Auction-Based Combinatorial Multi-Armed Bandit Mechanisms with Strategic Arms

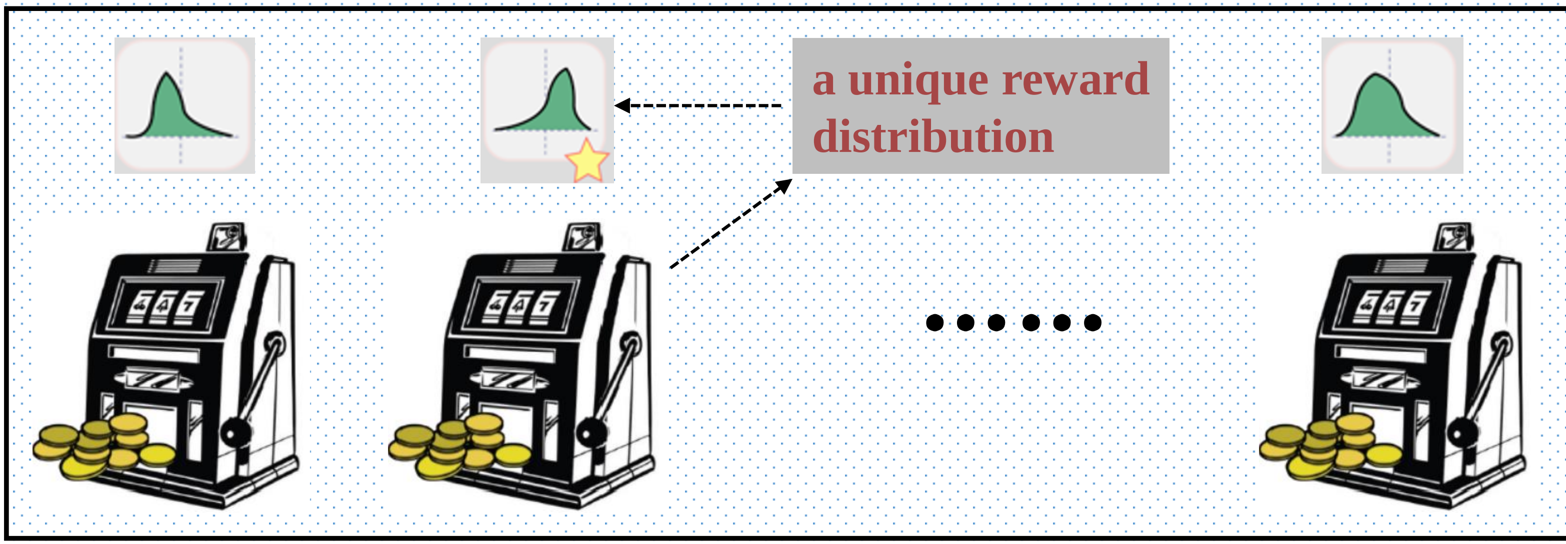
Guoju Gao, He Huang*, Mingjun Xiao*,
Jie Wu, Yu-E Sun, Sheng Zhang

Road Map

- Background & Motivation
- Model & Design Goal
- Solution
- Simulation
- Conclusion

