Ad Asset Portfolio Optimization via Policy Learning (AdKDD2026)

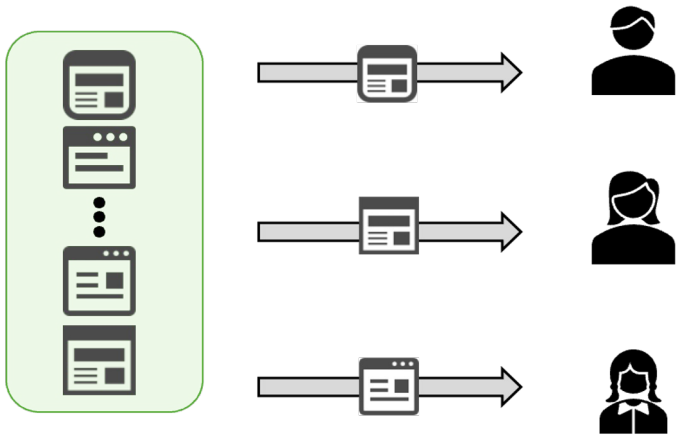
Koichi Tanaka[1], Zhi Wang[2], Masahiro Asami[2],
Kota Ishizuka[2], Kosuke Kawakami[2], Yuta Saito[3]

[1]Keio University, [2]HAKUHODO Technologies Inc., [3]Hanjuku-kaso, Co., Ltd.

Ad Asset Selection for Diverse Users

Many ad selection algorithms have been developed and applied in industrial systems.

Platforms (e.g., YouTube, TikTok, X)



Display ad assets for users

- Given a set of ad assets,
  a selection algo displays optimal ads
- Ultimate goal is the reward maximization
- However, even optimal algos might fail depending on a set of ad assets
  - ○ few or low-diversity ad assets
    → no room to optimize
  - ○ too many ad assets
    → substantial exploration costs

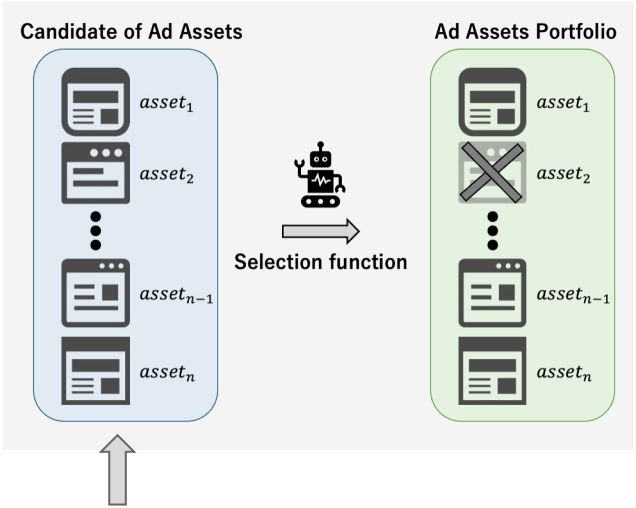
Question

Which ads should advertisers submit to platforms?

Ad Asset Portfolio Optimization (Ad-POP)

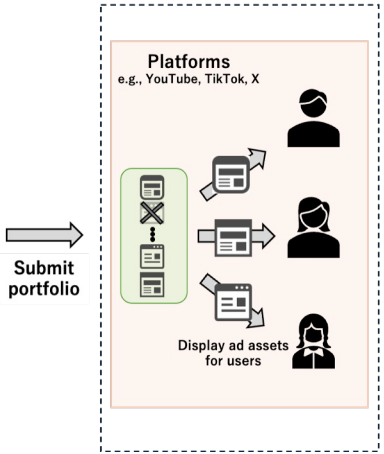
To tackle this challenge, we introduce and formulate
an underexplored problem, **"Ad Asset Portfolio Optimization (Ad-POP)"**

