# A Government Fund Allocation Mechanism Based upon Credibility Games

Thesis Submitted to

**Tsinghua University**

in partial fulfillment of the requirement

for the professional degree of

**Master of Finance**

by

**Kong Ruoyan**

Thesis Supervisor： Professor Michael R. Powers

**April, 2018**

# Abstract

In the past, the central government estimated the demands of local governments based on its history demand or expenditure, which will incentivize local governments to over-report their demands for the fund and cause wasteful uses of funds. How to design a government fund-allocation mechanism which can incentivize local governments to report their demands for fund honestly and uses fund appropriately to avoid wastes in fiscal expenditure is a problem left to be solved.

There are five challenges in designing a government fund-allocation mechanism. Firstly, how to estimate local governments' credibility without the knowledge of actual demand for the fund. Secondly, how to allocate fund according to the estimation of credibility, such that the allocation result can incentivize local governments to report their demands honestly. Thirdly, how to allocate fund according to different features of different areas and different uses of the fund. Fourthly, how to make the fund-allocation mechanism incentive compatible, such that the game between local governments will converge to an equilibrium where every local government would report their demands for fund honestly. Fifthly, how to balance the exploration and exploitation actions in this mechanism, such that the mechanism will cause a low burden to the society in the process of converging.

To solve the problems above, this paper builds a dynamic fund-allocation mechanism in Chapter 3. This mechanism is built on the techniques of algorithmic game theory, optimal mechanism design, and multi-armed bandit algorithms. This paper concludes the history, results and implementation, and research areas of the three fields above in Chapter 2.

This dynamic fund-allocation mechanism has the following five features. Firstly, it includes a dynamically learning mechanism on the credibility of local governments, where the credibility of a government will be estimated based on the data of other governments and a UCB-style indicator. Secondly, based on the estimation of credibility, the mechanism offers a fund-allocation plan which considers the randomness of data, such that the truth-telling local governments will not be deprived of the opportunity of receiving fund because of randomness. Thirdly, it collects the features of different areas and different uses of the fund and relates the features with actual demands through a linear model. Fourthly, it builds a game model between local governments and builds an incentive compatible mechanism which can induce the game such that it will converge to an equilibrium where every local government tells the truth. Fifthly, it offers a multi-armed bandit framework to balance the exploration and exploitation of this mechanism.

This work proves that under this mechanism, the precision of estimation on credibility can achieve O(1/n). This work also proves that under this mechanism, the game on the credibility indicator between all the local governments will have an $O(\sqrt{\ln T/T})$-Bayesian Nash Equilibrium, where all the local governments will not over-report their demands at this equilibrium.

**Key words:** funding allocation; optimal mechanism design; incentive compatible; Bayesian game; multi-armed bandit algorithm

# Contents

# Chapter 1 Introduction

The idea of this paper sources from an interesting phenomenon happened near my home. Each time when it comes to the end of that year, the road near my apartment would be demolished and rebuilt by the government. There is no reason to conduct such a construction project each year -- the road was extremely well-functioned each time, the project caused a lot of money, and the construction project would cause serious traffic jams and a lot of inconvenience for citizens. What is the incentive for the local government to conduct such an unnecessary and unbeneficial project each year? After studying the government funding allocation policy, I got the answer – as there was some construction funding left at the end of this year, if they don’t use up all the construction funding, not only will they need to return the surplus funding back to the central government, but also they will get less money in the coming years – because the central government will think that they just don’t need such a lot of money. As a consideration that the more the better, they don’t want to lose this part of the potential money and decide to use up the funding every year (though the project they invest in may be unnecessary). Then I began to wonder if there exists a funding allocation mechanism for the central government to allocate funding to local governments, such that local governments will have no incentives to over-report their demand or exaggerate their demand for funding through some unnecessary projects.

This paper offers a mechanism to solve the problem I talked above. The mechanism is an optimal-designed mechanism with a game on credibility simulated between local governments on a multi-armed bandit framework. The game on credibility incentivized local governments to compete on the degree of telling the truth. The optimal-designed mechanism makes sure that the mechanism we designed exactly generate the kind of game we need – the game which has an equilibrium where every local government chooses to tell the truth. The multi-armed bandit framework offers a method to make a tradeoff between exploration and exploitation in the whole process and also offers a UCB1 index to measure the credibility of each local government.

The rest of this paper is organized as the following: Chapter 1 introduces the funding allocation mechanism and the related research in the past; Chapter 2 introduces the three major techniques that we used in this mechanism – algorithmic game theory, optimal mechanism design, and multi-armed bandit algorithms; Chapter 3 extends the detail of this government funding allocation mechanism and offers a poof for its efficiency and feasibility. Chapter 4 concludes the contribution of this work and gives the possible future work based on this government funding allocation mechanism.

## 1.1 Resource Allocation Mechanism

Resource allocation mechanisms refer to the mechanism that decides whether allocate resource to an entity or not and how many resources to allocate. This kind of mechanisms is widely used in many areas, such as allocating resource in university systems to induce better performance from higher education institutions, allocating market funding in P2P systems, allocating spectrum resources to wireless networks, and allocating funding to green power programs or other public goods dominated by governments.

The goals of resource allocation mechanisms can be various. In research funding allocation problem, the goal can be reflecting the performance of tenure faculties as precise as possible [1]. In health resource allocation problem, the goal can be maximizing the service levels offered to patients [2] or improving the equity of allocating intra-regional health resource. In allocating funding aimed for public affairs like education or infrastructure, the goal can be minimizing the costs of providing equal education opportunities [3] or maximizing the cost-beneficial ratio when deciding whether to build a road or not [4].

A lot of studies showed that government funding tend to be fiscal mismanaged and abusive [5], and the goal of this paper will be to design a mechanism that can allocate the government funding efficiently and ruling out the over report behavior in this process.

The basis method has long been a major method used to make fiscal budgets in China [8], which means that the central government will assess the funding that local governments need based on the money they spent in the last year. However this method will cause a phenomenon called *Rush Spending,* which refers to that if local governments have surplus funding at the end of the year, instead of saving the money, they will choose to spend the money on some unnecessary projects such as fixing a well-functioning road, or painting a totally new government building. The incentive for this phenomenon is very simple: assume that a local government got 10 million dollars this year and it have only spent 8 million at the end of the year, if it uses up the rest 2 million dollars, then the money it will receive next year will be 10 million or more because the central government may think it still needs more money. However, if it keeps the rest 2 million dollars, not only will it need to submit the 2 million dollars back to the central government, but also it will only get 8 million next year because the central government may think 8 million dollars will be enough for this local government. This phenomenon caused huge waste in fiscal funding each year, actually according to the data of the Ministry of Finance of China, in 2017, all the local governments spent 1.96 trillion dollars in the last month of the year, approximately 1/9 of the total year budget of 2017. From Table 1.1 we can find out that approximately 1/6 ~ 1/9 fiscal expenditure was spent in the last month of the year.

Table 1.1 The Fourth Season's Fiscal Expenditure's Proportion to the Total Year's Fiscal Expenditure in China

| **Year** | **Fiscal Expenditure Proportion (%)** | | | **Fiscal Expenditure (billion)** | | |
|---|---|---|---|---|---|---|
| | **Oct** | **Nov** | **Dec** | **Oct** | **Nov** | **Dec** |
| 2014 | 6.43 | 8.73 | 16.93 | 829.7 | 1126.6 | 2184.9 |
| 2015 | 7.73 | 9.40 | 14.84 | 1160.6 | 1412.0 | 2229.1 |
| 2016 | 6.13 | 9.89 | 11.37 | 982.9 | 1587.2 | 1824.3 |
| 2017 | 5.21 | 8.11 | 11.29 | 903.8 | 1407.7 | 1958.1 |

Thus the goal of the funding allocation mechanism in this paper will be examining whether a local government declares a true number of its fiscal budget (demand) and should the central government allocate funding to it.

## 1.2 Methods of Building Resource Allocation Mechanism

Many methods have been used to build funding or resource allocation mechanisms. In [10], the author used a reverse iterative combinatorial auction to allocate the resource in Device-to-Device (D2D) communications. In [11], the author designed a distributed and threshold-based algorithm to allocate labor resources to a given task whose demand evolves dynamically over time. Performance based funding are used to link funding levels to some measure of outputs or outcome of students' performance or research in education funding allocation [12]. The author of [13] built a multilateral bargaining model to include valuing certainty in water resource management and resource allocation.

A lot of funding allocation methods are designed for P2P lending systems. A P2P lending system refers to the system that links lenders with borrowers through an online platform. They provide cheaper service than traditional financial institutions and lenders get higher returns compared to saving money in the bank [14]. A major problem in P2P lending is how to allocate the funding collected from lenders to borrowers efficiently, especially allocate to those borrowers who can reward a high return as well as maintain a reasonable risk level at the same time [15]. A traditional method is to build a credit score system and allocate money to the borrowers who has high credit scores [16]. In [17], the author proposed an alternative profit scoring system to solve the funding allocation problem.

## 1.3 Government Funding Allocation

Most nations have 3 levels of government: central government, local government, and regional governments. High-level governments often need to allocate funding and resource to lower level governments. Allocating resources to local governments is very important in providing basic social care for the citizens and enable local governments to offer necessary services for their areas. "Postal Code" is a common method, where the funding allocated to local authorities' will depend on its population and economy [18]. In this case, the mechanism should be designed to satisfy the local population's needs. Funding also needs to be capable of supporting the redistributive system of the welfare state and guarantying a degree of the territorial justice.

A very import goal of the government funding allocation mechanism is to ensure equity, but how to define and how to apply the concept equity is still in dispute. In [21], the author gives the following three goals of spatial equity: need, rights and effort. The three goals cannot be satisfied simultaneously and central governments need to make a tradeoff between these goals. And if the demand is just assessed by some indicators of local needs and if central governments allocate money according to such indicators, local governments will lack flexibility and autonomy in funding utilizing [22].

The following 4 major methods which are used by central governments in different nations to allocate funding:

1) political patronage
2) historical precedent
3) proposing auctions and bids declared by local authorities
4) local expenditure indicators

The formula funding approach are used to decide the weight of combination among the 4 methods above with the goal to make full use of the limited expenditure, and to allocate the limited expenditure with the greatest optimality, and to impose designed incentives on local governments.

Some funding provided by central governments to local governments might have a specific function, like for sustaining health systems, or for paying educational expenditure. And the central governments need to assess the extent of specific needs of local governments. More often than not, central governments may ask local governments to meet certain requisites of services and use funding in services that have national priorities.

In 1966, the Local Government Act of United States built an allocation system based on rate support grants (RSGs). This system included an extended needs factor, a resources factor and a domestic factor aimed to decrease the rate burden of householders. The system allocates funding to local authorities differently with respect to their relative expenditure needs and taxable affairs. Since 1966, the system has been updated for many times. In 1981, the Local Government Act of United States built a block grant system to take the needs and taxable affairs factors into consideration at the same time, but the principle of allocating funding to local governments according to local circumstances has not been changed. And this principle was also adopted by the allocation of National Health Service fund.

Block grants are designed to represent the relative needs and costs of local governments when they need to provide services for their areas. But it is not intended to assess the direct amount needed to provide local services, instead it provides rules for selecting the formulae to use. The central government will ask local governments to achieve certain requisites of service. When the central government wants to control local government expenditure, they will use centrally-determined formulae to let local governments conform to the level of expenditure they want [22].

The formula-based approach is proposed to consider the local factors that affect the demand for government services, but which may not be completely controlled by local governments. The formula-based approach aims to provide enough resources to achieve either the same level on service for citizens in different areas, or the same level on potential to achieve the same service [23]. It should be noticed that achievable outcomes are influenced by demands. The amount of funding needed to generate a unit of outcome (or the potential to achieve outcomes) will need to be calculated. The formula method is still being improved these days. The idea of equivalence is the potential to provide services, but not to achieve certain outcomes.

No matter the service or outcome standards used, these approaches are built on the correlation between spending and performance induced. In the United States, the formulae approach are used to allocate the total national budget between over 150 local governments, which is proved by some data to be able to give a good result. Normative approaches will include resources being allocated according to some criteria about the designed goals for the system. These criteria can then decide adjustments being added to the relationships between spending and outcomes, which will change the principle of allocating resources between local governments. Actually, this process would possess local governments the resources needed to achieve the outcomes they should have achieved with respect to the normative criteria. This approach is not currently adopted in social care. It might point out a further step compared with the current practice-based formula systems.

# Chapter 2 Basis of Techniques

This chapter will introduce the major techniques used in this paper, including algorithmic game theory, auctions and optimal mechanism design, and multi-armed bandit algorithm. The algorithmic game theory is introduced to simulate the competition on credibility between local governments. Compared to the traditional game theory, algorithmic game theory calculated the complexity of finding equilibriums in a game and can deal with a large amount of data sourced from local governments. Optimal mechanism design provides us the method to design mechanisms where the game and local governments behavior's incentivized by this mechanism can achieve an equilibrium that exactly realizes the goal we impose on this mechanism, no matter what kind of features local governments might have. In other words, it ensures that the mechanism we design will reach an equilibrium where no local government will has the incentive to over report its demand. Auction is a special mechanism which sets rules to allocate items and charge bidders. The auction which doesn't need payment from bidders is called money-burning mechanism, this is very similar to the scenario of government funding allocation processes where local governments don't need to pay the central government. The multi-armed bandit framework provides a method to make a tradeoff between exploitation and exploration. As the government funding allocation has a big influence on the whole social members, we can't simply divide the data set into training set and test set because the funding allocation process is an interactive process and if we try several immature mechanisms in the training period, they may cause big negative impacts on the whole society, for example, the funding might not be allocated to the local government which has an urgent need. The multi-armed bandit framework offers us a dynamic method to gradually improve the mechanism, make the tradeoff between exploration and exploitation, and collect gains as much as possible in the whole allocation process.

## 2.1 Algorithmic Game Theory

### 2.1.1 Introduction

During the latest years, at the interface of game theory, computer science, and economic theory, an increasing number of research were done. This emerging field is largely motivated by the computational ability of the computer and the development of the Internet. Algorithmic Game Theory is the theory that aimed to find out the central ideas and results of this emerging area.

Game theory builds model for scenarios in which many players can interact or affect each other's gain or outcomes induced from this game [30]. Game theory has been widely used in solving problems in many fields, like building the most efficient routing choice to send messages between Internet Service Providers (ISPs), allocating resource in distributed multi-cell OFDMA systems, allocating cloud resource in a cloud computing environment via an imperfect information Stackelberg game and a hidden Markov model, financing and delivering health care and building trust between doctors and patients, preventing hazardous material from entering the nature in closed-loop supply chain (CLSC) management, allocating spatial public goods, allocating funding in transport infrastructure, and allocating goods and payments in auctions.

The game existing among competition for public goods points out the tragedy of commons, where each player will exist a selfish strategy (stable strategy) to try to maximize his or her gain from the game and the total gain of the game will actually be a very limited number. For example, controlling pollution actually brings large costs for a company (like the expenditure on buying new environmental friendly equipment) and it will only get little benefits (from the improvement of the environment), then companies will choose to still produce pollution. When all the companies adopt this negative strategy, they will end up bringing a large negative impact on the whole society.

**Definition** 1 A Simultaneous Move Game with complete information is a game that consists n players {1,…,n}, n possible strategy set ($S_i$ for each player i). Each player i will implement a strategy $s_i \in S_i$ in the game. $s = (s_1, .., s_n)$ is the strategy vector selected by all players and $S = \times S_i$ is the total strategy vector set. Each strategy vector is linked with one payoff function $u_i: S \to R$ that might be different for each player.

A game is interactive, which means that the payoff of each player will not only depend on his own strategy but also on other players' strategies at the same time.

**Definition** 2 A dominant strategy solution is a unique best strategy vector for all players, which is the best choice for each player whatever other players' strategies might be, i.e., strategy vector s satisfies the following inequality, $u_i(s_i, s'_{-i}) \geq u_i(s'_i, s'_{-i})$, $\forall i, s'$.

A dominant strategy is not equal to the strategy that gives the optimal payoff to all players as players may improve their payoffs by changing the strategies of all players at the same time. Mechanism design in the second part of this chapter has the goal to design the games that can have dominant strategy solutions, and also this solutions can lead to good outcomes which satisfy some features we want, like being socially desirable, or profitable for the mechanism designer.

Dominant strategy solutions has very strict criteria, actually we can use a less strict and widely acceptable concept – Nash Equilibrium.