SOOCHOW UNIVERSITY

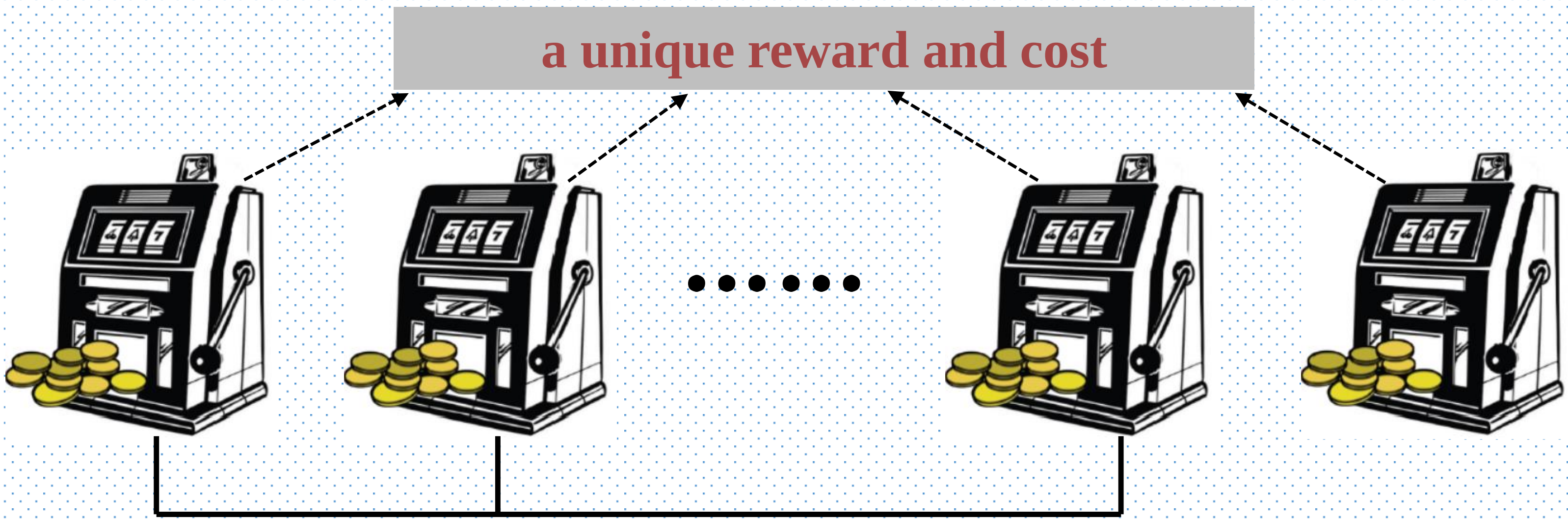
Multi-Arm Bandit (MAB) Model



How to select **one arm** in each round such that the **cumulative rewards** can be maximized under the round constraint?

SOOCHOW UNIVERSITY

Combinatorial MAB (CMAB)



How to select K arms in each round such that the cumulative rewards can be maximized under the budget constraint?

exploration **vs.** **exploitation**

SOOCHOW UNIVERSITY

Limitations: Strategic Arms

traditional

All arms are feelingless machines

novel

Each arm may strategically report its cost to maximize its own payoff.

What about the scenario where all arms are **rational and selfish** individuals?

