}\sum_{t=L+1}^{T} cE[1_{|A_t|<2f}] \\ &= 0 \end{aligned}$$

We note that the construction of A_t is done without knowing the number of pulls of arms of other clients. This is realized by using the homomorphic results, Theorem 5.2 as in (Asharov et al., 2012) under the universal composability framework.

Next, we proceed to bound the first term T_1. Note that

$$\begin{aligned} E[T_1|A] &\leq \sum_{t=L+1}^{T}\sum_{m\in M_H}(\mu_{i^*} - \mu_{a_m^t} 1_{b_t=1}) \\ &= (T-L)\cdot|M_H|\cdot\mu_{i^*} - \sum_{m\in M_H}\sum_{t=L}^{T} E[\mu_{a_m^t}|b_t=1]P(b_t=1) \end{aligned}$$

In the meantime, we obtain the following

$$\begin{aligned}
&E[\mu_{a_m^t}|b_t = 1] \\
&= E[\sum_{k=1}^{K} \mu_k \cdot 1_{a_m^t = k}|b_t = 1] \\
&= \sum_{k=1}^{K} E[\mu_k 1_{a_m^t = k}|b_t = 1] \\
&\geq \sum_{k=1}^{K} \mu_k \cdot \frac{1}{P(b_t = 1)} \cdot (E[1_{a_m^t = k}] - P(b_t = 0)).
\end{aligned}$$

This immediately gives us that

$$\begin{aligned}
&E[T_1|A] \\
&\leq (T-L)|M_H|\mu_{i^*} - \sum_{m \in M_H} \sum_{t=L}^{T} (\sum_{k=1}^{K} \mu_k \cdot \frac{1}{P(b_t = 1)} \cdot (E[1_{a_m^t = k}] - P(b_t = 0)))P(b_t = 1) \\
&= (T-L)|M_H|\mu_{i^*} - \sum_{m \in M_H} \sum_{t=L}^{T} (\sum_{k=1}^{K} \mu_k (E[1_{a_m^t = k}] - P(b_t = 0)) \\
&= (T-L)|M_H|\mu_{i^*} - \sum_{m \in M_H} \sum_{t=L}^{T} \sum_{k=1}^{K} \mu_k E[1_{a_m^t = k}] + \sum_{m \in M_H} \sum_{t=L}^{T} \sum_{k=1}^{K} \mu_k P(b_t = 0).
\end{aligned}$$

Based on Theorem 2 in (Lamport et al., 2019), the consensus is achieved, i.e. $b_t = 1$, as long as the digital signatures of the honest participants can not be forged. Based on our assumption, we have that the malicious participants can only perform existential forgery on the signatures of the honest participants and the attacks are adaptive chosen-message attack. Then based on the result, Main Theorem in (Goldwasser et al., 1988), the attack holds with probability at most $\frac{1}{Q(l)}$ for any polynomial function Q and large enough l where l is the length of the signature.

More precisely, we have that with probability at least $1 - \frac{1}{l^T}$, the signature of the honest participants can not be forged, and thus, the consensus can be achieved, i.e.

$$P(b_t = 1) \geq 1 - \frac{1}{l^T}.$$

Subsequently, we derive that

$$\begin{aligned}
(10) &\leq (T-L)|M_H|\mu_{i^*} - \sum_{m \in M_H} \sum_{t=L}^{T} \sum_{k=1}^{K} \mu_k E[1_{a_m^t = k}] + \sum_{m \in M_H} \sum_{t=L}^{T} \sum_{k=1}^{K} \mu_k (\frac{1}{l^T}) \\
&\leq \sum_{m \in M_H} \sum_{t=L}^{T} (\mu_{i^*} - \sum_{k=1}^{K} \mu_k E[1_{a_m^t = k}]) + |M_H| K l^{T-1} \\
&= \sum_{m \in M_H} \sum_{k=1}^{K} \Delta_k E[n_{m,k}(t)] + |M_H| K l^{T-1} \\
&\doteq T_{21} + |M_H| K l^{T-1}
\end{aligned}$$