**Definition** 3 A Nash equilibrium is an equilibrium where when all other players don't change their chosen strategies, no player (e.g. player i) will have the incentive to replace his chosen strategy $s_i$ by another strategy $s_i'$ as his payoff can't be improved through this way, i.e. strategy vector s satisfy the following inequality, $u_i(s_i, s_{-i}) \geq u_i(s_i', s_{-i}), \forall i, s_i'$.

It can be noticed that a dominant strategy will be a Nash equilibrium, thus a Nash solution may not be the optimal strategy as a dominant strategy solution may not be the optimal strategy too. A strict dominant strategy will be the unique Nash solution. Within a game, there might be no Nash solution, a single Nash solution or several Nash solutions which will have different payoff to different players in the game.

The Nash Equilibrium Solution we talked above is a pure Nash Equilibrium, where each player may only choose one strategy. Most games don't have a pure Nash Equilibrium. We introduce a less stringent equilibrium – Mixed Strategy Nash Equilibrium. In a Mixed Strategy Nash Equilibrium, players can implement several strategies with probabilities. For example, in the prison dilemma, the players can choose to confess with 0.5 probability and to keep silence with 0.5 probability. We assume that all players are risk-neutral, which means that they want to maximize their expect payoff. And then we can have the famous Nash theorem [33]:

**Theorem 1** A game whose size of set of players and size of set of strategies are finite has a Mixed Strategy Nash Equilibrium solution at least.

The game we talked about above is non-Bayesian game, where players have no private information. However, in reality, many players will hold private information, for examples, in an auction, only the bidder himself or herself knows the true value he or she has for the item; in government funding allocation, only the local governments know the true amount of money it needs to offer sustainable local service. We call such games where players can possess private information the Bayesian game [34].

**Definition** 4 A Bayesian Nash equilibrium is the equilibrium where under a strategy profile and beliefs about the features of the other players, each player's expected payoff cannot be improved if he changed his strategy (mixed strategy) given the beliefs about the other players' types within the same mixed strategies played by the other players.

Different beliefs on other players' types will result in different Bayesian Nash equilibrium.

In reality, the Bayesian Nash equilibrium is still a very stringent equilibrium. Instead, we can replace the Bayesian Nash equilibrium within approximate Bayesian Nash equilibrium in the application of algorithmic game theory, which means that in this equilibrium, for each player, if he replaces his strategy by a new strategy, the gain he can induce from this action will be controlled within a certain boundary if other players don't change their strategies.

### 2.1.2 Complexity of Calculating Nash Equilibrium

The Nash Theorem gives us the important fact that in every finite game there exists at least one mixed Nash equilibrium, and in this situation, no player will have the incentive to change his strategy. However, we need to make sure that this equilibrium can be reached in practice or in acceptable length of time, so we need a stable and convenient algorithm to calculate the equilibrium. Because if the calculation of Nash equilibrium is very complex and time-consuming, it may weaken the Nash equilibrium's power as a prediction of rational players' behaviors. For example, if a player needs to make a decision in one minute but the calculation of the Nash equilibrium strategy needs to consume one day, there is no outstanding possibility for we assume that the player will follow the Nash equilibrium strategy.

We should notice the following clues. Firstly, if each player's mixed strategy is the best response to the mixed strategies of the rest, a mixed strategy profile is a Nash equilibrium. Secondly, that a mixed strategy is the best response is equal to that all pure strategies within this mixed strategy are best responses. Then to find a Nash equilibrium is equal to find a way to combine pure best response strategies to make other players have a support of best responses that can sustain the equilibrium.

Actually finding a Nash equilibrium is the same as solving a combinatorial problem to find a matching support for every player. And most works in the past aimed to find Nash equilibrium via combinatorial methods. But no method until now has been proved to work in polynomial time (P problem).

Finding a Nash equilibrium is not an NP-completeness problem too (the problem which is both NP and NP-hard, NP means the decision problems that can be solved by a non-deterministic Turing machine within polynomial time, an NP-hard problem means that every NP problem can be reduced to this problem within polynomial time). Because by Nash theorem we know that every finite game at least has one Nash equilibrium while traditional NP-complete problem doesn't guarantee the existence of a solution. Suppose that finding a Nash equilibrium is NP-complete and there is an algorithm that can reduce from one NP-complete problem to the Nash equilibrium problem, then there will be a polynomial complexity function which can project the Boolean formula to games such that we can solve the NP-complete problem within polynomial time if and only if the corresponding Nash equilibrium satisfy some polynomial property. But we know that from an unsatisfiable formula, we can guess a Nash equilibrium and then conduct the check step, which actually means NP=coNP [35].

If we want to utilize Nash Theorem's proof to find a Nash equilibrium solution, we will find that the proof uses Brouwer's fix point theorem ( a continuous function f which projects the n-dimensional unit ball to itself will have a fix point x such that f(x)= x). Unfortunately, Brouwer's theorem has a nonconstructive feature and finding a fixed point is a hard problem too [36].

But we can think from the reverse direction: try to establish a Nash problem as hard as an NP-complete or an NP-hard problem. Actually we know that the following problems are NP-hard: In a two player game, if there exists two Nash equilibriums or not? If there exists one Nash equilibrium where a player has a lower bound on payoff or not? If there exists one Nash equilibrium where has a lower bound on total payoff or not? If there exists one Nash equilibrium where has a lower bound on the number of pure strategies within the mixed strategies or not? If there exists one Nash equilibrium whose mixed strategies include one certain pure strategy or not? If there exists one Nash equilibrium whose mixed strategies does not include one certain pure strategy or not?

The famous algorithm for finding a Nash equilibrium in a two-player game is Lemke–Howson algorithm. Lemke-Howson algorithm is a combinatorial algorithm and it provides an alternative proof for the Nash theorem. The Lemke-Howson algorithm can find n + m different Nash equilibriums in case of different choices of the initial-dropped label, where n is the size of players, and m is the size of strategy profile set.

The homotopy-based approach has the same effect as Lemke-Howson algorithm [50]. The homotopy-based approach selects an arbitrary pure strategy g first and add it to the game G, and gives the player who has that strategy a large number B of payment to play it. Then in the modified game G, the strategy g will be played with probability 1, correspondingly other players will choose his best response to g with probability 1. We let B continuously decrease to 0, then there will exist a path of Nash equilibrium which connects the unique equilibrium of the modified game, to an equilibrium of the game G. Actually the pure strategy g designated to give large payment is corresponding to the initially dropped label in the Lemke-Howson algorithm.

The Lemke-Howson algorithm will introduce a graph. The vertices of this graph will represent sets of inequalities, where all strategies are represented except a certain strategy n. The sets of vertices is actually a finite set of combinatorial objects. All vertices will have one or two edges which are incident upon them representing whether strategy n is represented in the vertex v. If there is another endpoint of the path except the zero point, then this endpoint will be a Nash equilibrium solution of the game.

Though the Lemke-Howson algorithm is simple to implement, the calculation in pivot steps in the algorithm may be exponential with corresponding to the number of pure strategies in the game. Actually, it is PSPACE-complete to find a solution that can be induced from the Lemke–Howson algorithm.

And this kind of problems can be concluded as a PPAD problem, which has the following features:

1) a directed graph with a finite and exponentially large set of vertices

2) it takes less than polynomial time to find out whether a string is a vertex of the graph

3) each vertex has at most one in degree and one out degree

4) it takes less than polynomial time to find out the neighbors of a vertex

5) it takes less than polynomial time to find out the predecessor and successor of a vertex

6) one original source with zero in degree

Then a vertex with zero out degree (a sink) or a non-original source will be a solution of the PPAD problem.

Solving a PPAD problem is through walking through the long path and arriving at one sink (or a non-original source) as fast as possible without rote traversal. Similarly, solving an NP problem means reducing to a solution from the exponential size of candidates without exponential times of search. Actually, P = NP implies PPAD = P because PPAD problem is a subset of NP problem, and also a Nash equilibrium solution can be verified within polynomial time if it was found. Nash problem is actually PPAD-complete.

PPAD-completeness is weaker in the intractability than NP-completeness problem: the result may be closed to P<=PPAD<=NP. And if a PPAD-complete problem could be solved in polynomial time, then all problems in PPAD can be solved in polynomial time too. Besides, since any algorithms for finding fix points (Brouwer's points) will assume the function as a black box, if PPAD=P, then there will be a method that can find fix points by exploring the properties of the function. Thus an efficient algorithm to solve the PPAD-complete problem will have to be very sophisticated in a specific sense.

Recently some coevolutionary algorithms are also suggested to detect all the Nash points of a multi-player normal form game at the same time.

### 2.1.3 Learning Algorithms and Equilibrium

Funding allocation mechanisms need to make decisions repeatedly in an environment with uncertainty. For example, the rainfall level of each local area will change each year, the central government needs to take the variation of rainfall level and also the strategies of local governments into consideration to allocate the funding as efficiently as possible. In this section, we introduce learning algorithms for this problem, which also consider the games and equilibriums between players in this systems.

Given several actions we can adopt at each turn, like the agent to allocate funding, the route to a destination, or the choice in the {rock, paper, scissors} game. At each turn. The algorithm will choose one action like selecting a route for driving (which is slightly different from funding allocation mechanisms where the algorithms normally choose several agents), and also the environment, like the congestion situation of each road) will also move. The algorithm will receive a loss from the difference between the action it chose and the optimal action. This process will be repeated for many times, so the algorithm should be adaptive that can adapt to the change of the environment and players.

Regret analysis is widely used to measure such loss. It sourced from the principle that if we sell our algorithm to a company, the difference between the loss that our algorithm incur and the minimal loss that actually can be induced from a much simpler alternative strategy can be controlled in an acceptable amount $\pi$.

The definition of regret is related to the definition of simple policy. A widely used concept is external regret which is based on the best single strategy (the strategy that always chooses the same action at each turn) in retrospect. External regret is also adopted in comparing the performance of online algorithms and the optimal offline algorithms.

Internal regret (swap regret) is also widely used to consider some simple modifications on online sequence of actions (like if you choose to give funding to government A, you should give that funding to government B instead). This modification rule ask the algorithm to replace action i with action j every time it was chosen. The internal regret will only allow changing one action, and the swap regret allows to change several actions through a mapping.

Assume that the number of selections we can choose from at each turn is N, and at each turn t, an online algorithm H will decide a distribution $\mathrm{p}^{\mathrm{t}}$ as the probability distribution used to choose actions. And the environment will return a loss vector $\mathrm{l}^{\mathrm{t}} \in [0,1]^N$ as the loss that the player will get from each action at time t. Then at turn t the algorithm H will get loss $\mathrm{l}_{\mathrm{H}}^{\mathrm{t}} = \sum_{i=1}^{N} p_i^t l_i^t$, the constant i-th action strategy will get loss $\mathrm{L}_{\mathrm{i}}^{\mathrm{T}} = \sum_{t=1}^{T} l_i^t$ after T turns, the H algorithm will get $\mathrm{L}_{\mathrm{H}}^{\mathrm{T}} = \sum_{t=1}^{T} l_H^t$ after t turns. Let $\Gamma$ be a comparison class of algorithm. The external regret method will try to find an algorithm H whose loss is most close to the best algorithm in $\Gamma$. The distance is measured as the external regret $\mathrm{R}_{\Gamma} = L_H^T - L_{\Gamma,min}^T$. The most widely used comparison space is the space X which contain all the constant algorithms (always choose the same action). Then the minimum loss in this comparison class will be $\mathrm{L}_{\mathrm{min}}^{\mathrm{T}} = \min_{\mathrm{i}} L_i^T$, and the external regret will be $\mathrm{R} = \mathrm{L}_{\mathrm{H}}^{\mathrm{T}} = L_{min}^T$.

When considering modification rules, we can use the comparison algorithm class which contain strategies we want to compare with. Suppose at turn t the original algorithm offer a probability distribution $\mathrm{p}^{\mathrm{t}}$ on the actions, the modification rule F will turn the original distribution $\mathrm{p}^{\mathrm{t}}$ into a new distribution $\mathrm{f}^{\mathrm{t}} = F^t(p^t), f_i^t = \sum_{j:F^t(j)=i} p_j^t$ on actions. The modified algorithm's loss will be $\mathrm{L}_{\mathrm{H,F}} = \sum_t \sum_i f_i^t l_i^t$. Then considering the space of modification rules M, the regret of H will be $\mathrm{R}_{\mathrm{M}} = \max_{F \in M} L_H^T - L_{H,F}^T$. In external regret, comparing with the constant algorithm class X is equivalent to consider a N modification rule class. And the internal regret is equivalent to consider a N(N-1)-size modification rules class. Swap regret is equivalent to consider a $N^N$-size modification rule class.

Minimizing external regret has been an important topic in this field. An important theorem in competitive analysis offers the conclusion that comparing with the overall optimal sequences, it is impossible to control the regret within some bound [73] [74]. Let $G_{all}$ be the overall algorithm space, we have the following theorem:

**Theorem 2** Let H be an online algorithm, then there is a loss vector sequence with length T such that $R_{G_{all}}$ will be greater than T(1-1/N).

This conclusion tell us avoiding from considering all algorithms because the regret can't be controlled in this case. In practice, we usually choose the constant algorithm class $G_a = \{g_i : i = 1, \ldots, N\}$ as a comparision.

A widely used algorithm is the greedy algorithm that at each turn it will select the action who has the minimal cumulative loss $L_i^{t-1}$. And in this case we can control loss within $L_{Greedy}^T \leq NL_{min}^T + (N-1)$. This bound is very weak as it only guarantee its loss will be less than N times the loss of the best action. And it has been proved that a constant algorithm will also have similar weakness, for any constant algorithm D, we can find out a loss sequence that satisfy $L_D^T = T, L_{min}^T = \lfloor T/N \rfloor$ (we assume that the loss will be either zero on one). The pure greedy algorithm will have a deterministic tie breaker, which means that it will only select one best action. Randomized Greedy algorithm overcomes this weakness by allowing the algorithm distributes the probability of being selected over several best actions. Comparing with the pure greedy algorithm, the randomized greedy algorithm has a significant performance improvement as the difference in loss between the optimal algorithm and the randomized greedy algorithm is only a O(logN) factor instead of O(N). For any loss sequence, the randomized greedy algorithm's loss can be controlled within $\ln N + (1 + \ln N)L_{min}^T$. By allocating weight to the actions which is currently near the best actions, we can prevent the big loss when the size of optimal selections are very big in the randomized greedy algorithm. Based on this idea, the randomized weighted majority algorithm (Littlestone and Warmuth algorithm) will change the weight of the actions with respect to its distance to the current optimal action. It will make a tradeoff between the exploration and current knowledge. For any loss sequence, the loss of randomized weighted majority algorithm can be controlled within $(1+\eta)L_{min}^T + \frac{\ln N}{\eta}$, where $\eta$ is the random coefficient and we assume that the loss will be either zero or one and T is known. “Guess and Double” approach can deal with the situation when T is unknown.

The polynomial weights algorithm is based on the randomized weighted algorithm to deal with the loss with [0, 1] interval instead of {0,1}. The polynomial weights algorithm changed the update step of the randomized weighted algorithm in the weight of total loss. The loss of the polynomial weights algorithm can be controlled within the boundary $\mathrm{L}_{\mathrm{k}}^{\mathrm{T}} + \eta Q_k^T + \frac{\ln(N)}{\eta}$. This result is very near optimal. We have the following theorem to show that if T is much smaller than N, it is not possible to induce sublinear regret.

**Theorem 3** If T<$\log_2 N$, there will be a way to generate losses stochastically such that $\mathrm{E}\left[\mathrm{L}_{\mathrm{R}}^{\mathrm{T}}\right] = \frac{T}{2}, L_{min}^T = 0$ for any online algorithm R. More specifically, if N=2, we could find a stochastic loss sequence such that $\mathrm{E}\left[\mathrm{L}_{\mathrm{R}}^{\mathrm{T}} - L_{min}^T\right] = \Omega(T)$.

This two results tell us that a $\mathrm{L}_{\mathrm{k}}^{\mathrm{T}} + \eta Q_k^T + \frac{\ln(N)}{\eta}$ boundary in the Littlestone and Warmuth algorithm is very near to the minimum regret we can get.