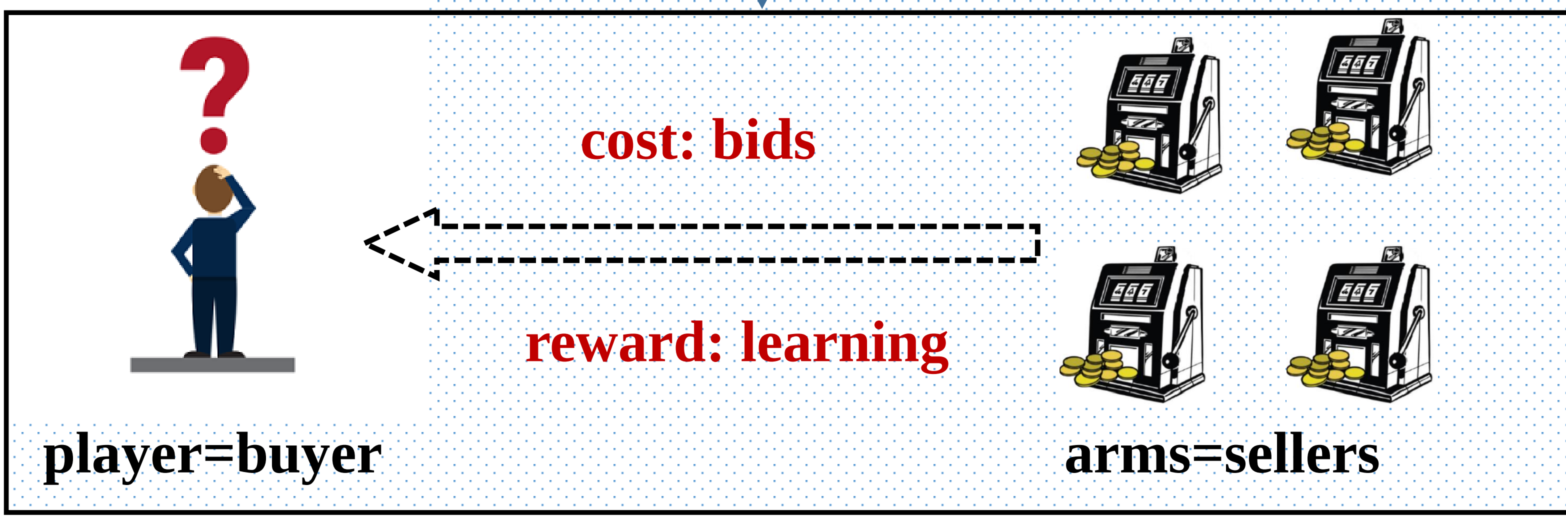
Our focus: CMAB model with strategic arms

SOOCHOW UNIVERSITY

Auction-Based CMAB (ACMAB)

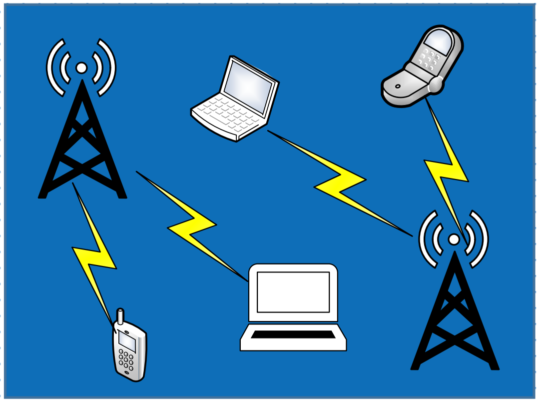
How to solve the strategic behaviors in CMAB?

combination of auction theory and CMAB

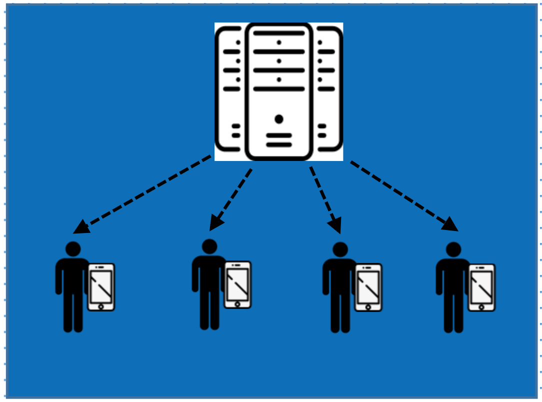


SOOCHOW UNIVERSITY

Application of the ACMAB Model



rate allocation in wireless network



user selection in crowdsensing

Ad recommendation in social network

Goals for ACMAB

- **Truthfulness**
  - each arm will truthfully bid its cost value
- **Individual rationality**
  - each arm's payoff must be greater than 0
- **Computational efficiency**
  - polynomial-time complexity
- **Good regret performance**
  - the difference in the total achieved rewards between the optimal policy and our proposed bandit-pulling policy

SOOCHOW UNIVERSITY

Existing Methods for ACMAB

First Exploring:

- uses a fraction of budget to learn arms' rewards
- determines the payment with the maximum value

⇩

Second Exploiting:

- uses remaining budget to select the top K "best" arms
- determines the critical payment (auction theory)
- the average sampling rewards will not update

