Exploiting Correlated Auxiliary Feedback in Parameterized Bandits

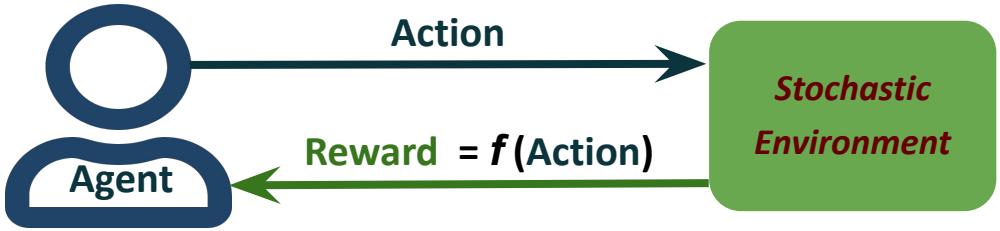
Arun Verma, Zhongxiang Dai, Yao Shu, Bryan Kian Hsiang Low

National University of Singapore

NeurIPS 2023

Parameterized Bandits

- In each round, an agent (or decision-maker) selects the next action.



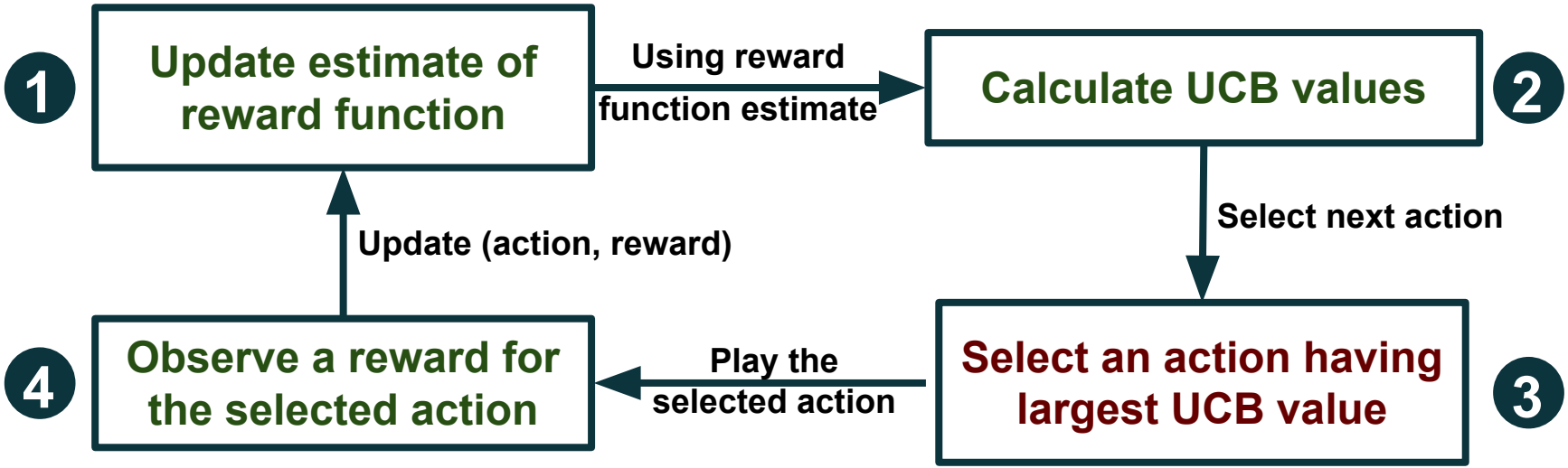
- Environment generates a stochastic reward, which is an unknown function (***f***) of the features of the selected action.
- Here, function ***f*** can be a non-linear, complex, and black-box function.

How to select the next action that maximizes the reward?