Normally an equilibrium can be achieved if all the players in a game follow a sublinear swap-regret strategy or other kinds of regret. Given a two-player constant sum (the sum of all player’s gain is a constant number) game G $=< \{1,2\}, (\mathrm{X_i}), (s_i) >, s_1(x_1, x_2) + s_2(x_1, x_2) = c$, where x is the strategy and s is gain, c is some constant. The well-defined value (there exists some strategies that player’s loss is at most this value no matter what strategies other players implemented) of the game is $(\mathrm{v}_1, v_2)$. Then if each player follow the regret-minimization algorithm H, and denote the external regret as R, in a two-player constant some game the average loss $\frac{\mathrm{L}_{\mathrm{on}}^{\mathrm{T}}}{T}$ is at most $\mathrm{v_i} + \frac{R}{T}$. When using the Littlestone and Warmuth algorithm, the average loss can be bounded within $\mathrm{v_i} + O(\sqrt{\frac{lnN}{T}})$. This external regret minimization algorithm also provides a method to prove the minimax theorem in the two-player zero-sum game.

Learning algorithms can also be designed to be able to learn to avoid playing strategies that are dominated by other players. This requisite can be satisfied under a sublinear swap regret procedure. Given a game G and a player i who implements a swap regret procedure, denote the swap regret as R and the number of turns as T, then the player i at most put R/eT weight at the set of e-dominated strategies (actions). The randomized weighted majority algorithm and PW algorithm also provide such guarantee, but having low external regret itself is not enough to provide such boundary.

A procedure who achieves low external regret can also be reduced to a procedure who can provide low swap regret. A widely used method is to instantiate N external regret procedures who will give us a probability vector at each turn. Then this probability vector will be used to generate a probability vector p. After receiving a loss vector l, we will return it back to procedure i with weight $p_i$, then procedure i's view on action j's loss will be $\sum_t p_i^t l_j^t$, which is the cost that could be induced if procedure i's probability was put on action j. The probability p is well defined such that the sum of the losses of procedure i will match the overall true losses. A normally way to design the R external regret procedure A is to let it satisfy that for any loss sequence $l^t$ with length T and any actions j, j=1,…,N, we have $L_A^T = L_j^T + R$. And algorithm H's swap regret can also be controlled within NR. Given an R external regret procedure, for every swap function F on {1,…,N}, the loss $L_H$ of online procedure H can be controlled within $L_{H,F} + NR$. When using the Littlestone and Warmuth algorithm, we can find an online algorithm H whose loss can be controlled within $L_{H,F} + O(N\sqrt{TlogN})$.

In the past, only a constant number T of the number of the turns was considered, and T is independent of the best action's performance, which is called zero-order bounds. Recently the first-order bound is studied more and more, which refers to the bound that will depend on the loss of the optimal action, and also the second bound which will depend on the sum of squares of the losses. It's unknown that how to get an external regret which will be proportional to the optimal action's empirical variance. How to skip the step to collect prior information in the regret minimization algorithm is also being studied [75].

The size of the pool of actions N was also be assumed to be small enough that they can all be listed, and the algorithms complexity is in proportional to N. However, sometimes N will be very large (like in government funding allocation problem, there may be a lot of local government to allocate funding compared to the time T), we need to find computationally efficient algorithms to solve this problem. The author in [76] gave an efficient algorithm when the set of actions is a subset of $R^n$ and the loss vectors l are $R^n$'s linear functions.

## 2.2 Auctions and Optimal Mechanism Design

### 2.2.1 Introduction

Mechanism design is widely used in resource or funding allocation problems. A mechanism refers to an economic and computational system where individuals are rational and try to optimize their strategy to achieve some selfish goals (e.g. a local government wants to find strategy of reporting demand to maximize the amount of funding received from the central government). The system will combine the strategies and actions of individuals and provide some outcomes (optimal mechanism design will optimize this outcome). This kind of mechanisms has good performance in many fields, including the student assignment mechanism which matches schools to students to make parents satisfied, the incentive mechanism in mobile phone sensing to incentivize users' participation, auction mechanisms which allocate goods to bidders and charge for the goods allocated, the computer system mechanism which allocates computational resources in a multi-agent software system, cloud computing mechanism design which selects appropriate cloud vendor and make dynamic pricing, and green pricing programs which aimed to allocate funding to renewable energy programs.

The economic factors in mechanism design are major sourced from the value function (preferences) of agents and the overall performance of the mechanism. For example, bidders in an auction want to maximize the value they can gain from bidding, while the performance of the mechanism or the performance from the perspective of the seller can be measured based on the total payments it collects.

The computational factors in mechanism design are major sourced from the agents who need to optimize their strategies, and the system which needs to satisfy some rules. For example, bidders in an auction must decide their bid strategies, and the auctioneer must decide a strategy to choose a winner and the rules of how to charge bidders. A single item auction may have simple computation (e.g. an auction for a painting). A lot of auctions may involve difficult computation (e.g., in the Federal Communications Commission (FCC) auctions, the bidders have a high level of complementarities).

The agents' strategy space is very complex and the mechanism's rule space can be very complex too. Finding the theoretical optimal individual strategy or mechanism rules can be very time-consuming and space-consuming in a complex environment and will lead to complex results. However, in the real world, the best agents' strategies or mechanism rules, more often than not, are very simple. This phenomenon may be caused by the limited capability to explore the mechanism rules and strategies for agents or to maintain reasonable robustness in different environments. Then the central consideration in mechanism design will be keeping the balance between optimality and other properties like robustness, simplicity, and computational tractability. Approximation theory can describe this tradeoff by measuring the decrease in performance of a simple approximation mechanism with compared to the complicated optimal mechanism. The theory points out that good mechanisms should provide a simple start for the agents.

A traditional problem in mechanism design is the single-item resource allocation problem. This problem needs to choose a single optimal agent to allocate only one item, like to allocate the item to the agent who has the highest value for this item.

**Definition** 5 Single-item resource allocation problem is the kind of allocation problem which has the following features:

1) a single item
2) n agents who use some bidding strategies to compete for the item
3) each agent can get value $v_i$ if he or she receives the item

The principle of solving single-item resource allocation problem is to maximize the social surplus.

**Definition** 6 Social surplus refers to the sum of the value of the auctioneer and agents collected from the allocation result.

In single-item resource allocation problem, the social surplus will be maximized if we allocate the item to the agent who has the highest value for the item, denoted $v_{(1)}$. The difficulty of designing the mechanism to achieve this goal is that we usually don't have information about the value of each agent, otherwise we could simply allocate the item to the agent who possess the highest value for the item. A simple idea on this problem is to let agents report their values through the following procedure:

1) let each agent i report his or her value $b_i$ for the item
2) select agent $i^*$ who has the higest $b_i$
3) allocate the item to $i^*$

However, this simple mechanism will incentivize dishonesty, as the higher number $b_i$ you report, the higher probability that you will get the item, thus the item may not be allocated to the agent who holds the highest value for the item and the social surplus cannot be maximized (social-inefficient). For example, agent 1 values the item as 5 dollars and agent 2 values the item as 10 dollars. If agent 1 reports 15 dollars and agent 2 reports 10 dollars, then the item will be allocated to agent 1 and the social surplus will be 5 dollars instead of the optimal value 10 dollars – the agent who over-reports wins. To solve this problem, we need to build a mechanism that can incentivize agents to tell the truth and then we can maximize the social surplus. In this case we will introduce money (pay mechanism) to the current mechanism, which is called auctions.

### 2.2.2 Auction

**Definition** 7 An auction is a dynamic allocation mechanism that charges payments to agents for bidding.

In an auction, if the payments is positively correlated to the agents' bids, it could keep low-valued agents from making fake high bids.

**Definition** 8 A single-item auction is an allocation solution to the single-item resource allocation problem that sets up bids, selects a winner to give the item (allocation), and determines the payments of the winner (pricing).

A natural way to build a single-item auction is charging the bidder the value he or she bids – the first price auction.

**Definition** 9 A first-price auction is a single-item auction that implements the following steps:

1) for each agent i, let agent i report his or her value $b_i$
2) select agent $i^*$ who has the highest reported value $b_i$
3) allocate the item to the agent $i^*$
4) charge agent $i^*$ the value of his or her bid $b_i$

However, the first-price auction is not a truth-telling mechanism. On the one hand, the agent who wins the auction will want to pay as little as possible (actually he or she wants to guess $v_{(2)}$ and reports a value which is close to $v_{(1)}$), then he or she will tend to bid a lower value. On the other hand, if the value that the agent bids are too small, then he or she probably lose the auction. And as the agents' strategies will not be reporting true values, the agent who has the highest value for the bid may not be selected. Then this mechanism is not truth-telling or social-efficient.

Nobel Prize Winner Professor William Vickrey proposed a solution to achieve the truth-telling and social-efficient goal: simulating the ascending-price auction (the auction that gradually arise the price until only 1 agent left and charge him or her for the second price) with sealed bids.

**Definition** 10 The second price auction (Vickrey auction) is a single-item auction that implements the following steps:

1) for each agent i, let agent i report his or her sealed bids value $b_i$
2) select agent $i^*$ who has the highest value $b_i$
3) allocate the item to the agent $i^*$
4) charge agent $i^*$ the value of the second highest bid $b_{(2)}$

Then the truthful bidding will be a dominant strategy in the second price auction. Social surplus can also be maximized in the dominant strategy equilibrium of second price auction.

Some mechanisms do not include monetary payments, like the funding allocation mechanism where agents just receive the money, and congestion control mechanisms where adding payments in the Internet would need to build financial infrastructure which is very difficult both in money saving and computation. This kind of mechanism is called as money-burning mechanisms.

In the Internet mechanism design, one solution for the money burning mechanism has been implemented for allocating computational resource, such as offering some non-monetary payments as “proofs of capability”. The value of the agent’s message needs to be proved through some way in such a procedure, for example, making his or her computer perform a verifiable, complex, but may be worthless computational task. Computational payments caused utility lost in the society because computational payments will not be transferred to the auctioneer, compared with the monetary payments. So such mechanism may not be adopted in the funding allocation mechanism.

Besides social surplus, residual surplus is another measure to assess a mechanism. The residual surplus is defined as the total value the agents get minus the payments collected. The residual surplus in a second price auction is $\mathrm{v}_1 - \mathrm{v}_{(2)}$. Second price auction does not maximize the residual surplus.

Second price auction may also be very slow and complicated too. A simpler approach is to set a take-or-leave price – uniform pricing auction.

**Definition** 11 The uniform pricing auction will allocate the item to the first agent who is willing to pay $\hat{v}$.

The price $\hat{v}$ needs to be carefully chosen. In some case we can assume that there will be prior knowledge known on the distribution of the value (for example, in Internet routing cases there will be a large number of messages to be transferred every moment which can provide useful information). Assume that agent i’s value will follow a cumulative distribution function F, then a method to choose $\hat{v}$ is to simulate the result of the second-price auction for many times.

### 2.2.3 Mechanism Design

Mechanism design aims to design protocols, laws, or some rules of interaction, like finding a common social choice in elections, the allocation of goods and money in markets, choosing the winner of an auction, making a single social choice in the government policy [38].

A principle in the mechanism design is that the mechanism can drive selfish behaviors (each agent will try to maximize his or her gain individually) to generate a good outcome. We call the selfish behavior rational behaviors. Agents play a game with rational behaviors and the equilibrium will generate the outcome we want to induce, like to maximize the social surplus, residual surplus or the total payments paid to the mechanism. We can also include other rules in the mechanism design. The widely-considered feature in mechanism design includes:

1) Predictability: The observed practical behavior of agents should follow the prediction of the mechanism.

2) Tractability: The rules of the mechanism should be calculated with reasonable time, such as polynomial time.

3) Custom: The mechanism will give a standard to measure the performance of a mechanism.

4) Information: It will consider the features of the real environment.

Optimality is not necessary. Optimal mechanism design is a normal optimization problem: under game theoretic strategizing's incentive constraints, environment's practice constraints, and agent preferences' distributions, try to optimize a certain objective. Incentive constraints and practice constraints in ideal environments may be very simple and the optimization problem will be very simple too. For example, finding the optimal set of messages to transfer is equivalent to the result of raising a second price auction. But most environments and objectives are very complex and it will be very complex to design a simple mechanism.

A central problem in mechanism design is to find the social choice where the majority of the voters will agree with the choice. Allocating funding to the local governments is actually making a social choice too. Condorcet's Paradox tells us that a simple majority vote mechanism cannot guarantee finding out the social choice. Arrow's theorem tells us that over a set of more than 2 candidates, every social welfare function that satisfies unanimity and independence of irrelevant alternatives will exist a dictator (dictatorship refers to the mechanism that one social member in the society has the power to decide the ending social choice). Gibbard–Satterthwaite's theorem clarifies that any incentive social compatible function with a candidate set larger than two will be a dictatorship too. Gibbard–Satterthwaite's theorem makes the possibility of designing an incentive compatible social choice function become very tiny, so the mechanism design tries to avoid this impossibility by modifying the model.

If the mechanism includes monetary payments, the mechanism will make the social choice and decide the payment at the same time. The social choice will be represented through the payment of money. Each agent's preference function is represented by a value function $\mathrm{v_i}: A \to R, v_i \in V_i$, where A is the candidate pool of social choices.

An important goal in the mechanism design is to incentivize the agents to tell the truth, for example, the central government will want the mechanism has the ability to let the local governments report their true demands for the funding instead of over reporting the demand. We use the concept incentive compatibility to describe this feature:

**Definition** 12 A mechanism $(\mathrm{f}, \mathrm{p}_1, \dots, p_n)$ is incentive compatible if for every agent i, every value function $\mathrm{v}_1, .., v_n$, social choice $a = f(\mathrm{v_i}, v_{-i}), a' = f(v_i', v_{-i})$, we have $\mathrm{v_i}(a) - p_i(v_i, v_{-i}) \geq v_i(a') - p_i(v_i', v_{-i})$, where f is the social choice function and $\mathrm{p_i}$ is the payment function.

This feature implies that agent i who has value $\mathrm{v_i}$ will prefer to tell the truth $\mathrm{v_i}$ to the mechanism rather than reporting any fake value $\mathrm{v_i'}$ because his or her gain won't be improved by this replacement.

In a general mechanism without money, any nontrivial mechanism won't be incentive compatible as we discussed above. But if we includes money in the mechanism and use the social welfare as the social choice function, there may be incentive compatible mechanism. We introduce the Vickrey-Clarke-Groves mechanism first:

**Definition** 13 A Vickrey-Clarke-Groves (VCG) mechanism is a mechanism that contains the following features:

1) the social choice function f maximizes the social welfare $\sum_{\mathrm{i}} v_i(a)$

2) there exists function $\mathrm{h_i}: V_{-i} \to R$ for bidder i satisfies that $\mathrm{h_i}$ is independent with $\mathrm{v_i}$ and $\mathrm{p_i}(v_1, \dots, v_n) = h_i(v_{-i}) - \sum_{j \neq i} v_i(f(v_1, .., v_n))$

Actually the VCG mechanism requires that each bidder needs to pay to the mechanism the sum of other bidders' values. And $\mathrm{v_i} - (-\sum_{j \neq i} v_i(f(v_1, .., v_n)))$ represents the total social welfare brought by $f(v_1, .., v_n)$, and this mechanism actually will maximize the social welfare when all bidders have the incentive to tell the truth. Then a VCG mechanism can be an incentive compatible mechanism.

In a VCG mechanism the part $h_i(v_{-i})$ will not influence bidder i's decision on strategy because it's not relevant, but it will influence the amount of the payment. A widely-used method to set $h_i(v_{-i})$ is the Clarket pivot rule: let $h_i(v_{-i}) := \max_{b \in A} \sum_{j \neq i} v_i(b)$. Under the Clarke pivot payments, the VCG mechanism will charge bidder i the value that other bidders could have induced without bidder i. And if $\mathrm{v_i}(a) \geq 0, \forall v_i \in V_i, a \in A$, the VCG mechanism will be individually rational, which means that the bidders will always get nonnegative utility. The second price auction (Vickrey auction) we discussed above belongs to the VCG mechanisms with Clarke pivot rule too. As finding the bidder who has the highest value for the item is actually maximizing $\sum_{\mathrm{i}} v_i$ as only one $\mathrm{v_i}$ will not be equal to zero.

Actually any general mechanisms can be converted to incentive compatible mechanisms. There is a famous revelation principle:

**Theorem 4** If an arbitrary mechanism's dominant strategy is f, then there will be a corresponding incentive compatible mechanism that also implements f and its payments will be equal to the payments at the equilibrium of the original mechanism.

We have only talked about pure mechanisms above, but in many cases players will not have all information about others. The worst case is having no information about the prior distribution. But once we have some information on prior distribution, we can build a Bayesian mechanism and optimize the solution with respect to the information like prior distributions a player has.