SOOCHOW UNIVERSITY

Our Proposed Solution

Combining exploration and exploitation:

- tailor-made upper confidence bound expression to balance the exploitation and exploration
- greedy and monotonic winner selection
- critical payment computation for winners
- update the sampling rewards in all rounds

good regret bound + truthfulness

Upper Confidence Bound (UCB)

exploitation **vs.** **exploration**

$$\beta_i(t) = \begin{cases} \beta_i(t-1)+1; & i \in \Phi^t \\ \beta_i(t-1); & i \notin \Phi^t \end{cases}$$

$$\overline{r}_i(t) = \begin{cases} \dfrac{\overline{r}_i(t-1)\beta_i(t-1) + r_i^t}{\beta_i(t-1)+1}; & i \in \Phi^t \\ \overline{r}_i(t-1); & i \notin \Phi^t \end{cases}$$

average reward

related to the regret bound

$$u_i(t) = \sqrt{\frac{(K+1)\ln t}{\beta_i(t)}}$$

UCB bonus

UCB-based reward: optimism in the face of uncertainty

$$\widehat{r}_i(t) = \overline{r}_i(t) + u_i(t)$$

SOOCHOW UNIVERSITY

Winning Arm Selection Procedure

- **Initialization phase**
  - selects all arm in the first round to initialize some parameters
  - determines the payment with the maximum value c_{max}
  - updates the remaining budget

- **Winning arm selection phase**
  - acquires all arms' UCB-based rewards in the previous round
  - computes the ratios of UCB-based rewards and bids
  - selects top K arms according to the sorted ratio values

Payment Determination Procedure

Myerson rule for auction mechanisms

- ✓ the winner selection process is monotonic
- ✓ each winner is paid with the critical value

$$p_i^t(b_i) = \min\{\frac{\widehat{r}_i(t-1)}{\widehat{r}_{K+1}(t-1)} \cdot b_{K+1}, c_{max}\}$$

- ➢ $\frac{\widehat{r}_i(t-1)}{\widehat{r}_{K+1}(t-1)} b_{K+1}$ means the critical payment
- ➢ $\min\{\cdot\}$ ensures the maximum payment
- ➢ updates the remaining budget

* For a winning arm, a bid larger than the critical payment will not win, but a smaller bid must win

SOOCHOW UNIVERSITY

Detailed Algorithm: AUCB

initialization

winner selection

payment computation

termination & output

update

Algorithm 1 Auction-based UCB Algorithm (AUCB)

Require: $\mathcal{N}$, $\mathcal{B}$, K, and B
Ensure: Φ, $r(B)$, $\tau(B)$, and $\mathcal{P}$

Initialization: $t=1$, $B(0)=B$, and $r(B)=0$, the player selects all arms in the first round, i.e., $\Phi^1=\mathcal{N}$;
Obtain the reward values r_i^1 for $\forall i\in\mathcal{N}$ in the first round;
Determine the payments for selected arms, i.e., $p_i^1=c_{max}$;
Update the parameters: $\overline{r}_i(t)$, $\widehat{r}_i(t)$, $B(t)=B(t-1)-N\cdot c_{max}$, and $r(B)=r(B)+\sum_{i\in\Phi^t} r_i^t$;
while $true$ **do**
    $t\Leftarrow t+1$, $\Phi^t=\phi$, and $p_i^t(b_i)=0$ for $\forall i\in\mathcal{N}$;
    Sort the arms according to the value $\frac{\widehat{r}_i(t-1)}{b_i}$;
    Consider $\frac{\widehat{r}_{i_1}(t-1)}{b_{i_1}}\geq\cdots\geq\frac{\widehat{r}_{i_j}(t-1)}{b_{i_j}}\cdots\geq\frac{\widehat{r}_{i_N}(t-1)}{b_{i_N}}$;
    Select the top K arms, denoted as Φ^t;
    Compute the payments for each selected arm in Φ^t, i.e., $p_{i_j}^t(b_{i_j})=\min\{\frac{\widehat{r}_{i_j}(t-1)}{\widehat{r}_{i_{K+1}}(t-1)}\cdot b_{i_{K+1}}, c_{max}\}$;
    if $\sum_{i\in\Phi^t} p_i^t(b_i)\geq B(t-1)$ **then**
        return Terminate and output Φ, $r(B)$, $\tau(B)=t$, $\mathcal{P}$;
    Obtain the rewards r_i^t for $\forall i\in\Phi^t$;
    Update the parameters: $\overline{r}_i(t)$, $\widehat{r}_i(t)$, $B(t)=B(t-1)-\sum_{i\in\Phi^t} p_i^t(b_i)$, and $r(B)=r(B)+\sum_{i\in\Phi^t} r_i^t$;