And for each honest participant, they are using the estimators based on the validated estimators, as long as the block is approved. Consider the following event, $A = \{\forall 1 \leq t \leq T, b_t = 1\}$. Based on (10) and the Bonferroni's inequality, we obtained that

$$
\begin{aligned}
P(A) &= P(\forall 1 \le t \le T, b_t = 1) \\
&= 1 - P(\exists 1 \le t \le T, b_t = 0) \\
&\le 1 - \sum_{t=1}^{T} P(b_t = 0) \\
&\le \frac{1}{l^{T-1}}.
\end{aligned}
$$

On event A, the blockchain always gets approved, and then all the honest participants follow the validated estimators from the validators. By (10) and Lemma 2 in (Zhu et al., 2023), we have that the validated estimator $\tilde{\mu}_i(t)$ can be expressed as

$$
\tilde{\mu}_i(t) = \sum_{j \in A_t \cap M_H} w_{j,i}(t) \bar{\mu}_{j,i}(t)
$$

which is also equivalent to $\tilde{\mu}_{m,i}(t)$. Here the weight $w_{j,i}(t)$ meets the condition

$$
\sum_{j \in A_t \cap M_H} w_{j,i}(t) = 1,
$$

which immediately implies that

$$
E[\tilde{\mu}_i(t)] = \mu_i.
$$

We note that the variance of $\tilde{\mu}_i(t)$, $var(\tilde{\mu}_i(t))$, satisfies that,

$$
\begin{aligned}
var(\tilde{\mu}_i(t)) &= var(\sum_{j \in A_t \cap M_H} w_{j,i}(t) \bar{\mu}_{j,i}(t)) \\
&\le |A_t \cap M_H| \sum_{j \in A_t \cap M_H} w_{j,i}(t)^2 var(\bar{\mu}_{j,i}(t))) \\
&\le |A_t \cap M_H| \sum_{j \in A_t \cap M_H} w_{j,i}^2(t) \sigma^2 \frac{1}{n_{j,i}(t)} \\
&\le |A_t \cap M_H| \sum_{j \in A_t \cap M_H} w_{j,i}^2(t) \sigma^2 \frac{k_i}{n_{m,i}(t)} \\
&= |M_H| \frac{k_i}{n_{m,i}(t)} \sum_{j \in M_H} w_{j,i}^2(t) \sigma^2 \\
&\le |M_H| \frac{k_i \sigma^2}{n_{m,i}(t)}
\end{aligned}
$$

where the inequality holds by the Cauchy-Schwarz inequality, the second inequality holds by the definition of sub-Gaussian distributions, the third inequality results from the construction of A_t, and the last inequality is as a result of $(a+b)^2 \ge a^2 + b^2$.

Subsequently, we have that

$$
\begin{aligned}
& P(\tilde{\mu}_i^m(t) - \sqrt{\frac{C_1 \log t}{n_{m,i}(t)}} > \mu_i, n_{m,i}(t-1) \geq l) \\
& \leq \exp\{-\frac{(\sqrt{\frac{C_1 \log t}{n_{m,i}(t)}})^2}{2var(\tilde{\mu}_i^m)}\} \\
& \leq \exp\{-\frac{(\sqrt{\frac{C_1 \log t}{n_{m,i}(t)}})^2}{2|M_H|\frac{k_i\sigma^2}{n_{m,i}(t)}}\} \\
& = \exp\{-\frac{C_1 \log t}{2|M_H|k_i\sigma^2}\} \leq \frac{1}{t^2}
\end{aligned}
$$

where the first inequality holds by Chernoff bound, the second inequality is derived by plugging in the above upper bound on $var(\tilde{\mu}_i^m(t))$, and the last inequality results from then choice of C_1 that satisfies $\frac{C_1}{2|M_H|k_i\sigma^2} \geq 1$.

