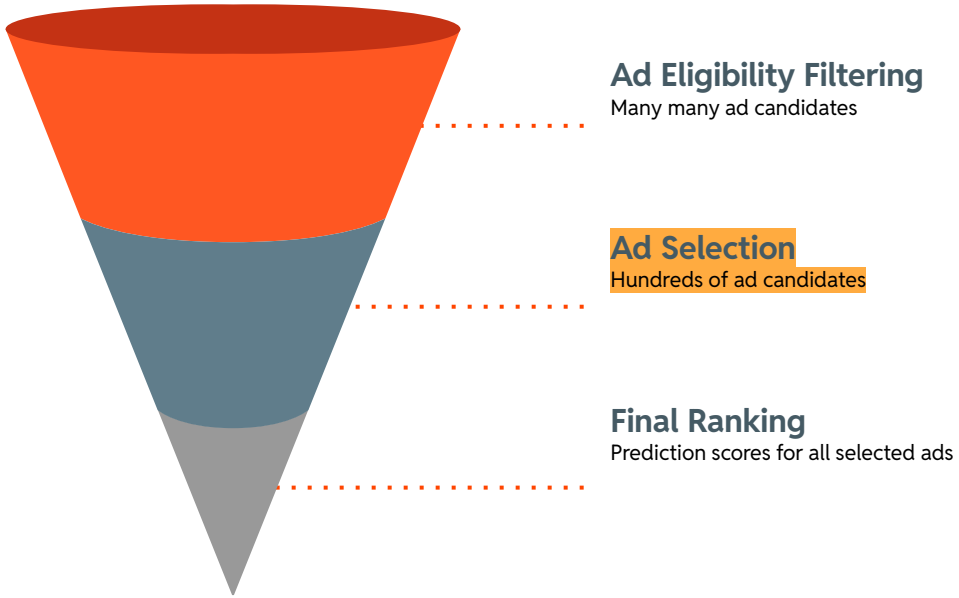
Augmented Two-Stage Bandit Framework: Practical Approaches for Improved Online Ad Selection

Seowon Han, Ryan Lakritz, Hanxiao Wu

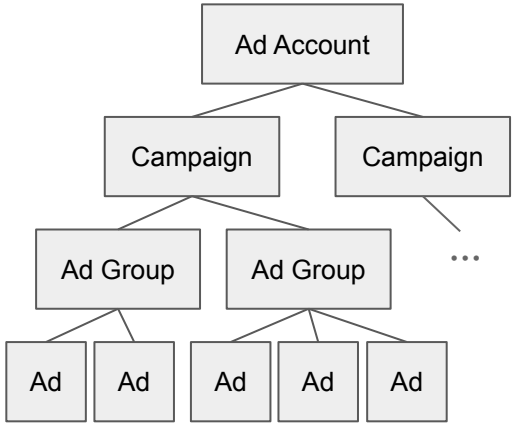
Ad Auction Funnel

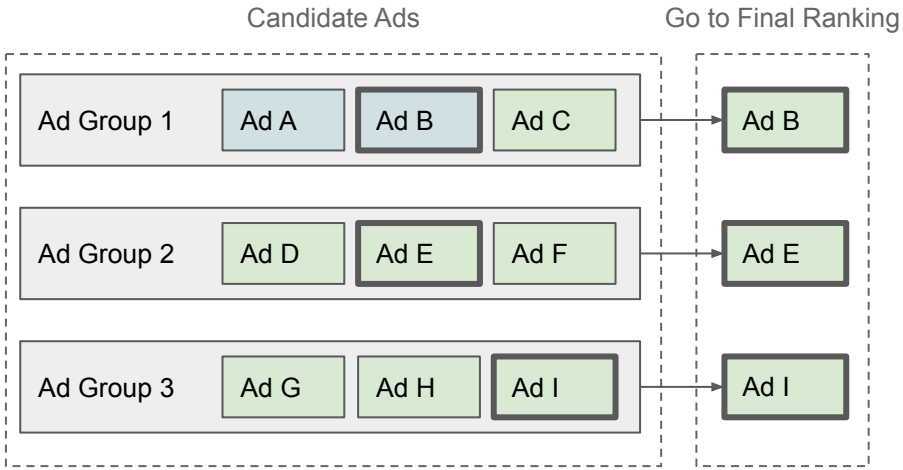
- Multi-faceted final auction utility score is based on:
  - Bid
  - Predicted CTR (pCTR)
  - Bid Modification and Utility Boosting factors
- Each level of the funnel filters based on advertiser preferences and indicators of success in the final auction
- Final Ranking (pCTR model) is a very computationally expensive model
- **Ad Selection** plays a crucial role in filtering, optimization, and exploration