SOOCHOW UNIVERSITY

SOOCHOW UNIVERSITY

Properties of the AUCB Algorithm

- Upper bound on regret (Theorem 1)
  - The expected regret of AUCB is bounded as
    $$O\Big(NK^3\ln(B+NK^2\ln(NK^2))\Big)$$
- Truthfulness in each round (Theorem 2)
- Individual rationality (Theorem 3)
- Computational efficiency (Theorem 4)
  - The computational overhead of AUCB is
    $$O(NB+N^2K^2\ln(NK^2))$$

Simulation Settings

- **Compared algorithms**
  - optimal: arms' expected rewards are known in prior; the extremely-critical payment equals to the bid.
  - separate[1]: tailor-made exploration budget and exploitation budget; payment in each round is fixed.
  - ε-first[2]: ε*budget for randomness, (1-ε)*budget for the exploitation; payment is based on the average rewards.
- **Settings**

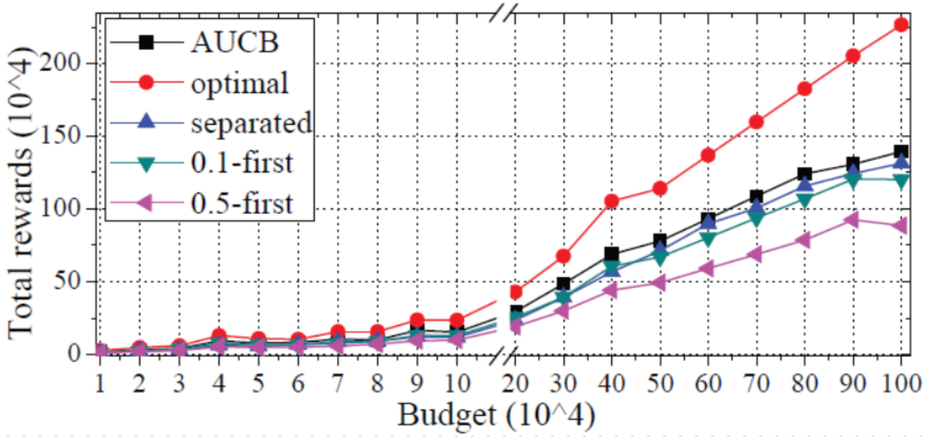
parameter name	range
budget, B	10^4 – 10^6 (5×10^5 in default)
number of arms, N	50 – 100 (60 in default)
number of selected arms, K	10 – 50 (20 in default)
expected reward, r_i	0.1 – 1
variance of reward, σ_i	0 – $\min\{r_i/3, (1-r_i)/3\}$
cost, c_i and bid, b_i	0.1 – 1