**Definition** 14 A Bayesian mechanism is a mechanism that has the following features:

1) n players
2) n type space $\mathrm{T_1}, \ldots, T_n$
3) n prior distributions on the type space: $\mathrm{D_1}, \ldots, D_n$
4) n action spaces $\mathrm{X_1}, \ldots, X_n$
5) an outcome space A
6) player's values functions $\mathrm{v_i}: T_i \times A: \rightarrow R$
7) an outcome function $\mathrm{a}: \mathrm{X_1} \times \ldots \times X_n \rightarrow A$
8) n payoff functions $\mathrm{p_i}: X_1 \times \ldots \times \mathrm{X_n} \rightarrow R$

Besides truthful mechanisms we talked above, mechanisms are designed to achieve some goals, and a kind of important goal is to maximize profits (optimal mechanism design). The goal of this mechanism is to maximize the profit of auctioneer which equals to $\sum_{\mathrm{i}} p_i - c(x)$ while keeping that mechanism truthful at the same time, c(x) is the cost of producing x for the auctioneer. In a single item auction we set c(x)=0 to represent that there is only a single item to be allocated. If we know some prior information about the bidder's values, we can design more profitable mechanisms than the Vickrey auction whose auctioneer's profit will be $v_{(2)}$.

The mechanism which optimizes the Bayesian mechanism such that the values of players will comply with a prior distribution is called a Myerson mechanism: a truthful mechanism that maximizes the expected profit of the auctioneer over the supposed distribution of the players' values.

In order to find out the Myerson mechanism, the concepts virtual valuations and virtual surplus are introduced:

**Definition** 15 $v_i \in [0,h], \forall i,$ $F_i$ is the distribution function of bidder i's value, the density function is $f_i(z) = \frac{d}{dz}F_i(z)$, the virtual valuation of player i is defined as $\phi_i(v_i) = v_i - \frac{1-F_i(v_i)}{f_i(v_i)}$, the virtual surplus of allocation x is defined as $\sum_i \phi_i(v_i)x_i - c(x)$.

The surplus of allocation x is equal to $\sum_i v_i x_i - c(x)$, it can be seen that the virtual surplus is the projection of original surplus with replacing valuations with virtual valuations.

After introducing virtual valuation and virtual surplus, it will be very convenient to calculate the profit as it can be proved that the expected virtual surplus of a truth mechanism is equal to its expected profit. Then the task becomes finding truthful mechanisms. We have the following theorem:

**Theorem 5** The virtual surplus maximization mechanism is truthful if and only if $\phi_i(v_i)$ is monotone and non-decreasing with respect to $v_i$.

Then we will have the following algorithm to calculate Myerson's optimal mechanism Myef(b):

1) calculate virtual bids $b_i' = \phi_i(b_i)$ for the bids b and F
2) set up a VCG auction on virtual bids $b'$ to get allocation $x'$ and payment $p'$
3) return allocation $x = x'$ and payment $p_i = \phi_i^{-1}(p_i')$

To maximize the social surplus of a single item auction, we will allocate the item to the bidder with the highest value and the auctioneer won't keep the item. But results may be different in maximizing the virtual surplus as if a bidder has positive valuation but negative virtual valuation, we won't allocate the item to the bidder. Then at some circumstances the auctioneer will keep the item. We have the following theorem:

**Theorem 6** Under values drawn from distribution F, the optimal single item auction is the Vickrey with the reservation price $\phi^{-1}(0) = VA_{\phi^{-1}(0)}$.

The mechanism we talked above includes money. Actually in some cases there might be no money and no payment (like in the funding allocation problem, the local governments will just receive funding from the central government but need not to pay for the funding). In this case, as we talked at the beginning of this section, this mechanism will need the players use some way to prove their valuation without transferring money to the auctioneer. The mechanism which burns arbitrary payments and maximize residual surplus (the total value the players get minus the sum of burnt payments) is called as monetary burning mechanisms. In [39], the author proved that for the agent who wishes to maximize the total utility, the payments they make to the auctioneer should be efficiently burnt under a weak control condition. As the optimization problem in the grand coalition is similar to the money-burning setting problem of the auctioneers, these weak cartel results will be similar to the results in a multiunit auction problem [40]. Through an ironing procedure we can transform a possible non-monotone virtual function in money burning mechanisms into an monotone ironed virtual valuation function, and the optimal allocation for the ironed virtual valuation is identical to the optimal allocation in the original virtual valuations [41]. Given cost function c and valuation profile v, valuations' distribution F, iron virtual valuation function $\mathrm{u_i}$, the mechanism that satisfies the allocation rule that $\mathrm{x(v)} \in \mathrm{argmax}_{x'} \sum_i u_i(v_i)x_i - c(x')$ and $\frac{\mathrm{d}}{\mathrm{dv_i}} u_i(v_i) = 0 => \frac{d}{dv_i} x_i(v_i) = 0$ will be optimal from the expected residual surplus perspective. And the money-burning mechanism will be O(1+ $\log_{\mathrm{n}} k$) times larger in surplus compared to the corresponding k-unit auction.

Another important type of mechanism is applied in strategic learning [42] – the incentive-compatible experimental design, which refers to designing experiment or learning mechanisms to evaluate treatments, which are played by selfish agents who want to get the highest evaluation in the experiment or learning process. This scenario is also existing in funding allocation mechanisms, where the local governments want to let the central government believe that they have reported the true demand for the funding. This mechanism induces a game between agents (like between local governments) where each agent may make a selection in multiple treatment it plays. And the action of one agent will have an influence on the actions of all other agents (strategic interference). In the incentive-compatible experiment, an agent will select his or her natural strategy to maximize its performance in the evaluation process if there was free of competition. To set up this kind of incentive compatible mechanism, one needs to find a certain statistic to assess agents' performance first, like to build a rating model for the outcome and asymptotic results of the maximum-likelihood estimator, and also calculate the covariance matrix of the statistic. Then it needs to find out a transformation function f to satisfy the incentive compatible equation. Possible transformation functions might be identity function, reciprocal function and the function needs to satisfy the ratio between the denominator and nominator in the incentive compatibility equation.

Strategic learning is applied in many area. In [43], the author builds a bandit framework to acquire more precise data dynamically from strategic data resources which might have large variations such as the data collected from crowdsourcing platforms like Amazon Turk. This paper gives us the hint to explore the idea of introducing a bandit framework in the funding allocation mechanism to have a more precise understanding of the demand collected from local governments. In [42], the author tried to use this kind of mechanism to evaluate two marketing agents like advertisement companies and to select the company who has higher efficiency in viral marketing. In [44], the author simulates the effects of risk taking and the responsiveness of market pioneering. In [45], the author proposed a strategic learning mechanism that can estimate the parameters in a linear regression while protects the privacy of agent at the same time. In [46], the author builds a strategic learning mechanism combined with a multi-armed bandit framework to estimate the quality of user-generated content and incentivize attention-motivated agents who can get benefits when the content they generated is displayed with a relatively lower cost to generate high quality content at the same time.

## 2.3 Multi-armed Bandit Problem

In this paper we will explore a multi-armed bandit algorithm to calculate a score that represents whether local governments tended to report the true demand for the funding in the history. Thus we introduce the basis of multi-armed bandit problem in this section first.

### 2.3.1 Introduction

Multi-armed bandit problems was introduced by Robbins in 1952 [47]. Multi-armed bandit problems were widely used to describe situations where the agent will need to make a compromise between gaining new knowledge via exploring the environment and exploiting the knowledge he or she already got. There are many problems of this type in the real world, like adapting to the intermediate results of advertising tests and deciding the percentage of impression brought by each advertisement [48], allocating treatments in clinical trials [49], allocating capacity in the channelized dynamic spectrum routing problem [50], directing discrete dynamic vehicle [51], online experiment design [52]. Compared to reinforcement learning which is also used to deal with the tradeoff between a dynamic problem's exploration and exploitation, the multi-armed bandit algorithm offers a simple and clear method to describe the trade-offs (unlike a black box).

A bandit problem normally contains K probability distributions $(\mathrm{D}_1, \dots, D_K)$ and the expected values of these distributions $(\mu_1, \dots, \mu_K)$ and also their variances $(\sigma_1^2, \dots, \sigma_K^2)$. A player of this game has no knowledge on the parameters of distributions and wants to find out which distribution has the highest expected value and also collect as much reward as possible at the same time. These distributions are viewed as arms of a slot machine and the player will selects an arm j(t) in each turn t. Then the player will receive the reward $\mathrm{r(t)} \sim \mathrm{D}_{\mathrm{j(t)}}$. The multi-armed bandit algorithms will offer strategies that lead the player to choose an arm j(t) at each turn t.

The normal assumption in a multi-armed bandit problem is that the random variable reward of each arm are independent and identically distributed. For each selection i, there will be a distribution $\mathrm{D}_\mathrm{i}$ where reward will be generated from, and the reward will be simulated from this distribution independently. The algorithm wants to find out each distribution $\mathrm{D}_\mathrm{i}$ during T turns.

The simplest form of reward distribution is Bernoulli reward, whose value can only be zero or one ("failure" or "success"). The distribution can be completely represented by the mean reward or the probability of a one reward. The problem above can be described only by the number of turns T and the mean of rewards.

A normal measure to estimate the performance of a multi-armed bandit algorithm is total expected regret, which is defined as $\mathrm{R_T} = T_{\mu^*} - \sum_{t=1}^{T} \mu_{j(t)}$ for T turns, where $\mu^* = \max_i \mu_i$ is the mean of the arm which has the highest expected reward. The regret can also be written as $\mathrm{R_T} = T_{\mu^*} - \sum_{k=1}^{K} \mu_k E(T_k(T))$, where $T_k(T)$ is the number of selections of arm k during the first T turns.

Robbins [53] gave a famous result about $T_k(T)$ in 1985: $\mathrm{E}\big(T_k(T)\big) \geq lnT / D(p_k||p^*)$ where $D(p_k||p^*)$ is the Kullback-Leibler divergence.

Recent years, many theoretical bounds of multi-armed bandit problem have been proved [55]. Some latest experiments implemented an extensive empirical experiment on the bounds of different multi-armed problems. It is evaluated that how well the UCB family of algorithms and some other important algorithms like pursuit or reinforcement comparison can perform in the real environment.

The bandit problem is equal to a Markov decision problem with one state. The zero strategy of a multi-armed problem means that the regret of this arm will convergence to 0 as the number of turns played increases to infinity. The zero-regret strategy will converge to an optimal strategy when the number of turns goes to infinity too.

There are some different formulations of the multi-armed bandit problem. In the binary multi-armed bandit problem, the slot will return a one reward with probability p, and zero otherwise with probability 1-p. Another kind of multi-armed bandit problem contains a Markov process. At turn t when an arm is chosen, the state of the slot will transfer at the next turn t+1. And the selection of the strategy must take the Markov state transition into consideration. The reward of the arm will be dependent on the current state of the slot, which is called as “restless bandit problem” too [63]. There is also a variation of the multi-armed bandit problem where the number of arms that the players can select will increase with the number of turn played [64].

Normally we will divide multi-armed bandit problems into worst-case assumption analyzing, minimizing regret in finite time horizons with stochastic arms’ rewards, infinite time horizon with stochastic arms’ rewards, finite time horizon with non-stochastic arms’ rewards, and infinite time horizon with non-stochastic arms’ rewards.

### 2.3.2 Multi-armed Bandit Algorithms

Many research constructed optimal arm selection strategies which can convergence to the arm with the highest mean and the maximum uniformly convergence rate to solve the multi-armed problem. In [53], the author built the convergent arm selection strategies under the one-parameter exponential family to simulate the arm reward distributions. The author in [65] extend the proof into normal distributions with public variation information. Burnetas et al. then extend the case into situations where the distributions of arms’ rewards will depend on an unknown parameter vector. They also constructed a strategy which has the maximum convergence rate [66]. They also gave a solution to find out the optimal arm when the distribution of arms' rewards are drawn from some arbitrary discrete and univariate distributions.

In [67], Burnetas studied the Markov Decision Process model under the partial information case where the expectation of single turn rewards will depend on some unknown parameters. This work constructed an analytical form for the total expected reward in finite horizon with finite number of states and irreducible transitions. The selection of arms is based on the inflations of the right side of the equation of estimated average rewards, which is called as the optimistic approach [68][69].

There are six major kinds of algorithms in multi-armed bandit problems. In the following, $\widehat{\mu_i}(t)$ is defined as the mean of reward of arm i in the first t turns. $\mathrm{p_i}(t)$ is defined as the probability of selecting arm i at turn t.

1) ε-greedy Algorithm

ε-greedy algorithm is very easy to realize. It has wide generalizations for different sequential decision problems, and is widely adopted. The ε-greedy algorithm will always select the arm with the highest empirical mean at each turn. And in order to contain some extent of exploration and randomness, this action will be implemented with probability $1 - \varepsilon$, and with probability ε the algorithm will select a random arm instead. Denote the empirical means of each arm at the beginning of the procedure as $\widehat{\mu_i}(0)$, then the probability is set as:

$$\mathrm{p_i}(t+1) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{\mathrm{k}}, if\ i = argmax_j \widehat{\mu_j}(t) \\ \frac{\varepsilon}{\mathrm{k}}, else \end{cases} \qquad (2\text{-}1)$$

With regard to the performance of this algorithm, if ε is a constant number during the turns, then the expected regret will only achieve a linear bound. In [56], the author improved the algorithm with a decreasing ε and generated a poly-logarithmic bound. But in some recent research, it is found out that this algorithm didn't show outstanding improvements in the empirical experiment.

2) Boltzmann Exploration Algorithm

Boltzmann exploration algorithm is designed based on the theorem of choice of Luce [57]. It picks an arm with a probability which is positively correlated with its average reward. Then high empirical means' arms are selected with relatively high probabilities too. It uses the Boltzman distribution to select the arms. Denote the empirical means of each arm at the beginning of the procedure as $\widehat{\mu_i}(0)$, then the probability is set as:

$$\mathrm{p_i}(t+1)=\frac{e^{\frac{\widehat{\mu_i}(t)}{\tau}}}{\sum_j e^{\frac{\widehat{\mu_j}(t)}{\tau}}}, i=1,\dots,n \tag{2-2}$$

, τ is called as the temperature parameter which can control the degree of randomness. And when τ is equal to 0, the Boltzmann exploration algorithm will be simplified to a ε-greedy algorithm. If τ is big enough, the algorithm can be viewed as selecting each arm with the same probability.

Variations of the Boltzmann exploration algorithm can also achieve polylogarithmic regret bounds, but in empirical studies, there is no outstanding improvement achieved by this algorithm too.

3) Pursuit Algorithm

Instead of selecting arms based on estimating the empirical means of each arm, the pursuit algorithm implements an explicit strategy on selecting the arms [58]. Pursuit algorithms use empirical means to do updates too but the updates will be performed separately. At the beginning of the procedure, the algorithm selects each arm with a uniform probability 1/k. Then at each turn t, it will selects the arm with probability

$$\mathrm{p_i}(t+1)=\begin{cases} p_i(t)+\beta(1-p_i(t)), if\ i=argmax_j\widehat{\mu_j}(t) \\ \quad p_i(t)+\beta(-p_i(t)), else \end{cases} \tag{2-3}$$

, β stands for the learning rate for this algorithm. Actor-critic algorithm, a variation of the pursuit algorithm, is used in solving the sequential decision problems in reinforcement learning. When learning automata, it has a PAC-style convergence rate [59].

4) Reinforcement Comparison Algorithm

The reinforcement comparison algorithm also maintains a distribution over selections which are not calculated directly from the empirical means like what the pursuit algorithm does. Instead, it keeps an average expected reward ar(t). It will select an arm with the probability computed by comparing ar(t) with empirical means. It will increase the probability to choose an arm if its average expected reward is above its empirical mean, and decrease otherwise. The reinforcement learning algorithms is suitable for the case where the means of the arms are very similar.

The algorithm records a set of preferences $\mathrm{\pi_i}(t)$ for each arm. And at each turn t, it will use the Boltzmann distribution to calculate the probability:

$$\mathrm{p_i}(t)=\frac{e^{\pi_i(t)}}{\sum_j e^{\pi_j(t)}}, i=1,\dots,n \tag{2-4}$$

The preference of the selected arm $\pi_{\mathrm{j(t)}}(t)$ will also be updated as the following:

$$\pi_{\mathrm{j(t)}}(t+1) = \pi_{\mathrm{j(t)}}(t) + \beta(\mathrm{r(t)} - \mathrm{ar(t)})$$

The means of the reward will be updated to:

$\mathrm{ar(t+1)} = (1-\alpha)\mathrm{ar(t)} + \alpha\mathrm{r(t)},$

$\alpha$ and $\beta$ are learning rates given.