Ad-POP (our focus)



may be generated by LLMs

Conventional Ad Selection

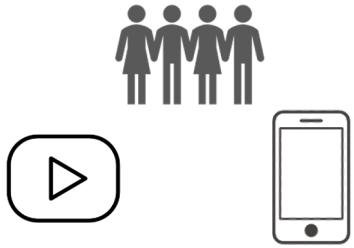


- Ad-POP considers which ads we should submit to platforms
- Ad-POP is increasingly important
  - LLMs can generate a vast number of ads automatically

Goal of Ad-POP

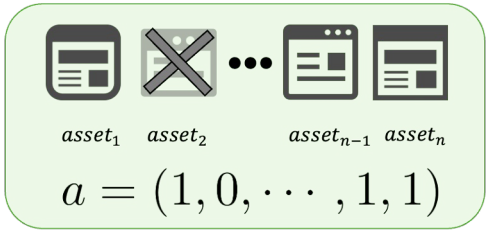
Identifying optimal ad asset portfolios for downstream reward maximization

Problem Formulation (contextual combinatorial bandit)



$$x \sim p(x)$$

observe context
(platforms, campaigns, target, etc.)

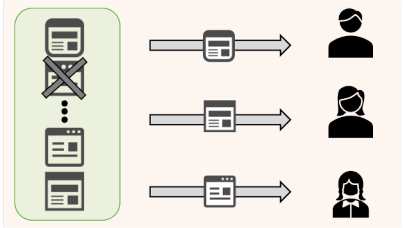


$$a \sim \pi(a|x)$$

select an ad asset portfolio

submit portfolio

platform



$$r \sim p(r|x, a)$$

observe reward
(click, conversion)

Goal of Ad-POP : learning a new policy to maximize rewards

$$\max_{\theta \in \Theta} V(\pi_\theta) := \mathbb{E}_{p(x)\pi_\theta(a|x)}[q(x, a)]$$

only using logged data $\mathcal{D} = \{(x_i, a_i, r_i)\}_{i=1}^n \sim \pi_0 \neq \pi_\theta$

Existing Approaches

	Policy-based Approach	Reg-based Approach
Exact Approach	IPS-PG High Variance	Reg-based High Bias
Independent Approach	PI-PG High Bias	Independent Reg-based High Bias

Exact Approach

- Reg-based method
  - estimate expected reward using conventional machine learning methods
  - difficult to learn rewards for all possible ad asset combinations
- Policy-based method (IPS-PG)
  - use iterative gradient ascent
  - high variance due to importance weights for ad asset combinations

Independent Approach

- Reg-based method (Independent Reg-based)
  - learns the value of each ad asset independently and constructs ad asset portfolios by adding high-value ad asset
  - can avoid combinatorial explosion, but cannot deal with interaction effects
- Policy-based (PI-PG)
  - also suffer from high bias due to interaction effects

Ad Asset Portfolio Optimization via Meta Information (Ad-POM)

Policy Decomposition via Meta Information

The Policy Decomposition via Meta Information

We decompose a policy into 1st-stage and 2nd-stage policies, using meta information

$$\pi_{\theta,\phi}^{overall}(a|x) = \sum_{m \in \mathcal{M}} \pi_{\theta}^{1st}(m|x) \cdot \pi_{\phi}^{2nd}(a|x,m)$$

- $\mathcal{m}$ represents meta information
  - the number of ad assets, the degree of diversity, etc.
- $\pi_{\theta}^{1st}(m|x)$ selects meta information
  - the 1st-stage policy is optimized by the policy-based method
- $\pi_{\phi}^{2nd}(a|x,m)$ selects an ad asset portfolio based on $\mathcal{m}$ sampled by 1st-stage policy
  - the 2nd-stage policy is optimized by the Reg-based method