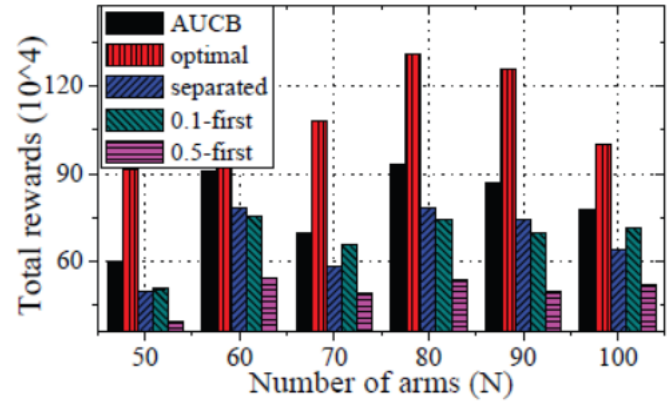
[1] A. Biswas, S. Jain, D. Mandal, and Y. Narahari, "A truthful budget feasible multi-armed bandit mechanism for crowdsourcing time critical tasks," in International Conference on Autonomous Agents and Multiagent Systems, 2015, pp. 1101–1109.

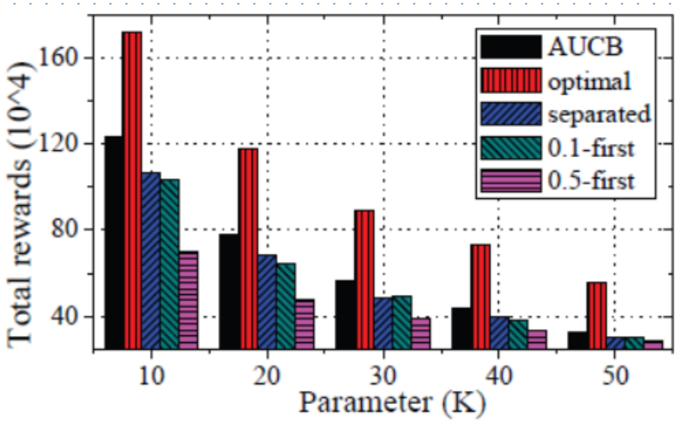
[2] L. Tran-Thanh, A. Chapman, E. M. de Cote, A. Rogers, and N. R. Jennings, "Epsilon–first policies for budget–limited multi-armed bandits," in Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.