Likewise, by symmetry, we have

$$
P(\tilde{\mu}_i^m(t) + \sqrt{\frac{C_1 \log t}{n_{m,i}(t)}} < \mu_i, n_{m,i}(t-1) \geq l) \leq \frac{1}{t^2}.
$$

Meanwhile, we have that

$$
\sum_{t=L+1}^{T} P(\mu_i + 2\sqrt{\frac{C_1 \log t}{n_{m,i}(t-1)}} > \mu_{i^*}, n_{m,i}(t-1) \geq l) = 0
$$

if the choice of l satisfies $l \geq [\frac{4C_1 \log T}{\Delta_i^2}]$ with $\Delta_i = \mu_{i^*} - \mu_i$.

Based on the decision rule, we have the following hold for $n_{m,i}(T)$ with $l \geq [\frac{4C_1 \log T}{\Delta_i^2}]$,

$$
\begin{aligned}
n_{m,i}(T) \leq l + & \sum_{t=L+1}^{T} 1_{\{a_t^m = i, n_{m,i}(t) > l\}} \\
\leq l + & \sum_{t=L+1}^{T} 1_{\{\tilde{\mu}_i^m - \sqrt{\frac{C_1 \log t}{n_{m,i}(t-1)}} > \mu_i, n_{m,i}(t-1) \geq l\}} \\
& + \sum_{t=L+1}^{T} 1_{\{\tilde{\mu}_{i^*}^m + \sqrt{\frac{C_1 \log t}{n_{m,i^*}(t-1)}} < \mu_{i^*}, n_{m,i}(t-1) \geq l\}} \\
& + \sum_{t=L+1}^{T} 1_{\{\mu_i + 2\sqrt{\frac{C_1 \log t}{n_{m,i}(t-1)}} > \mu_{i^*}, n_{m,i}(t-1) \geq l\}}.
\end{aligned}
$$

By taking the expectation over $n_{m,i}(t)$, we obtain

$$
\begin{aligned}
E[n_{m,i}(t)] \leq l &+ \sum_{t=L+1}^{T} P(\tilde{\mu}_i^m(t) - \sqrt{\frac{C_1 \log t}{n_{m,i}(t)}} > \mu_i, n_{m,i}(t-1) \geq l) \\
&+ \sum_{t=L+1}^{T} P(\tilde{\mu}_i^m(t) + \sqrt{\frac{C_1 \log t}{n_{m,i}(t)}} < \mu_i, n_{m,i}(t-1) \geq l) \\
&+ \sum_{t=L+1}^{T} P(\mu_i + 2\sqrt{\frac{C_1 \log t}{n_{m,i}(t-1)}} > \mu_{i^*}, n_{m,i}(t-1) \geq l) \\
\leq l &+ \sum_{t=L+1}^{T} \frac{1}{t^2} + \sum_{t=L+1}^{T} \frac{1}{t^2} + 0 \\
\leq l &+ \frac{\pi^2}{3} = [\frac{4C_1 \log T}{\Delta_i^2}] + \frac{\pi^2}{3}
\end{aligned}
$$

where the second inequality holds by using (10), (10), and (10).

Then by the definition of T_{21}, we derive

$$
\begin{aligned}
E[T_{21}|A] &= \sum_{m \in M_H} \sum_{k=1}^{K} \Delta_k E[n_{m,k}(t)] \\
&\leq \sum_{m \in M_H} \sum_{k=1}^{K} \Delta_k([\frac{4C_1 \log T}{\Delta_i^2}] + \frac{\pi^2}{3})
\end{aligned}
$$

where the inequality results from (10).

Consequently, we obtain

$$
\begin{aligned}
(10) &\leq E[T_{21}|A] + |M_H| K l^{T-1} \\
&\leq \sum_{m \in M_H} \sum_{k=1}^{K} \Delta_k([\frac{4C_1 \log T}{\Delta_i^2}] + \frac{\pi^2}{3}) + |M_H| K l^{T-1}.
\end{aligned}
$$

Furthermore, we have

$$
\begin{aligned}
(10) &\leq (c+1) \cdot L + E[T_1|A] + E[T_2|A] \\
&\leq (c+1) \cdot L + \sum_{m \in M_H} \sum_{k=1}^{K} \Delta_k([\frac{4C_1 \log T}{\Delta_i^2}] + \frac{\pi^2}{3}) + |M_H| K l^{T-1} + 0
\end{aligned}
$$

which completes the proof.

□