UCB-based algorithm for Parameterized Bandits

To select the next action, the UCB-based bandit algorithm

- uses a ***suitable function estimator*** to model the unknown reward function (e.g., Gaussian process to model a non-linear function) and
- selects an action that maximizes the ***Upper Confidence Bound (UCB)*** (using the estimated function) to balance ***exploration*** and ***exploitation***.



Auxiliary Feedback in Parameterized Bandits

Online food delivery platform:




- ***Different*** restaurants can be recommended to a user.
  Goal: Recommend a restaurant that has the highest user rating.
- Here, the *food delivery time* can be **auxiliary feedback** as it influences the user's rating.

Other similar problems:

- Showing best online sellers to users by e-commerce platform.
- Selecting the best cab for the rider by online cab aggregator.

How to use the Correlated Auxiliary Feedback to learn the best action quickly?

Image source: Internet

Control Variates

- Let μ be the unknown quantity that needs to be estimated.
- y be an unbiased estimator of μ , i.e., $\mathbb{E}[y] = \mu$.
- Any random variable w with known mean ω is a control variate if it is correlated with y.

For any β, define a new unbiased estimator as:

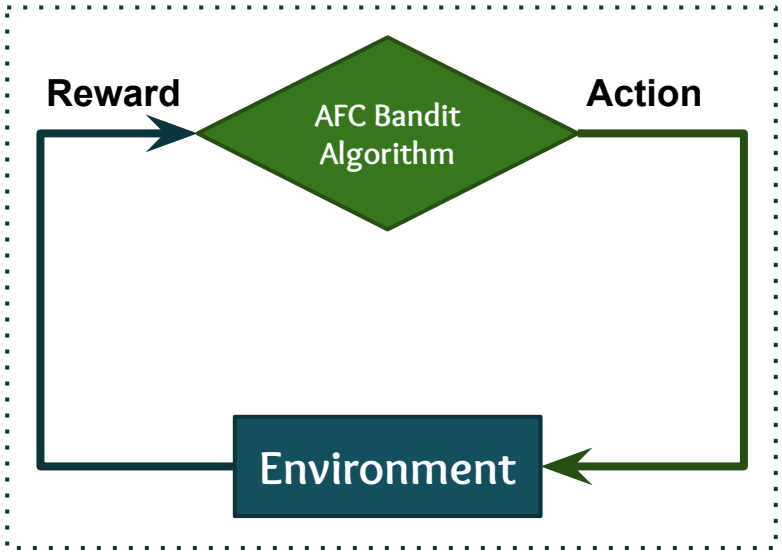
$$z = y + \beta(\omega - w).$$

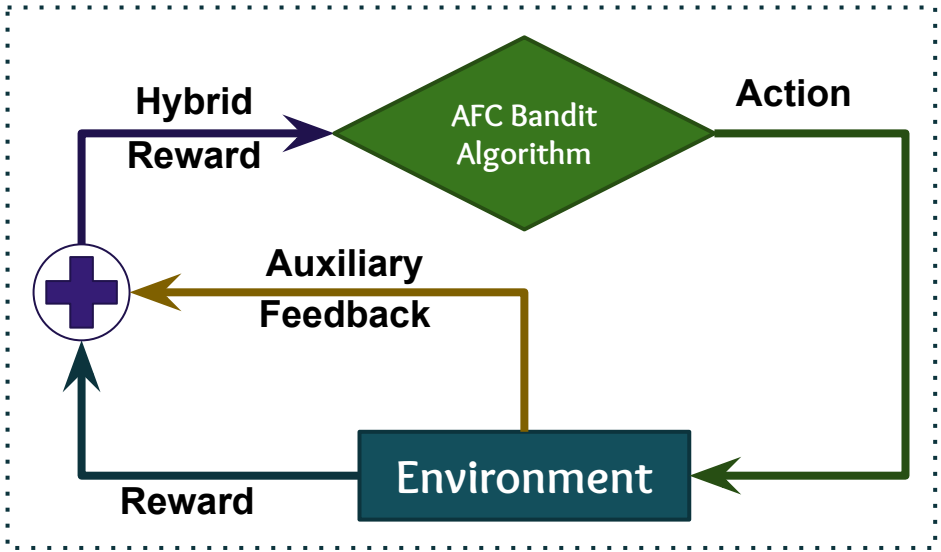
- For $\beta = \dfrac{\mathrm{Cov}(y, w)}{\mathrm{Var}(w)}$, $\mathrm{Var}(z) = (1 - \rho^2)\mathrm{Var}(y)$ is minimum, where ρ is the correlation coefficient between y and w.