Ad Selection

- ➔ Our ad management framework contains a hierarchy, including "**Ad Group**" which contains a set of *similar* Ads.
- ➔ In the Ad Selection stage, the system selects one **Ad per Ad Group**





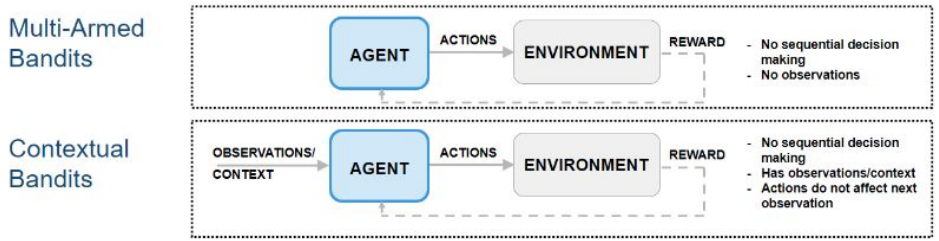
Motivation

Constraints and requirements

- Real-time adaptability for new ads
- Low latency (< 10s of ms), low infrastructure overhead
- Exploration is important; exploring within a range of some confidence bound is equally important

Bandit algorithms is well suited!

- Agent observes an impression with associated feature vectors
- Agent chooses an ad a from the set of eligible ads based on the learned policy
- Agent observes the clicks generated by the impression
- Agent updates the policy



Limitations of existing methods

Non-contextual Multi-armed bandit:

- No contextual information is considered when choosing the action
  - The reward function: $r_{MAB}(a_t)$
- Though it may achieve fast convergence, especially for new ads, personalization is limited where rewards are not optimal at each feature level - “one ad fits all (features)”

Contextual bandit

- Considers contextual information when choosing the action
  - The reward function: $r_{CB}(s_t, a_t)$
- More personalization, eg, time of the day, device; can achieve higher total rewards
- Suffer from data sparsity and excessive exploration at the initial stage of learning

In practice, contextual bandit tends to perform worse than multi-armed bandit at the beginning but catch up over time

Our Proposal

We want to achieve higher total rewards with personalization in the long run while preserving performance at early stage

Proposal: Augmented **Two-staged bandit framework**

Motivation: The best performing ad for the overall marketplace is likely a better-than-average candidate in each context.
Proposal: Initially relying on the context-free policy's rewards when the context information is sparse, and then transitioning to the context-aware policy's rewards once it outperforms the context-free bandit policy.

$$r_{TS}(s_t, a_t) = \begin{cases} r_{MAB}(a_t) & \text{if } Var_t(s,a) > \tau \\ r_{CB}(s_t, a_t) & \text{otherwise} \end{cases}$$

- $r_{MAB}(a_t)$ context-free reward function
- $r_{CB}(s_t, a_t)$ contextual reward function
- $Var_t(s,a)$ variance of expected contextual rewards
- τ threshold tuned using offline evaluation and online experimentation

Our Proposal

We want to achieve higher total rewards with personalization in the long run while preserving performance at early stage

Proposal: **Augmented** Two-staged bandit framework

Motivation: Knowledge distillation of heavy ranking pCTR model

- Real-time pCTR has high accuracy but not feasibility in early ranking
- Previous day's pCTR is accessible immediately and has high correlation with actual CTR
- Incorporating this knowledge will further mitigate data-sparsity issues for new agents