The difference between the algorithms above is the different opinions they have on dealing with the tradeoff between exploration and exploitation. The principle that the number of selections of an arm needs to be positively correlated with the empirical mean of that arm is utilized in the Boltzmann exploration algorithm. The principle that keeping arms' explicit probability distributions and searching the candidate pool of probability distributions directly is utilized in the pursuit algorithm. Though these methods offer simple approaches to handle the tradeoff between exploration and exploitation, these methods don't have a statistical mean theoretically and can't be understood analytically. To solve this problem, people designed two algorithms – UCB1 and UCB1-Tuned.

The UCB1 algorithm and UCB1-Tuned algorithm offer detailed mathematical meanings which can provide verifiable theoretical expected regret bounds. The UCB1 algorithm's regret will only have a constant factor difference from the optimal algorithm's regret.

5) Upper Confidence Bounds (UCB) Algorithm

In 2002, Auer et al. proposed a UCB family of algorithms to describe the optimism ideas analytically under uncertainty [60]. In [61], the author proposed a sequential tree algorithm and proved its high efficiency empirically in some playing programs like Go. UCB1 is the simplest algorithm in the UCB family. It will record the number of selections of each arm i before time t as $\mathrm{n_i}(t)$ besides the empirical means. To generate the initial value of empirical means, at the beginning each arm will be played for at least one time.

The UCB1 algorithm is a kind of greedy algorithm, which means that at turn t, it will select arm j(t) with probability 1, j(t) is defined as $\mathrm{j(t)} \coloneqq \mathrm{argmax_i}(\hat{\mu}_i + \sqrt{\frac{2lnt}{n_i}})$.

It can be proved that UCB1's expected regret can be bounded by $8\sum_{\mathrm{i:\mu_i}<\mu^*}\frac{lnt}{\delta_i} + \left(1+\frac{\pi^2}{3}\right)\sum_i \delta_i, \delta_i \coloneqq \mu^* - \mu_i$. This regret bound is O(log n), and Lai et al. proved an $\Omega(\log n)$ bound. Then UCB1's difference on regret compared to the optimal algorithm is only a multiplicative constant factor. It fairly solves the multi-armed bandit problem.

6) Upper Confidence Bounds (UCB1) –Tuned Algorithm

The UCB1-Tuned algorithm is proved to work better than UCB1 algorithm in practice. It has no theoretical bound. Compared with the UCB1 algorithm, the UCB1-Tuned algorithm considers each arm's variance besides its empirical mean. At turn t, UCB1-tuned algorithm will select j(t) as:

$$\mathrm{j(t)} = \mathrm{argmax_k}(\widehat{\mu_i} + \sqrt{\frac{lnt}{n_i}\min\left(\frac{1}{4}, V_i(n_i)\right)}) \tag{2-5}$$

and $\mathrm{V_i}(t) = \widehat{\sigma_i}^2(t) + \sqrt{\frac{2lnt}{n_i}}$.

Variance $\widehat{\sigma_i}^2(t)$ can be estimated by recording the empirical sum of the reward's squares besides the empirical mean. In [62], the author provided the variance-based UCB algorithm's regret bounds, which is very similar to the UCB1-Tuned algorithm.

In the following we discuss the detail of the lower bound of the regret in multi-armed bandit algorithms. There is an initial conclusion first:

**Theorem 7** Given the number of turns T and the number of arms K, there is a mean distribution instance such that $\mathrm{E[R(T)]} \geq \Omega(\sqrt{\mathrm{KT}})$.

This bound is met in the worst scenario and it's fairly possible that an algorithm may have a regret lower than this bound. To achieve this goal, an important concept KL-divergence was introduced,

KL-divergence is widely used in information theory. In a space $\Omega$ which has finite samples and two probability distributions p, q, the Kullback-Leibler divergence is defined as $\mathrm{KL(p,q)} = \sum_{\mathrm{x}\in\Omega} p(x)\ln(\frac{p(x)}{q(x)}) = E_p[\ln\left(\frac{p(x)}{q(x)}\right)]$, which can be viewed as the distance between two distributions. It satisfies the Gibb's Inequality Chain rule, Pinsker's equality, Random coins, but it's not symmetrical. KL-divergence is actually the expectation of the log-likelihood ratio of finding out several samples which belongs to distribution p or distribution q.

Given a coin which gives 1 with probability u and 0 with probability 1-u, $\mathrm{u} \in \{\mathrm{u_1}, u_2\}$, we want to find out the estimated value of u. In a bandit problem with K arms, each arm represents a biased coin with mean which is private information. We want to find out the arm which has the highest mean after T turns as precise as possible, denote the prediction on the arm as $\mathrm{y_T}$. Then we can get that for number of turns T and number of arms K, in a bandit algorithm, choose an arm from a uniform distribution, then $\mathrm{E[R(T)]} \geq \Omega(\sqrt{\mathrm{KT}})$.

Another important lower bound of regret complements the log(T) upper bound in UCB1. It states a log(T) regret with an instance-dependent constant. We have the following theorem:

**Theorem 8** The regret of an algorithm has a lower bound $\mathrm{o}(\mathrm{c}_{\mathrm{I}}\log(t))$ for instances I, $\mathrm{c}_{\mathrm{I}}$ is a constant which only depends on instance I.

This theorem implies that there exists at least one instance that has a high regret. But if we want to find out the upper bound for all instances, we need to guarantee the performance of the algorithm to some extent, we use the following theorem:

**Theorem 9** If a multi-armed bandit problem has K arms and $\mathrm{E[R(t)]} \leq \mathrm{O}(\mathrm{C}_{\mathrm{I},\alpha}t^{\alpha})$ for each instance I and $\alpha$. For an arbitrary problem instance I, there will exist a time t' such that for all t>=t', $\mathrm{E[R(t)]} \geq \mathrm{C}_{\mathrm{I}}\mathrm{ln}(t)$.

We can also define the Bayesian regret [72] if we know the posterior distribution on arms as $\mathrm{p}_{\mathrm{t}}(a) = P(a = a^{*}|H_{t})$, where a* is the arm with the highest empirical reward for the instance and $\mathrm{H}_{\mathrm{t}}$ is the history information of reward before turn t. The Thompson algorithm [71] explores a distribution $\mathrm{p}_{\mathrm{t}}$ to simulate an arm $\mathrm{a}_{\mathrm{t}}$. If a* is a random variable which is dependent on the problem instance, then $\mathrm{a}_{\mathrm{t}}{\sim}a^{*}$ if $\mathrm{H}_{\mathrm{t}}$ is fixed. We want to prove an upper bound on Thompson sampling's Bayesian regret, which is defined as $\mathrm{E}_{\mathrm{u\sim prior}}[E_{rewards}[R(t)|u]]$. It should be noticed that a normal regret bound for all problem instances also holds in case of the Bayesian regret. There is a major result about the Bayesian regret:

**Theorem 10** In a bandit problem with over k arms, $\mathrm{BR(T)} = \mathrm{O(KTlog(T))}$

### 2.3.3 Incentive Compatible Multi-armed Bandit Algorithms

As we talked above, the central goal in mechanism design is to set allocation and payment rules to achieve some optimization goals, like maximizing social surplus, residual surplus, or the profit of the auctioneer. An important principle to design the mechanism to achieve these goals is to set payment rules that can incentivize players to tell their true value to the mechanism (truthful telling mechanism).

Multi-armed bandit algorithms provide a method to estimate preference when values are unknown by making people show their values for an item through the interaction with mechanism [70], like an auction mechanism in keyword auction is bidding on a set of keywords and the bidders can learn the value of the advertisements through the expected click-through rate (CTR). But at the beginning the advertiser didn't know the click through rate (the auctioneer didn't know neither) and he or she needs to estimate CTR in the following process. The auction mechanism should also take the position of slots into consideration as at each turn many slots are allocated to different advertisers. Actually in the keyword auction of Google, they designed a mechanism to decide the prices to create a truthful mechanism like the second price auction we talked about in the former section.

Suppose there are k bidders in an incentive compatible multi-armed bandit problem for the advertisement auction. Bidder i wants to put one advertisement and get a private value $v_i$ for every click. The multi-armed bandit mechanism will solicit bids from every bidder first and will have T turns. The mechanism will select one bidder at each turn and put his or her advertisement on the slot and get an empirical click rate. Payments will be decided after T turns. Each bidder will try to maximize his or her utility, which equals to the value collected from these clicks minus his or her payments. The mechanism should also incentivize each bidder to bid truthfully under the incentive of maximizing his or her utility.

# Chapter 3 The Government Funding Allocation Mechanism with a Multi-armed Bandit Framework

## 3.1 Introduction

This chapter is aimed to build a government funding allocation mechanism to rule out the extra money requests from the local governments under a multi-armed bandit framework. This chapter's theories prove the feasibility and efficiency of the mechanism. From the proofs we get the upper boundary of extra gain that the local governments can induce if they over report that their true demand through a repeated Bayesian game. We provide a detailed scenario for this government funding allocation mechanism as follows.

Every year / month / week, the central government will have some funding to be allocated to local governments according to the demands that the local governments declared and the feature data of the local governments that the central government can observe. For example, at the beginning of each year, the central government will receive the reports from local governments declaring how much money they may need to fix the buildings broken by water disasters this year (or the money they declared that they have spent in fixing buildings broken by water disasters in the last year). The central government can also observe the rainfall during the rainy seasons in the past years. Then the central government will pay the local governments the number they declared for them to prepare to fix the broken buildings during this year.

However, this mechanism has an obvious flaw – the local governments will be incentivized to declare a number larger than their true demands (over report) because of corruption, or fearing of receiving less money in the coming years if they don't report big enough numbers this year, etc.

To solve this problem, we introduce a multi-armed bandit frame mechanism to introduce an extra step to the original mechanism – at the beginning of each turn, for each local government, we estimate to how much extent it had told the truth in the history and give scores to each local government according to the estimation of its history performance on credibility. Then we will set a standard to select the local governments who have high scores to give funding in each turn. As different cities or states may have different features (for example, some areas may suffer a lot more rainfall frequently than other areas and thus need more money for fixing and rebuilding), the feature data collected in the past years will also be concluded in this mechanism as a source of information to help the central government judge whether a local government tended to over report its demand in the past.

We proved that under this mechanism, there will be an approximate Bayesian Nash Equilibrium (BNE) such that at this BNE, every local government will have no incentive to over report its demand.

The rest of this chapter will be organized as follows: in Section 3.2 we give the assumption of the general model and the statistically identical model, in Section 3.3 we give the algorithm of the mechanism, in Section 3.4 we prove the feasibility and efficiency of the mechanism.

## 3.2 The Assumptions of the Model

**The General Model**

We have 1 auctioneer (the central government) and $\mathrm{N}$ bidders $\mathrm{U} = \{1, \dots, \mathrm{N}\}$ (local governments). The mechanism will run for $\mathrm{T}$ times $(\mathrm{t} = 1, \dots, \mathrm{T})$. At each turn t, the auctioneers will observe feature $\mathrm{X_i(t)}, i = 1, \dots, N$.

$\{\mathrm{X_i(t)}\}_{\mathrm{i=1,t=1}}^{i=N,t=T} \sim Uniform(B^d(0,1))$, so we assume that features are independent (e.g. the rainfall of this year/ this city will be uncorrelated with the rainfall of the past years / other cities, time-dependent features such like population should be avoided).

Each $\mathrm{X_i(t)}$ is associated with a true demand $\mathrm{d_i(X_i(t))}$, which cannot be observed directly by the auctioneer. Let $\mathrm{d_i}\big(\mathrm{X_i(t)}\big) = a_i'\mathrm{X_i(t)},\ a_i \in R^d,\ \big||a_i|\big| \leq M,\ a_i$ is bidder i's private information. Other bidders will assume that $a_i \sim N(a, \epsilon),\ a \in R^d$.

At the beginning of each turn $\mathrm{t}$, the auctioneer will select $\mathrm{s(t)}$ (the set of bidders to give funding) according to their scores.

After begin selected, bidder $\mathrm{i}$ will declare $\widetilde{d_i}\big(X_i(t)\big) = a_i'\mathrm{X_i(t)} + f_i(t), f_i(t)\ \epsilon\ [-Z, Z]$, and its profit will be $f_i(t)$.

**The Statistically Identical Model**

We assume that $a_1 = \cdots = a_N = a$, but the auctioneer doesn't know the exact value of $a$. So he has to estimate $a$ according to the history data.

## 3.3 The Algorithm for Statistically Identical Case

Let $\tilde{a}_{-i}^{t}$ be the linear square estimator trained using the data from bidders $\{j\}_{j\neq i}$ up to time t-1. Let

$$I_i(t) \triangleq b_1\sqrt{\frac{\ln(t)}{n_i(t)}} + R_i(t) \qquad (3\text{-}1)$$

$$R_i(t) \triangleq \frac{1}{n_i(t)}\sum_{n=1}^{t-1} 1\,(i \in s(n))[b_2 - b_3(\tilde{d}_i(X_i(n)) - \tilde{a}_{-i}(t)'X_i(n))] \qquad (3\text{-}2)$$

, where $n_i(t)$ is the times of being selected of bidder i before turn t.

**Lemma 1** $b_2, b_3$ can be chosen to ensure $R_i(t)$ bounded and positive.

*Proof:*

$$\begin{aligned} & |\tilde{d}_i(X_i(n)) - \tilde{a}_{-i}(t)'X_i(n))| \\ \leq \quad & ||a' - \tilde{a}_{-i}(t)'||_2 \ (Holder\ Inequality) \\ \leq \quad & ||a'||_2 + ||\tilde{a}_{-i}(t)'||_2 \ (Minkowski\ Inequality) \\ \leq \quad & 2M^{\frac{1}{2}} \end{aligned} \qquad (3\text{-}3)$$

Then choose $b_2, b_3$ s.t. $b_2 > 2M^{\frac{1}{2}}b_3$. We finish the proof of **Lemma 1**.

**Algorithm UCB-I** Bidder Selection

**Step 1** For each bidder i, train the linear estimator $\tilde{a}_{-i}(t)$ by data $\{\mathrm{X_j(n)}\}_{\mathrm{j\neq i\ and\ j\in s(n),n=1}}^{n=t-1}$(the data collected from all the bidders except bidder i before turn t).

**Step 2** For each bidder i, calculate $I_i(t)$ at the beginning of turn t.

**Step 3** Select the bidders in $\mathrm{s(t)}$ to give funding at time t, where

$$s(t) := \{j: I_j(t) \geq \max_i I_i(t) - e(t)\} \qquad (3\text{-}4)$$

$$e(t) := O(\sqrt{\frac{ln(t)}{t}}) \qquad (3\text{-}5)$$

$e(t)$ is the noise added to avoid rejecting competing bidders. It was set as the maximum error induced by the second part (exploration part) of $I_i(t)$ in order to avoid rejecting bidders who had a good performance in telling the true value in the past but still gets a little bit lower score because of the form of the index. We also assume $\mathrm{N} \gg \mathrm{T}$, for example, we have a lot of local governments of cities, villages that we select $|\mathrm{s(t)}| > \mathrm{T}$ for each time.

## 3.4 Feasibility and Efficiency of UCB-I

Let $\begin{matrix} f(t) := & \{f_1(t), \ldots, f_N(t)\} \\ f_{-i}(t) := & \{f_j(t)\}_{j\neq i} \end{matrix}$.

**Definition 16** $\{f_i(t)\}_{i=1,t=1}^{N,T}$ is a $\rho-$BNE if

$$\frac{1}{T}E[\sum_{t=1}^{T} f_i(t)\, 1(i \in s(t))|\{f(n)\}_{n\le t}] \ge \frac{1}{T}E[\sum_{t=1}^{T} \tilde{f_i}(t)\, 1(i \in s(t))|\{\tilde{f_i}(n), f_{-i}(n)\}_{n\le t}] - \rho \tag{3-6}$$

$\forall i, \{\tilde{f_i}(n)\}_{t=1}^{T}$

**Theorem 11** Under UCB-I, if the running mechanism of UCB-I is a public information to all the bidders, then there exists an $O\left(\sqrt{\frac{\ln(t)}{t}}\right)-BNE$ such that $f_i(t) = 0, \forall i \in s(t), t$.

We will prove from the following logic: Firstly, $f_i(t) < 0, \forall i \in s(t), t$ will be an unbeneficial strategy as it will only add limited number of selections compared to the $f_i(t) = 0, \forall i \in s(t)$ case and will receive an overall negative gain. Secondly, $f_i(t) > 0, \forall i \in s(t), t$ will be an unbeneficial strategy too as the number of selections will decrease dramatically in this case (there will be great probability to be judged as lying). Thirdly, $f_i(t) < 0$ for some i and $f_i(t) > 0$ for others will be an unbeneficial strategy too (through proving some general forms of theorems).