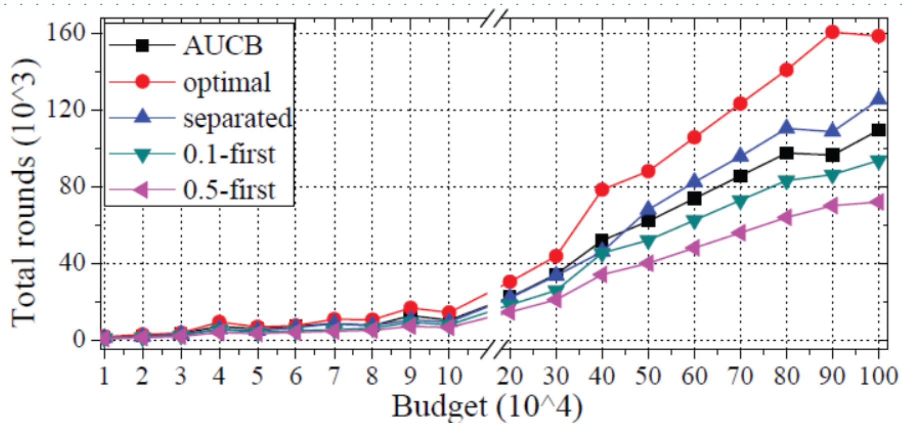
SOOCHOW UNIVERSITY

Results: Total Rewards



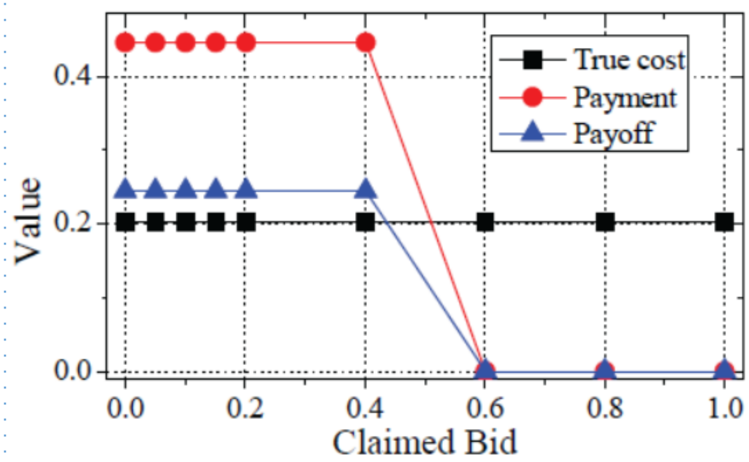


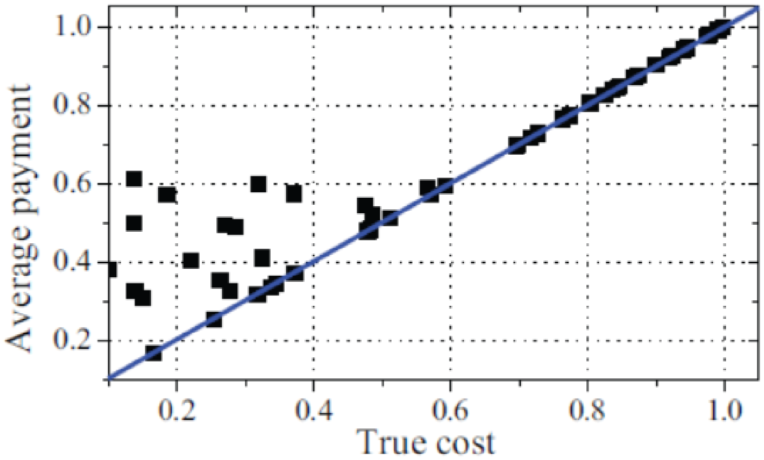


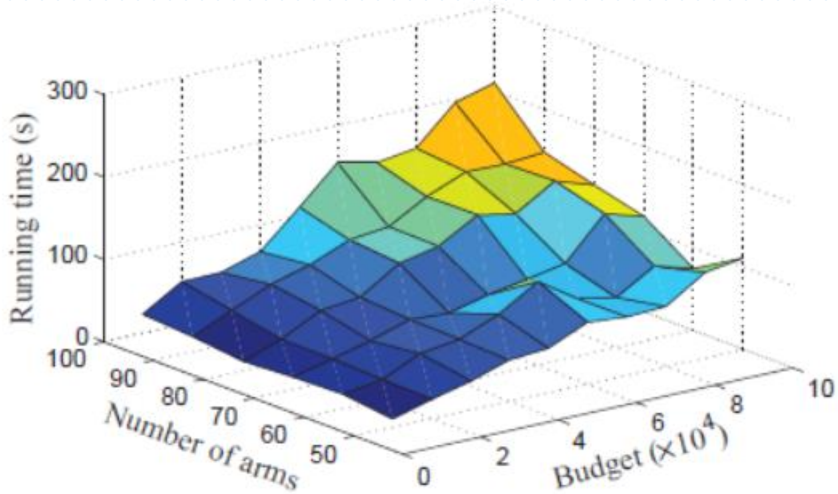


SOOCHOW UNIVERSITY

Results: Auction Properties







Conclusion

- Simulation results show that the total rewards achieved by AUCB are at least 12.49% higher than those of state-of-the-art (e.g., “exploration-separate”) algorithms.

- AUCB can ensure the truthfulness and individual rationality of the strategic arms.

- The computational overload of AUCB is polynomial.

- Both the theoretical analysis and simulation results show that AUCB has a good regret bound.

SOOCHOW UNIVERSITY

@ Contact Me

SOOCHOW UNIVERSITY

Thank You!

Q & A

gjgao@suda.edu.cn