Proposal: Augmentate the previous day's pCTR scores as weights to the context-free reward

$$r_{A-MAB}(a_t) = r_{MAB}(a_t) * pCTR(a_t)$$

- $r_{MAB}(a_t)$ context-free reward function
- $pCTR(a_t)$ previous day's pCTR

Our Proposal

We want to achieve higher total rewards with personalization in the long run while preserving performance at early stage

Augmented Two-staged bandit framework

$$r_{A-TS}(s_t, a_t) = \begin{cases} r_{A-MAB}(a_t) & \text{if } Var_t(s, a) > \tau \\ r_{CB}(s_t, a_t) & \text{otherwise} \end{cases}$$

Experiment Setup

Experiment Setup

Experiment Hypothesis: The proposed Augmented Two-Stage Bandit framework produces measurable performance improvement, especially in cold-start and data-scarce scenarios.

Duration: 7 days

Evaluation Metric: Click Through Rate (CTR)

Experiment Variants

Control: Base Linear Thompson Sampling Model

Two-Stage Bandit: Thompson Sampling in first stage; Linear Thompson Sampling in second stage

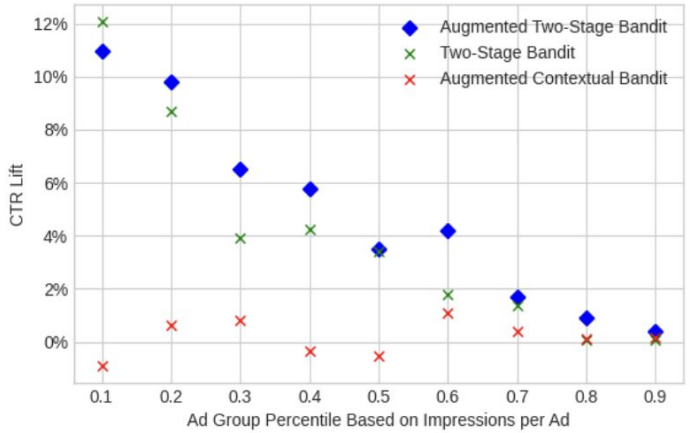
Augmented Contextual Bandit: Linear Thompson Sampling with pCTR augmented rewards

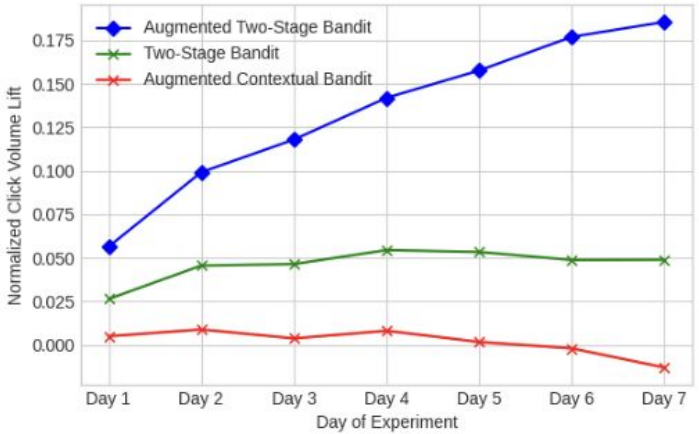
Augmented Two-Stage Bandit: Our proposed framework; combination of the two mechanisms above

Experiment Results

- ➔ Aggregate experiment results showed modest significant lift in CTR
- ➔ Cold-Start and Data-Scarce scenarios, represented by impression percentiles, showed substantial improvements in CTR
- ➔ Overall Click Volume improved for advertisers

	CTR Lift
Control	-
Augmented Two-Stage Bandit	0.97%
Two-Stage Bandit	0.49%
Augmented Contextual Bandit	-0.12%





Conclusion

- ➔ The two-stage augmented bandit framework provides a set of improvements on top of Contextual Bandit problem formulations
- ➔ This framework particularly addresses the cold-start that is present in contextual bandits and general online Ad Selection models
- ➔ Our implementation offers practical application to online serving with low latency requirements
- ➔ The online experiment results showed significant improvements in key performance metrics, with particular improvement in cold-start and data-scare scenarios