Firstly we want to prove that for every bidder i, the expectation of the second part of UCB-I score will be very close to a certain value which we view as an exact value of its credibility if we have enough samples to estimate the second part, and the value will be negatively correlated with the over reporting amount $f_i(t)$. We want to prove this because if the UCB-I score is negatively correlated with $f_i(t)$, we can induce that there will exist a competition on the level of over reporting amount $f_i(t)$ among the bidders. Thus we prove the following lemma first.

**Lemma 2** For each turn t, denote the number of i.i.d samples to calculate $\tilde{a}_{-i}(t)$ as n, then

$$|E[R_i(t)] - (b_2 - b_3 f_i(t))| \le O(\frac{1}{n}), \forall i, t \tag{3-7}$$

$$|R_i^1(t)| \le O(\frac{1}{n}) \tag{3-8}$$

, with probability $1 - e^{-Cn}$ for some constant $C > 0$.

*Proof:*

$$\tilde{d}_i(X_i(n)) - \tilde{a}_{-i}(t)'X_i(n) = (a' - \tilde{a}_{-i}(t)')X_i(n) + f_i(t) \tag{3-9}$$

We define $\begin{matrix} u_{i,t}^1(n) & := (a' - \tilde{a}_{-i}(t)')X_i(n) \\ u_{i,t}^2(n) & := f_i(t) \end{matrix}$, and $R_i^k(t) := -b_3 \frac{\sum_{n=1}^{t-1} 1(i\in s(n)) u_{i,t}^k(n)}{n_i(t)}$, then we have

$$R_i(t) := b_2 + R_i^1(t) + R_i^2(t) \qquad （3\text{-}10）$$

Firstly,

$$|u_{i,t}^1(n)| = |(\tilde{a}_{-i}(t)' - a')X_i(n)| \leq \ ||\tilde{a}_{-i}(t)' - a'||_2(Holder\ Inequality) \qquad (3\text{-}11)$$

Then we have the following lemma to ensure the closeness of $\tilde{a}_{-i}(t)'$ to $a'$ once we have enough samples:

**Lemma 3** Denote the number of i.i.d samples to calculate $\tilde{a}_{-i}(t)$ as n, then $||\tilde{a}_{-i}(t)' - a'||_2 \leq Z(d+2)\frac{1+z}{n(1-z)^2}, z \in (0,1)$, with probability $1 - e^{-Cn}$ for some constant $\mathrm{Z, C,}$ z.

By **Lemma 3** we can induce that

$$|R_i^1(t)| \leq b_3||\tilde{a}_{-i}(t)' - a'||_2 \leq b_3 Z(d+2)\frac{1+z}{n(1-z)^2} \qquad （3\text{-}12）$$

, and $E[u_i^2(t)] = f_i(t)$.

Then we have $R_i(t) = b_2 - b_3 f_i(t) + \mathrm{O}(\frac{1}{\mathrm{n}})$ and $|E[R_i(t)] - b_2 + b_3 f_i(t)| \leq \mathrm{O}(\frac{1}{\mathrm{n}})$. We finish the proof of **Lemma 2**.

As the credibility score can be estimated with a high precision and at each turn, the UCB-I algorithm will only choose the part of bidders whose credibility scores are very close to the highest score in the candidate pool, the bidders will have a competition on the level of credibility score in order to be chosen. To win this game and to be selected by the auctioneer, they will be incentivized to lower down the level of extra money that they added to their true demand.

*Proof* of **Lemma 3**:

Suppose we have n i.i.d samples and we represent them by $X \in R^{n\times d}$ with their corresponding declared demand $\tilde{d} \in R^n$, then we can do a linear estimation that $\tilde{a}_{-i}(t) = (X'X)^{-1}X'\tilde{d}$.

Then

$$\begin{aligned} ||\tilde{a}_{-i}(t) - a||_2 = &\quad ||(X'X)^{-1}X'\tilde{d} - (X'X)^{-1}X'd||_2 \\ = &\quad (trace((X'X)^{-1}X'(\tilde{d} - d)(\tilde{d} - d)'X(X'X)^{-1}))^{\frac{1}{2}} \\ = &\quad ||(X'X)^{-1}X'(\tilde{d} - d)(\tilde{d} - d)'X(X'X)^{-1}||_2 \\ \leq &\quad ||(X'X)^{-1}||_2^2||X'(\tilde{d} - d)(\tilde{d} - d)'X||_2 \end{aligned} \qquad （3\text{-}13）$$

As we want to bound 3−13, we utilize a famous theorem in random matrix analysis:

**Theorem 12** Denote $||\cdot||$ as the spectral norm, $\mathrm{z} \in (0,1), \mathrm{t} \geq 1$, if $\{\mathrm{X_i}\}_{\mathrm{i}=1}^n$ have d dimension and is generated i.i.d from $Uniform(B^d(0,1))$, then when $n \geq C(\frac{t}{z})^2(d+2)log(d)$ for some constant C and z,

$$||X'X|| \leq \frac{1+z}{2+d}n$$

$$||(X'X)^{-1}|| \le \frac{1}{(1-z)\frac{n}{d+2}}$$

, with probability $1-\mathrm{d}^{-\mathrm{t}^2}$ (Corollary 5.52 in [77]).

Then we can induce that

$$\begin{aligned}
||\tilde{a}_{-i}(t)-a||_2 &\le \Big(\frac{1}{(1-z)\frac{n}{d+2}}\Big)^2\big(trace(X'(\tilde{d}-d)(\tilde{d}-d)'X)\big)^{\frac{1}{2}} \\
&\le \Big(\frac{1}{(1-z)\frac{n}{d+2}}\Big)^2 Z(\textstyle\sum X_i'X_i)^{\frac{1}{2}} \\
&\le Z\Big(\frac{1}{(1-z)\frac{n}{d+2}}\Big)^2 ||X'X||_2 \\
&\le Z\Big(\frac{1}{(1-z)\frac{n}{d+2}}\Big)^2 \frac{1+z}{d+2}n \\
&= Z(d+2)\frac{1+z}{n(1-z)^2}
\end{aligned}$$

, we finish the proof of **Lemma 3**.

After proving that the expectation of the second part of UCB-I score will be very close to a certain value and with enough samples, we can estimate the bidders' credibility scores precisely, we want to use this conclusion to prove that if every bidder add no extra money to their true demand, the total number of selections of each bidder will be very close to t at each turn (which means that a truth-telling bidder will fairly be chosen to be allocated funding at each turn), or the following lemma:

**Lemma 4** If every bidder does not over report his or her true demand ($\mathrm{f_i(t)} = 0, \forall \mathrm{t}$), then $\mathrm{E[n_i(t)]} \ge t - C$ for some $\mathrm{C}$.

To prove **Lemma 4**, we first need to prove that if every bidder does not over report his or her true demand, the number of being selected of any bidder cannot be greater than another with a constant ratio, or the following lemma:

**Lemma 5** If every bidder does not over report their true demand ($\mathrm{f_i(t)} = 0, \forall \mathrm{t}$), then there exists $\sigma > 0$, s.t. $\mathrm{n_i(t)} \le (1+\sigma)n_j(t)$, $\forall \mathrm{i} \ne \mathrm{j}$, with probability $1 - \mathrm{O}(1/\mathrm{t}^2)$.

*Proof:*

If at any time the condition $\mathrm{n_i(t)} \le (1+\sigma)n_j(t)$ is not satisfied, we can assume that there exists some time t, at time the condition changes to $\mathrm{n_i(t)} > (1+\sigma)n_j(t)$, then we need to bound $\Pr\big[I_i(t) \ge I_j(t)\big]$ first. To finish the proof, actually we want the probability to be controlled in $\mathrm{O}(\frac{1}{\mathrm{t}^2})$. Firstly, we have that

$$Pr[I_i(t) \geq I_j(t)] \qquad = Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}}]$$

$$\leq Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{(1+\sigma)ln(t)}{n_i(t)}}]$$

$$\leq Pr[R_i(t) - R_j(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}}]$$

Denote $\mathrm{Z}_1 = b_3 Z(d+2)\frac{1+z}{(1-z)^2}$ in **Lemma 2**, then with probability $1 - e^{-C\sum_{k\neq i} n_k(t)}$, $|R_i^1(t)| \leq \frac{Z_1}{\sum_{k\neq i} n_k(t)}$ and with probability $1 - e^{-C\sum_{k\neq j} n_k(t)}$, $|R_j^1(t)| \leq \frac{Z_1}{\sum_{k\neq j} n_k(t)}$. Notice that we have $\sum_{k\neq i} n_k(t) \geq \sum_{n=1}^{t} |\, s(n)| - t \geq tT - t = t(T-1) \geq t^2$ and $\sum_{k\neq j} n_k(t) \geq \sum_{n=1}^{t} |\, s(n)| - t \geq tT - t = t(T-1) \geq t^2$. Then we conclude that with probability $1 - e^{-Ct^2}$, $max\{|R_i^1(t)|, |R_j^1(t)|\} \leq \frac{Z_1}{t^2}$.

Then as we have $E[R_i^2(t)] = E[R_j^2(t)]$,

$$Pr[R_i(t) - R_j(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}}]$$

$$\leq \qquad Pr[R_i^2(t) - R_j^2(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}]$$

$$\leq \qquad Pr[R_i^2(t) - E[R_i^2(t)] \geq \frac{\sqrt{1+\sigma} - 1}{2} b_1\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2}]$$

$$+Pr[R_j^2(t) - E\left[R_j^2(t)\right] \leq -\frac{\sqrt{1+\sigma} - 1}{2} b_1\sqrt{\frac{ln(t)}{n_i(t)}} + \frac{Z_1}{t^2}]$$

We notice that $-b_3 Z \leq R_i^2(t) \leq b_3 Z$, then via Hoeffding inequality [6] we have

$$Pr[R_i^2(t) - E[R_i^2(t)] \geq \frac{b_1(\sqrt{1+\sigma} - 1)}{2}\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2}]$$

$$\leq \qquad exp(-\frac{2(\frac{b_1(\sqrt{1+\sigma} - 1)}{2}\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2})^2 n_i(t)^2}{4b_3^2 Z^2 n_i(t)})$$

$$\leq \quad exp(-2\frac{\left(\left(\frac{b_1\left(\sqrt{1+\sigma} - 1\right)}{2}\right)^2 \ln(t)\right)}{4b_3^2 Z^2})exp(4\frac{\frac{b_1\left(\sqrt{1+\sigma} - 1\right)}{2}\sqrt{\frac{\ln(t)}{n_i(t)}}\frac{Z_1 n_i(t)}{t^2}}{4b_3^2 Z^2})\exp(-\frac{\frac{2Z_1^2}{t^4} n_i(t)}{4b_3^2 Z^2})$$

$$\leq \qquad \frac{1}{t^2}$$

when $\sigma, b_1, t$ are large enough such that $\left(\frac{b_1(\sqrt{1+\sigma}-1)}{2}\right)^2 \geq 4b_3^2Z^2$ and $4\frac{\frac{b_1(\sqrt{1+\sigma}-1)}{2}\mathrm{Z}_1}{4b_3^2Z^2} \leq 1$.

Similarly $Pr[R_j^2(t) - E\left[R_j^2(t)\right] \leq -\frac{\sqrt{1+\sigma}-1}{2}b_1\sqrt{\frac{\ln(t)}{n_i(t)}} + \frac{Z_1}{t^2}] \leq O(\frac{1}{t^2})$.

Then we conclude that $Pr[I_i(t) \geq I_j(t)] \leq O(\frac{1}{t^2})$.

As at time t it changes to $\mathrm{n_i(t)} > (1+\sigma)n_j(t)$, then $\exists\, \mathrm{t}_- < t < t_+$ s.t. $\mathrm{n_i(t_-)} \leq (1+\sigma)n_j(t_-)$ and $\mathrm{n_i(t_+)} > (1+\sigma)n_j(t_+)$ , $\mathrm{n_i(t)} \geq (1+\sigma)n_j(t) - 1 \geq (1+\sigma-1)n_j(t)$, when $\sigma$ is large enough such that $Pr[I_i(t) \geq I_j(t)] \leq O(\frac{1}{t^2})$.

At time $t_-$, i is selected but not j, otherwise the ratio $\frac{\mathrm{n_i(t)}}{(1+\sigma)n_j(t)}$ will go down, so we can induce that $I_i(t_-) \geq I_j(t_-)$. We discuss in the following 2 cases:

Case 1 $t_- \in [\frac{t}{2}, t]$

Then $Pr[I_i(t_-) \geq I_j(t\_)]$ can be upper bounded by $O\left(\frac{1}{\left(\frac{t}{2}\right)^2}\right) = O(\frac{1}{t^2})$.

Case 2 $t_- \in [0, \frac{t}{2})$

Then consider the bidder who was selected for most times during $[\frac{t}{2}, t]$, denote the bidder as bidder k.

If $\mathrm{n_k(t)} \leq (1+\sigma)n_j(t)$, then $n_j(t) \geq \frac{t}{2N(1+\sigma)}$, then

$$\begin{aligned} n_i(t)/n_j(t) &\leq \frac{t/2 + (1+\sigma)n_j(t)}{n_j(t)} \\ &\leq \frac{t/2}{\frac{t}{2N(1+\sigma)}} + (1+\sigma) \\ &= (1+\sigma)(N+1) \\ &= 1 + (N + (N+1)\sigma) \end{aligned}$$

, in this case we can redefine $\sigma = N + (N+1)\sigma$.

If $\mathrm{n_k(t)} > (1+\sigma)n_j(t)$ , then $\exists\, \mathrm{t}^*$ and $\mathrm{t}^* > \frac{t}{2}, n_k(\mathrm{t}^*) \geq (1+\sigma)n_j(\mathrm{t}^*) - 1 \geq (1+\sigma-1)n_j(\mathrm{t}^*)$, when $\sigma$ is large enough this probability can be upper bounded by $\mathrm{O}(\frac{1}{\mathrm{t}^2})$.

We finish the proof of **Lemma 5**.

After proving that if every bidder add no extra money to their true demand, no bidder can be selected more that another with a constant ratio, we can induce the total number of selections of each bidder in the following.

Proof of **Lemma 4**:

From **Lemma 5**, we have that with probability $1 - \mathrm{O}\left(\frac{1}{\mathrm{t}^2}\right)$, $\mathrm{n_i(t)} \geq \frac{t}{(1+\sigma)(N-1)+1}$ and $|\mathrm{R_i}(t) - E[|\mathrm{R_i}(t)]| \leq \frac{b_1(\sqrt{1+\sigma}-1)}{2}\sqrt{\frac{\ln(t)}{n_i(t)}} + \frac{Z_1}{t^2}$, then

$$
\begin{aligned}
& |I_i(t) - b_2| \\
= \quad & |R_i(t) + b_1\sqrt{\frac{\ln(t)}{n_i(t)}} - b_2| \\
\leq \quad & |R_i(t) - E[R_i(t)]| + |E[R_i(t)] - b_2| + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \\
\leq \quad & \frac{\sqrt{1+\sigma}+1}{2} b_1 \sqrt{\frac{((1+\sigma)(N-1)+1)ln(t)}{t}} + 2\frac{Z_1}{t^2} \\
\leq \quad & \frac{\sqrt{1+\sigma}+1}{2} b_1 \sqrt{\frac{(1+\sigma)Nln(t)}{t}} + 2\frac{Z_1}{t^2}
\end{aligned}
$$

Let

$$
e(t) := 2\left(\frac{\sqrt{1+\sigma}+1}{2} b_1 \sqrt{\frac{(1+\sigma)Nln(t)}{t}} + 2\frac{Z_1}{t^2}\right)
$$

then $\Pr\left[I_j(t) \leq \max_i I_i(t) - e(t)\right] \leq \mathrm{O}(\frac{1}{\mathrm{t}^2})$, i.e. $\Pr[j \in s(t)] \geq 1 - \mathrm{O}\left(\frac{1}{\mathrm{t}^2}\right), \forall j, t$, which means that if every bidder add no extra money to their true demand, each bidder is being selected for nearly each time.

And we can induce that if every bidder does not over report their true demand for the funding, the total number of selections of each bidder will be very close to t in each time as the following

$$
E[n_i(T)] = E[\sum_{t=1}^{T} 1\,(i \in s(n))] = \sum_{t=1}^{T} Pr[i \in s(n)] \geq T - O(\sum_{n=1}^{T} \frac{1}{n^2}) \geq T - C
$$

for some constant C. We finish the proof of **Lemma 4**.