Using Auxiliary Feedback in Bandit Algorithm

- **Hybrid rewards:** combination of reward and its auxiliary feedback, which leads to an unbiased reward estimator with a smaller variance than using only rewards.

- **Auxiliary Feedback Compatible (AFC) bandit algorithm:**
  Any bandit algorithm that can use hybrid rewards instead of only observed rewards.

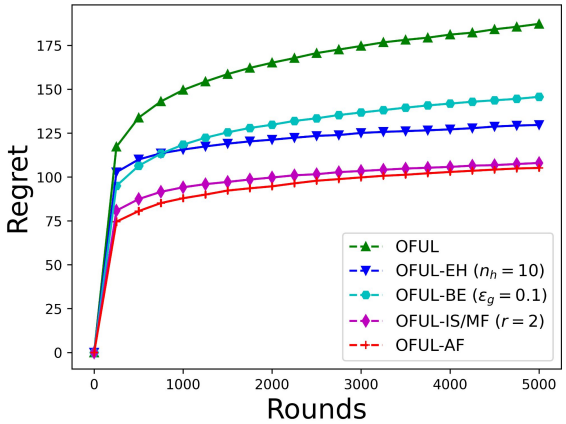




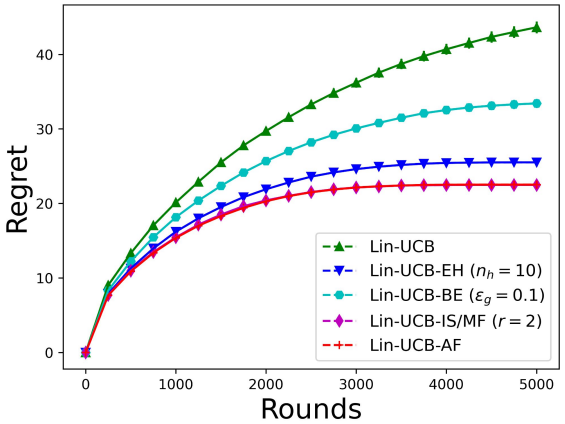
Results

Visit our poster or check our paper for more details.

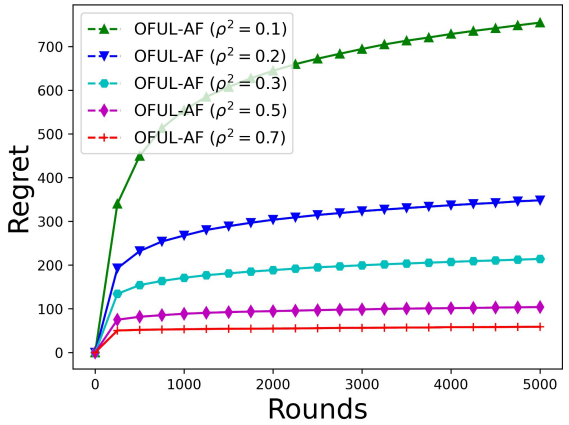
- Let ρ be the correlation coefficient between reward and its auxiliary feedback.
- Instantaneous regret of any AFC bandit algorithm using hybrid rewards is smaller by a factor $O((1-\rho^2)^{1/2})$ compared to when it only uses observed rewards.



Linear Bandit (OFUL)



Linear Contextual Bandits (Lin-UCB)



Influence of correlation (OFUL)