After proving that if every bidder add no extra money to their true demand, the total number of selections of each bidder will be very close to t in each time, we want to use this conclusion to find the number of selections of a bidder if he or she does over report its demand for the funding and how much gain he or she can induce from this dishonest strategy. If we can prove that the gain from over reporting in the short run will be weaken by a decrease of number of selections in the long run, we will be very close to the $O\left(\sqrt{\frac{\ln(t)}{t}}\right) - BNE$. Thus before proving the $O\left(\sqrt{\frac{\ln(t)}{t}}\right) - BNE$, we prove the following lemma first.

**Lemma 6** Let $0 < \mathrm{w} < \mathrm{w}^* < 1$, $\delta := \sqrt{\frac{ln(t)}{t^w}}$, there is a bidder i who implements the strategy $f_i(t) = \delta, \forall t$, then bidder i will be selected no more than $\mathrm{O(T^{w^*})}$ in the whole process, i.e. if at time t bidder i has been selected more than $\mathrm{O(T^{w^*})}$ times, then $E[n_i(T); n_i(t) \geq O(T^{w^*})] \leq O(T^w) < O(T^{w^*})$.

*Proof:*

For bidder i, we need to find the bound of its number of being selected, which is equal to limit $\Pr[I_i(t) \geq \underset{j}{max} I_j(t) - e(t)]$, so we only need to consider the bidder i, j who has the highest and second highest UCB-I score $I_j(t)$ and $I_i(t)$ as others will have lower $\Pr[I_i(t) \geq \underset{j}{max} I_j(t) - e(t)]$. In other word, we need to bound $\Pr[I_i(t) \geq \mathrm{I_j}(t) - e(t)]$.

As $R_i^k(t) := -b_3 \frac{\sum_{n=1}^{t-1} 1(i \in s(n)) u_{i,t}^k(n)}{n_i(t)}$ and $u_{i,t}^2(n) = f_i(t)$, for bidder i who implements the strategy $f_i(t) = \delta, \forall t$, we can have that $E[R_j^2(t)] \geq E[R_i^2(t)] + b_3\delta$.

We notice that $Pr[I_i(t) \geq I_j(t) - e(t)] = Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}} - e(t)]$, and $\{R_i(t) + b_1\sqrt{\frac{\ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{\ln(t)}{n_j(t)}} - e(t)\}$ actually implies that at least one of the following inequalities will be meet:

$$R_i^2(t) - E[R_i^2(t)] \geq b_1\sqrt{\frac{ln(t)}{n_i(t)}}$$

$$R_j^2(t) - E[R_j^2(t)] \leq -b_1\sqrt{\frac{ln(t)}{n_j(t)}}$$

$$b_3\delta \leq 2b_1\sqrt{\frac{ln(t)}{n_i(t)}} + e(t) + \frac{Z_1}{\sum_{k\neq i} n_k(t)} + \frac{Z_1}{\sum_{k\neq j} n_k(t)}$$

As otherwise there will exist a contradiction:

$$
\begin{aligned}
& R_j(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}} - e(t) \\
> \quad & b_2 + R_j^2(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}} - e(t) - \frac{Z_1}{\sum_{k\neq j} n_k(t)} \\
> \quad & b_2 + \ E[R_j^2(t)] - e(t) - \frac{Z_1}{\sum_{k\neq j} n_k(t)} \\
\geq \quad & b_2 + E[R_i^2(t)] + b_3\delta - e(t) - \frac{Z_1}{\sum_{k\neq j} n_k(t)} \\
\geq \quad & b_2 + E[R_i^2(t)] + 2b_1\sqrt{\frac{ln(t)}{n_i(t)}} + \frac{Z_1}{\sum_{k\neq i} n_k(t)} \\
\geq \quad & b_2 + R_i^2(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} + \frac{Z_1}{\sum_{k\neq i} n_k(t)} \\
\geq \quad & R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}}
\end{aligned}
$$

We notice that $-b_3 Z \leq R_{i,t}^2(n) \leq b_3 Z$, then via Hoeffding inequality [6] we have

$$
\begin{aligned}
& Pr[R_i^2(t) - E[R_i^2(t)] \geq \mathrm{b}_1\sqrt{\frac{ln(t)}{n_i(t)}}] \\
\leq \quad & exp(-\frac{2\left(\mathrm{b}_1\sqrt{\frac{\ln(t)}{n_i(t)}}\right)^2 n_i(t)^2}{4b_3^2 Z^2 n_i(t)}) \\
\leq \quad & exp(-4\ln(t)\frac{b_1^2}{8b_3^2 Z^2}) \\
\leq \quad & \frac{1}{t^4}
\end{aligned}
$$

when $b_1$ are large enough such that $\frac{b_1^2}{8b_3^2 Z^2} \geq 1$.

After proving $Pr\left[\left|R_i^2(t) - E[R_i^2(t)]\right| \geq b_1\sqrt{\frac{\ln(t)}{n_i(t)}}\right] = O\left(\frac{1}{t^4}\right)$, we will establish $\mathrm{n_i(t)} \leq (1+\sigma)n_j(t)$, then we can get $\sum_{k\neq i} n_k(t) \geq n_j(t) \geq O(T^{w^*})$.

If at any time the condition $\mathrm{n_i(t)} \leq (1+\sigma)n_j(t)$ is not satisfied, we can assume that there exists some time t, at time the condition changes to $\mathrm{n_i(t)} > (1+\sigma)n_j(t)$, then we need to bound $\mathrm{Pr}\left[\mathrm{I_i}(t) \geq I_j(t)\right]$. To finish the proof, actually we want the probability to be controlled in $\mathrm{O}(\frac{1}{\mathrm{t}^2})$. Firstly, we have that

$$\begin{aligned}
Pr[I_i(t) \ge I_j(t)] \quad &= Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \ge R_j(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}}] \\
&\le Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \ge R_j(t) + b_1\sqrt{\frac{(1+\sigma)ln(t)}{n_i(t)}}] \\
&\le Pr[R_i(t) - R_j(t) \ge \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}}]
\end{aligned}$$

Denote $Z_1 = b_3 Z(d+2)\frac{1+z}{(1-z)^2}$ in **Lemma 2**, then with probability $1 - e^{-C\sum_{k\neq i} n_k(t)}$, we have $|R_i^1(t)| \le \frac{Z_1}{\sum_{k\neq i} n_k(t)}$ and with probability $1 - e^{-C\sum_{k\neq j} n_k(t)}$, we have $R_j^1(t)| \le \frac{Z_1}{\sum_{k\neq j} n_k(t)}$. Notice that we have $\sum_{k\neq i} n_k(t) \ge \sum_{n=1}^{t} |s(n)| - t \ge tT - t = t(T-1) \ge t^2$ and $\sum_{k\neq j} n_k(t) \ge \sum_{n=1}^{t} |s(n)| - t \ge tT - t = t(T-1) \ge t^2$. Then we conclude that with probability $1 - e^{-Ct^2}$, $max\{|R_i^1(t)|, |R_j^1(t)|\} \le \frac{Z_1}{t^2}$.

Then as we have $E[R_i^2(t)] = E[R_j^2(t)]$,

$$Pr[R_i(t) - R_j(t) \ge \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}}]$$

$$\le \quad Pr[R_i^2(t) - R_j^2(t) \ge \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}]$$

$$\le \quad Pr[R_i^2(t) - E[R_i^2(t)] \ge \frac{\sqrt{1+\sigma} - 1}{2}b_1\sqrt{\frac{ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}]$$

$$+Pr[R_j^2(t) - E\left[R_j^2(t)\right] \le -\frac{\sqrt{1+\sigma} - 1}{2}b_1\sqrt{\frac{ln(t)}{n_i(t)}} + 2\frac{Z_1}{t^2}]$$

We notice that $-b_3 Z \le R_i^2(t) \le b_3\delta$, denote $U := b_3(z+\delta)/2$, then via Hoeffding inequality [6] we have

$$Pr[R_i^2(t) - E[R_i^2(t)] \geq \frac{b_1(\sqrt{1+\sigma}-1)}{2}\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2}]$$

$$\leq \quad exp(-\frac{2(\frac{b_1(\sqrt{1+\sigma}-1)}{2}\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2})^2 n_i(t)^2}{4\mathrm{U}^2 n_i(t)})$$

$$\leq \quad exp(-2\frac{\left(\left(\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\right)^2 \ln(t)\right)}{4\mathrm{U}^2})exp(4\frac{\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\sqrt{\frac{\ln(t)}{n_i(t)}}\frac{Z_1 n_i(t)}{t^2}}{4\mathrm{U}^2})\exp(-\frac{\frac{2Z_1^2}{t^4}n_i(t)}{4U^2})$$

$$\leq \quad \frac{1}{t^2}$$

when $\sigma, b_1, t$ are large enough such that $\left(\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\right)^2 \geq 4\mathrm{U}^2$ and $4\frac{\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\mathrm{Z}_1}{4\mathrm{U}^2} \leq 1$.

Similarly $Pr[R_j^2(t) - E\left[R_j^2(t)\right] \leq -\frac{\sqrt{1+\sigma}-1}{2}b_1\sqrt{\frac{\ln(t)}{n_i(t)}} + \frac{Z_1}{t^2}] \leq O(\frac{1}{t^2})$.

Then we conclude that $Pr[I_i(t) \geq I_j(t)] \leq O(\frac{1}{t^2})$.

As at time t it changes to $\mathrm{n_i(t)} > (1+\sigma)n_j(t)$, then $\exists\ \mathrm{t}_- < t < t_+$ s.t. $\mathrm{n_i(t_-)} \leq (1+\sigma)n_j(t_-)$ and $\mathrm{n_i(t_+)} > (1+\sigma)n_j(t_+)$, $\mathrm{n_i(t)} \geq (1+\sigma)n_j(t) - 1 \geq (1+\sigma-1)n_j(t)$, when $\sigma$ is large enough such that $Pr[I_i(t) \geq I_j(t)] \leq O(\frac{1}{t^2})$.

At time $t_-$, i is selected but not j, otherwise the ratio $\frac{\mathrm{n_i(t)}}{(1+\sigma)n_j(t)}$ will go down, so we can induce that $I_i(t_-) \geq I_j(t_-)$. We discuss in the following 2 cases:

Case 1 $t_- \in [\frac{t}{2}, t]$

Then $Pr[I_i(t_-) \geq I_j(t_-)]$ can be upper bounded by $O\left(\frac{1}{\left(\frac{t}{2}\right)^2}\right) = O(\frac{1}{t^2})$.

Case 2 $t_- \in [0, \frac{t}{2})$

Then consider the bidder who was selected for most times during $[\frac{t}{2}, t]$, denote the bidder as bidder k.

If $\mathrm{n_k(t)} \leq (1+\sigma)n_j(t)$, then $n_j(t) \geq \frac{t}{2N(1+\sigma)}$, then

$$n_i(t)/n_j(t) \leq \quad \frac{t/2+(1+\sigma)n_j(t)}{n_j(t)}$$

$$\leq \quad \frac{t/2}{\frac{t}{2N(1+\sigma)}} + (1+\sigma)$$ , in this case we can redefine $\sigma = N + (N + 1)\sigma$.

$$= \quad (1+\sigma)(N+1)$$

$$= \quad 1 + (N + (N+1)\sigma)$$

If $n_k(t) > (1+\sigma)n_j(t)$, then $\exists\ t^*$ and $t^* > \frac{t}{2}, n_k(t^*) \geq (1+\sigma)n_j(t^*) - 1 \geq (1+\sigma-1)n_j(t^*)$, when σ is large enough this probability can be upper bounded by $O(\frac{1}{t^2})$.

Then we get $\sum_{k\neq i} n_k(t) \geq n_j(t) \geq \frac{n_i(t)}{1+\sigma} \geq O(T^{w^*})$, when $n_i(t) \geq O(T^{w^*})$.

Because $O(T^{w^*}) \geq lnT \geq lnt$, with probability $1 - e^{cO(T^{w^*})} \geq 1 - O(\frac{1}{t^4})$, we have

$$2b_1\sqrt{\frac{ln(t)}{n_i(t)}} + e(t) + \frac{Z_1}{\sum_{k\neq i} n_k(t)} + \frac{Z_1}{\sum_{k\neq j} n_k(t)} \leq 2b_1\sqrt{\frac{ln(t)}{n_i(t)}} + e(t) + O(\frac{1}{T^{w^*}})$$

And then select t, T large enough to let $e(t) + O\left(\frac{1}{T^{w^*}}\right) < \frac{b_1\delta}{2}, n_i(t) \geq \frac{(2b_1)^2\ln(t)}{\left(\frac{b_3\delta}{2}\right)^2}$, then

$$2b_1\sqrt{\frac{ln(t)}{n_i(t)}} + e(t) + \frac{Z_1}{\sum_{k\neq i} n_k(t)} + \frac{Z_1}{\sum_{k\neq j} n_k(t)} \leq 2b_1\sqrt{\frac{\ln(t)}{n_i(t)}} + e(t) + O\left(\frac{1}{T^{w^*}}\right) < b_1\delta$$

Then

$$Pr\left[I_i(t) \geq \mathrm{I_j}(t) - e(t) | n_i(t) \geq \frac{(2b_1)^2\ln(t)}{\left(\frac{b_3\delta}{2}\right)^2}\right] \leq O(\frac{1}{t^4})$$

Like the proof of $\mathrm{E}[n_i(T)]$ in general UCB1 [7], we can induce that for $\gamma = \frac{(2b_1)^2\ln(t)}{\left(\frac{b_3\delta}{2}\right)^2}$, for some constant C, we have

$$\begin{aligned}
n_i(t) \leq & \gamma + \sum_{s=\gamma+1}^{t} \mathbf{1}(R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}} - e(t)) \\
\leq & \gamma + \sum_{s=\gamma+1}^{t} \mathbf{1}(\max_{\gamma<n<s} R_i(n) + b_1\sqrt{\frac{ln(t)}{n}} \geq \min_{0<n^*<s} R_j(n^*) + b_1\sqrt{\frac{ln(s)}{n^*}} - e(s), j \in \{1,2\}) \\
\leq & \gamma + \sum_{j\in\{1,2\}, s=1, n^*=1, n=\gamma}^{s=\infty, n^*=s-1, n=s-1} O(\frac{1}{s^4}) \\
\leq & \frac{(2b_1)^2 ln(t)}{(\frac{b_3\delta}{2})^2} + C \\
= & O(\frac{ln(T)}{\delta^2}) \\
= & O(T^w) \\
\leq & O(T^{w*})
\end{aligned}$$

We finish the proof of **Lemma 6**.

After proving that the number of selections will be limited if a bidder implements a positive deviated strategy, we are very close to the $O\left(\sqrt{\frac{\ln(t)}{t}}\right) - BNE$ now.

*Proof* of **Theorem 11**:

Firstly, $f_i(t) < 0, \forall t$ will be a bad strategy as bidder i's average gain will be negative, which is sure less than zero (the gain that this bidder can get if he implements the truth-telling strategy).

Secondly, $f_i(t) = O\left(\sqrt{\frac{\ln(t)}{t^w}}\right), 0 < w < 1, \forall t$ will be a bad strategy too. Because via **Lemma 6**, we can induce that the average gain of bidder i will be bounded by $\frac{O\left(\sqrt{\frac{\ln(T)}{T^w}}\right)T^{w^*}}{T} = O\left(\sqrt{\frac{\ln(T)}{T^{w-2w^*+2}}}\right) < O\left(\sqrt{\frac{\ln(T)}{T}}\right)$ when $\mathrm{w} < \mathrm{w}^* < \frac{1+w}{2}$, which is still within the $O\left(\sqrt{\frac{\ln(t)}{t}}\right) - BNE.$

Thirdly, $f_i(t) < 0$ for some t and $f_i(t) > 0$ for others (this scenario happens when a bidder uses $f_i(t) < 0$ first to increase his or her UCB-I score and thus increases his or her probability to be selected, and then asks for extra money when being selected in the following turns, will be a bad strategy too. We prove the infeasibility of this strategy in the following 2 cases:

If $\frac{\sum_{t=1}^{T} f_i(t)}{T} < 0$, then bidder i's average gain will be negative, which will surely be a bad strategy.

If $\delta = \frac{\sum_{t=1}^{T} f_i(t)}{T} > 0$, and $\delta = \sqrt{\frac{ln(t)}{t^w}}$, we need to prove that the similar edition of **Lemma 5** still hold in case, then we will have $E[n_i(T); n_i(t) \geq O(T^{w^*})] \leq O(T^w) < O(T^{w^*})$ (we only need to notice that

$$Pr\left[R_i^2(t) - R_j^2(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{\ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}\right] = Pr\left[R_i^2(t) - E\left[R_j^2(t)\right] + E\left[R_j^2(t)\right] - R_j^2(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{\ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}\right] \leq Pr\left[R_i^2(t) - E[R_i^2(t)] + E\left[R_j^2(t)\right] - R_j^2(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{\ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}\right]$$

in **Lemma 5**, and $E[R_j^2(t)] \geq E[R_i^2(t)] + b_3\delta$ still hold in **Lemma 6**).

**Lemma 7** If every bidder add no extra money ($\mathrm{f_j(t) = 0, \forall t, j \neq i}$) except bidder i and $\delta = \frac{\sum_{t=1}^{T} f_i(t)}{T} > 0$, then there exists $\sigma > 0$, s.t. $\forall \mathrm{i} \neq \mathrm{j}, \mathrm{n_i(t)} \leq (1+\sigma)n_j(t)$, with probability $1 - \mathrm{O}(\frac{1}{\mathrm{t}^2})$.

*Proof:*

Assume that at time t it changes to $\mathrm{n_i(t)} > (1+\sigma)n_j(t)$, then we want to bound $\Pr\left[I_\mathrm{i}(t) \geq I_j(t)\right]$.

$$
\begin{aligned}
Pr[I_i(t) \geq I_j(t)] \quad &= Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}}] \\
&\leq Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{(1+\sigma)ln(t)}{n_i(t)}}] \\
&\leq Pr[R_i(t) - R_j(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}}]
\end{aligned}
$$

Denote $\mathrm{Z}_1 = b_3Z(d+2)\frac{1+z}{(1-z)^2}$ in **Lemma 2**, then with probability $1 - e^{-C\sum_{k\neq i}n_k(t)}$, $|R_i^1(t)| \leq \frac{Z_1}{\sum_{k\neq i}n_k(t)}$ and with probability $1 - e^{-C\sum_{k\neq j}n_k(t)}$, $|R_i^1(t)| \leq \frac{Z_1}{\sum_{k\neq j}n_k(t)}$. Notice that we have $\sum_{k\neq i}n_k(t) \geq \sum_{n=1}^{t}|s(n)| - t \geq tT - t = t(T-1) \geq t^2$ and $\sum_{k\neq j}n_k(t) \geq \sum_{n=1}^{t}|s(n)| - t \geq tT - t = t(T-1) \geq t^2$. Then we conclude that with probability $1 - e^{-Ct^2}$, $max\{|R_i^1(t)|, |R_j^1(t)|\} \leq \frac{Z_1}{t^2}$.

Then

$$
Pr[R_i(t) - R_j(t) \geq (\sqrt{1+\sigma} - 1)\sqrt{\frac{ln(t)}{n_i(t)}}]
$$

$$
\begin{aligned}
\leq \quad & Pr[R_i^2(t) - R_j^2(t) \geq (\sqrt{1+\sigma} - 1)\sqrt{\frac{ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}] \\
= \quad & Pr[R_i^2(t) - E[R_j^2(t)] + E[R_j^2(t)] - R_j^2(t) \geq (\sqrt{1+\sigma} - 1)\sqrt{\frac{ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}] \\
\leq \quad & Pr[R_i^2(t) - E[R_i^2(t)] + E[R_j^2(t)] - R_j^2(t) \geq (\sqrt{1+\sigma} - 1)\sqrt{\frac{ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}] \\
\leq \quad & Pr[R_i^2(t) - E[R_i^2(t)] \geq \frac{\sqrt{1+\sigma} - 1}{2}\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2}] \\
& + Pr[R_j^2(t) - E\left[R_j^2(t)\right] \leq -\frac{\sqrt{1+\sigma} - 1}{2}\sqrt{\frac{ln(t)}{n_i(t)}} + \frac{Z_1}{t^2}]
\end{aligned}
$$

We notice that $-b_3Z \leq R_i^2(t) \leq \mathrm{b_3T\delta}$, denote $\mathrm{U} \coloneqq \mathrm{b}_3(Z + T\delta)/2$, then via Hoeffding inequality [6] we have

$$
\begin{aligned}
& Pr[R_i^2(t) - E[R_i^2(t)] \geq \frac{b_1(\sqrt{1+\sigma}-1)}{2}\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2}] \\
\leq \quad & exp(-\frac{2(\frac{b_1(\sqrt{1+\sigma}-1)}{2}\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2})^2 n_i(t)^2}{4\mathrm{U}^2 n_i(t)}) \\
\leq \quad & exp(-2\frac{\left(\left(\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\right)^2 \ln(t)\right)}{4\mathrm{U}^2}) exp(4\frac{\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\sqrt{\frac{\ln(t)}{n_i(t)}}\frac{Z_1 n_i(t)}{t^2}}{4\mathrm{U}^2}) \mathrm{exp}(-\frac{\frac{2Z_1^2}{t^4}n_i(t)}{4U^2}) \\
\leq \quad & \frac{1}{t^2}
\end{aligned}
$$

when $\sigma, b_1, t$ are large enough such that $\left(\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\right)^2 \geq 4\mathrm{U}^2$ and $4\frac{\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\mathrm{Z}_1}{4\mathrm{U}^2} \leq 1$.

Similarly $Pr[R_j^2(t) - E[R_j^2(t)] \geq -\frac{\sqrt{1+\sigma}-1}{2}b_1\sqrt{\frac{\ln(t)}{n_i(t)}} + \frac{Z_1}{t^2}] \leq O(\frac{1}{t^2})$.

Then we conclude that $Pr[I_i(t) \geq I_j(t)] \leq O(\frac{1}{t^2})$.

As at time t it changes to $\mathrm{n_i(t)} > (1+\sigma)n_j(t)$, then $\exists\ \mathrm{t_-} < t < t_+$ s.t. $\mathrm{n_i(t_-)} \leq (1+\sigma)n_j(t_-)$ and $\mathrm{n_i(t_+)} > (1+\sigma)n_j(t_+)$ , $\mathrm{n_i(t)} \geq (1+\sigma)n_j(t) - 1 \geq (1+\sigma-1)n_j(t)$, when σ is large enough so that $Pr[I_i(t) \geq I_j(t)] \leq O(\frac{1}{t^2})$.

At time $t_-$, i is selected but not j, otherwise the ratio $\frac{\mathrm{n_i(t)}}{(1+\sigma)n_j(t)}$ will go down, so we can induce that $I_i(t_-) \geq I_j(t_-)$. We discuss in the following 2 cases:

Case 1 $t_- \in [\frac{t}{2}, t]$

Then $Pr[I_i(t_-) \geq I_j(t_-)]$ can be upper bounded by $O\left(\frac{1}{\left(\frac{t}{2}\right)^2}\right) = O(\frac{1}{t^2})$.

Case 2 $t_- \in [0, \frac{t}{2})$

Then consider the bidder who was selected for most times during $[\frac{t}{2}, t]$, denote the bidder as bidder k.

If $\mathrm{n_k(t)} \leq (1+\sigma)n_j(t)$, then $n_j(t) \geq \frac{t}{2N(1+\sigma)}$, then

$$
\begin{aligned}
n_i(t)/n_j(t) \leq \quad & \frac{t/2+(1+\sigma)n_j(t)}{n_j(t)} \\
\leq \quad & \frac{\frac{t/2}{2}}{\frac{2}{2N(1+\sigma)}} + (1+\sigma) \\
= \quad & (1+\sigma)(N+1) \\
= \quad & 1 + (N + (N+1)\sigma)
\end{aligned}
$$

, in this case we can redefine $\sigma = N + (N+1)\sigma$.

If $\mathrm{n_k(t)} > (1+\sigma)n_j(t)$, then $\exists\ \mathrm{t^*}$ and $\mathrm{t^*} > \frac{t}{2}, n_k(\mathrm{t^*}) \geq (1+\sigma)n_j(\mathrm{t^*}) - 1 \geq (1+\sigma-1)n_j(\mathrm{t^*})$, when σ is large enough this probability can be upper bounded by $\mathrm{O}(\frac{1}{\mathrm{t}^2})$. We finish the proof of **Lemma 7**.

We can also prove a similar edition of **Lemma 6** (**Lemma 8)**.

**Lemma 8** Let $0 < \mathrm{w} < \mathrm{w}^* < 1$, $\delta := \sqrt{\frac{ln(t)}{t^w}}$, there is a bidder i who implements the strategy $\frac{\sum_{t=1}^{T} f_i(t)}{T} = \delta$. If at time t bidder i has been selected more than $\mathrm{O}(\mathrm{T}^{\mathrm{w}^*})$ times, bidder i will be selected no more than $\mathrm{O}(\mathrm{T}^{\mathrm{w}^*})$ in the whole process, i.e. $E[n_i(T); n_i(t) \geq O(T^{w^*})] \leq O(T^w) < O(T^{w^*})$.

*Proof:*

For bidder i, we need to bound its number of selections, which is equal to bound $\Pr[I_i(t) \geq \underset{j}{max} I_j(t) - e(t)]$, so we only need to consider the bidder i, j who has the highest and second highest UCB-I score $I_j(t)$ and $I_i(t)$ as others will have lower $\Pr[I_i(t) \geq \underset{j}{max} I_j(t) - e(t)]$. In other word, we need to bound $\Pr[I_i(t) \geq I_{\mathrm{j}}(t) - e(t)]$.

Recall $R_i^k(t) := -b_3 \frac{\sum_{n=1}^{t-1} 1(i \in s(n)) u_{i,t}^k(n)}{n_i(t)}$ and $u_{i,t}^2(n) = f_i(t)$, as bidder i who implements the strategy $\frac{\sum_{t=1}^{T} f_i(t)}{T} = \delta$, we can have that $E[R_j^2(t)] \geq E[R_i^2(t)] + b_3\delta$.

We notice that $Pr[I_i(t) \geq I_j(t) - e(t)] = Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}} - e(t)]$, and $\{R_i(t) + b_1\sqrt{\frac{\ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{\ln(t)}{n_j(t)}} - e(t)\}$ actually implies at least one of the following inequalities will be satisfied:

$$R_i^2(t) - E[R_i^2(t)] \geq b_1\sqrt{\frac{ln(t)}{n_i(t)}}$$

$$R_j^2(t) - E[R_j^2(t)] \leq -b_1\sqrt{\frac{ln(t)}{n_j(t)}}$$

$$b_3\delta \leq 2b_1\sqrt{\frac{ln(t)}{n_i(t)}} + e(t) + \frac{Z_1}{\sum_{k \neq i} n_k(t)} + \frac{Z_1}{\sum_{k \neq j} n_k(t)}$$

As otherwise there will be a contradiction:

$$
\begin{aligned}
& R_j(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}} - e(t) \\
> \quad & b_2 + R_j^2(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}} - e(t) - \frac{Z_1}{\sum_{k\neq j} n_k(t)} \\
> \quad & b_2 + \ E[R_j^2(t)] - e(t) - \frac{Z_1}{\sum_{k\neq j} n_k(t)} \\
\geq \quad & b_2 + E[R_i^2(t)] + b_3\delta - e(t) - \frac{Z_1}{\sum_{k\neq j} n_k(t)} \\
\geq \quad & b_2 + E[R_i^2(t)] + 2b_1\sqrt{\frac{ln(t)}{n_i(t)}} + \frac{Z_1}{\sum_{k\neq i} n_k(t)} \\
\geq \quad & b_2 + R_i^2(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} + \frac{Z_1}{\sum_{k\neq i} n_k(t)} \\
\geq \quad & R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}}
\end{aligned}
$$

We notice that $-b_3Z \leq R_i^2(t) \leq -\mathrm{b}_3\mathrm{T}\delta$, denote $\mathrm{U} \coloneqq \mathrm{b}_3(Z + T\delta)/2$, then via Hoeffding inequality [6] we have

$$
\begin{aligned}
& Pr[R_i^2(t) - E[R_i^2(t)] \geq \frac{b_1(\sqrt{1+\sigma}-1)}{2}\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2}] \\
\leq \quad & exp(-\frac{2(\frac{b_1(\sqrt{1+\sigma}-1)}{2}\sqrt{\frac{ln(t)}{n_i(t)}} - \frac{Z_1}{t^2})^2 n_i(t)^2}{4\mathrm{U}^2 n_i(t)}) \\
\leq \quad & exp(-4\frac{\left(\left(\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\right)^2 \ln(t)\right)}{8\mathrm{U}^2}) exp(4\frac{\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\sqrt{\frac{\ln(t)}{n_i(t)}}\frac{Z_1 n_i(t)}{t^2}}{4\mathrm{U}^2}) \exp(-\frac{\frac{2Z_1^2}{t^4}n_i(t)}{4U^2}) \\
\leq \quad & \frac{1}{t^4}
\end{aligned}
$$

when $\sigma, b_1, t$ are large enough such that $\left(\frac{b_1\left(\sqrt{1+\sigma}-1\right)}{2}\right)^2 \geq 8\mathrm{U}^2$ and $4\frac{\frac{b_1(\sqrt{1+\sigma}-1)}{2}\mathrm{Z}_1}{4\mathrm{U}^2} \leq 1$.

After proving $Pr\left[\left|R_i^2(t) - E[R_i^2(t)]\right| \geq b_1\sqrt{\frac{\ln(t)}{n_i(t)}}\right] = O\left(\frac{1}{t^4}\right)$, we will establish $\mathrm{n_i(t)} \leq (1+\sigma)n_j(t)$, then we can get $\sum_{k\neq i} n_k(t) \geq n_j(t) \geq O(T^{w^*})$.

If at any time the condition $\mathrm{n_i(t)} \leq (1+\sigma)n_j(t)$ is not satisfied, we can assume that there exists some time t, at time the condition changes to $\mathrm{n_i(t)} > (1+\sigma)n_j(t)$, then we need to bound $\Pr\left[\mathrm{I_i}(t) \geq I_j(t)\right]$. To finish the proof, actually we want the probability to be controlled in $\mathrm{O}(\frac{1}{\mathrm{t}^2})$. Firstly, we have that

$$Pr[I_i(t) \geq I_j(t)] \qquad = Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{ln(t)}{n_j(t)}}]$$

$$\leq Pr[R_i(t) + b_1\sqrt{\frac{ln(t)}{n_i(t)}} \geq R_j(t) + b_1\sqrt{\frac{(1+\sigma)ln(t)}{n_i(t)}}]$$

$$\leq Pr[R_i(t) - R_j(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}}]$$

Denote $\mathrm{Z}_1 = b_3Z(d+2)\frac{1+z}{(1-z)^2}$ in **Lemma 2**, then with probability $1 - e^{-C\sum_{k\neq i}n_k(t)}$, $|R_i^1(t)| \leq \frac{Z_1}{\sum_{k\neq i}n_k(t)}$ and with probability $1 - e^{-C\sum_{k\neq j}n_k(t)}$, $|R_j^1(t)| \leq \frac{Z_1}{\sum_{k\neq j}n_k(t)}$. Notice that we have $\sum_{k\neq i}n_k(t) \geq \sum_{n=1}^{t}|s(n)| - t \geq tT - t = t(T-1) \geq t^2$ and $\sum_{k\neq j}n_k(t) \geq \sum_{n=1}^{t}|s(n)| - t \geq tT - t = t(T-1) \geq t^2$. Then we conclude that with probability $1 - e^{-Ct^2}$, $max\{|R_i^1(t)|, |R_j^1(t)|\} \leq \frac{Z_1}{t^2}$.

Then as we have $E[R_i^2(t)] = E[R_j^2(t)]$,

$$Pr[R_i(t) - R_j(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}}]$$

$$\leq \qquad Pr[R_i^2(t) - R_j^2(t) \geq \left(\sqrt{1+\sigma} - 1\right)b_1\sqrt{\frac{ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}]$$

$$\leq \qquad Pr[R_i^2(t) - E[R_i^2(t)] \geq \frac{\sqrt{1+\sigma} - 1}{2}b_1\sqrt{\frac{ln(t)}{n_i(t)}} - 2\frac{Z_1}{t^2}]$$

$$+Pr[R_j^2(t) - E\left[R_j^2(t)\right] \leq -\frac{\sqrt{1+\sigma} - 1}{2}b_1\sqrt{\frac{ln(t)}{n_i(t)}} + 2\frac{Z_1}{t^2}]$$

We notice that $-b_3Z \leq R_i^2(t) \leq \mathrm{b_3T\delta}$, denote $\mathrm{U} \coloneqq \mathrm{b}_3(Z + T\delta)/2$, then via Hoeffding inequality [6] we have