Optimizing the 1st-Stage Policy

1st-stage policy is optimized via the policy-based method

The Ad-POM Gradient Estimator

$$\nabla_\theta \widehat{V}_{\mathrm{Ad-POM}}(\pi^{overall}_{\theta,\phi}; \mathcal{D})$$

$$:= \frac{1}{n}\sum_{i=1}^{n}\left\{ \frac{\pi^{1st}_\theta(m_{a_i}|x_i)}{\pi^{1st}_0(m_{a_i}|x_i)} \left(r_i - \hat{f}(x_i, a_i)\right)(x_i, m_{a_i}) + \mathbb{E}_{\pi^{1st}_\theta(m|x_i)}\left[\hat{f}^{\pi^{2nd}_\phi}(x_i, m)(x_i, m)\right] \right\}$$

where $f^{\pi^{2nd}_\phi}(x, m) = \mathbb{E}_{\pi^{2nd}_\phi(a|x,m)}[f(x, a)]$

meta information importance weight

- Ad-POM gradient estimator is unbiased under reasonable conditions
- applies importance weighting with respect to only the meta information space
  - reduce variance compared to the exact policy-based approach
- accurate $\hat{f}(x, a)$ provides more variance reduction

Optimizing the 2nd-Stage Policy

Process

1. We estimate the structure of the expected reward function

   If m includes the number of ads d and degree of similarity λ

$$\tilde{q}(x,a) = \sum_{l=1}^{|\mathcal{B}|} \eta(x, b_l)\, \mathbb{I}\{b_l = 1\} + \lambda \cdot \mathrm{Div}(a)$$

independent reward

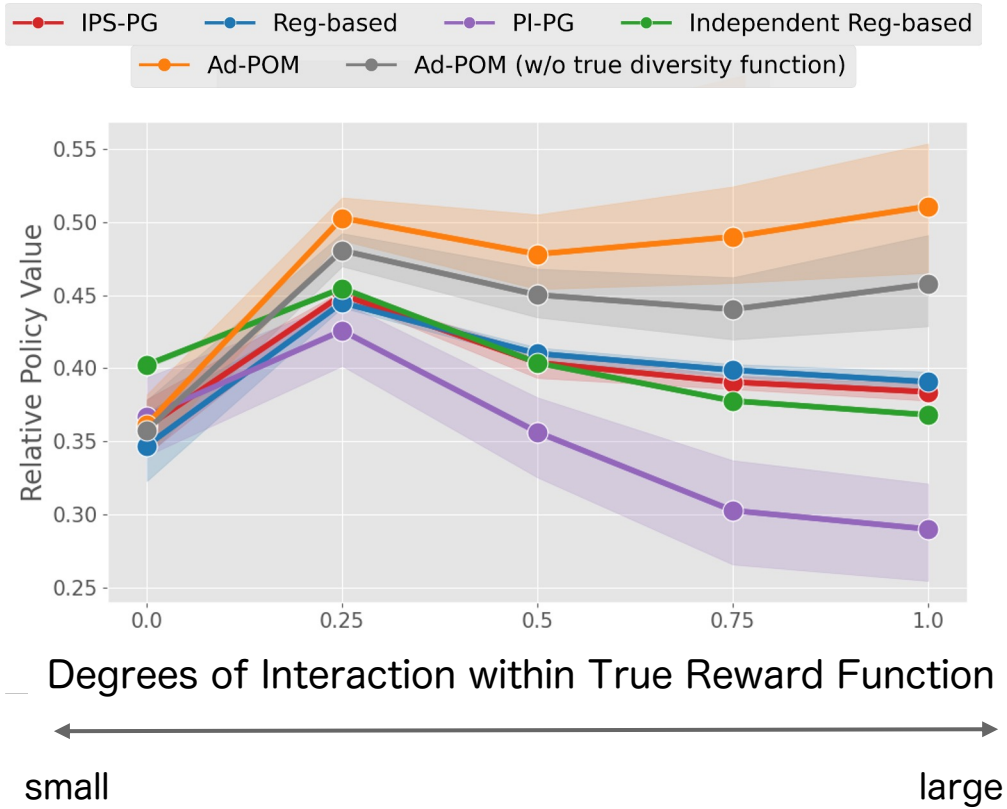
diversity function
(e.g., cosine similarity)

2. 2nd-stage policy selects d ad assets with highest values

- Considers the asset diversity as well as the independent value of assets
  - ○ low bias compared to independent approach
- Need not estimate the exact structure of the expected reward function

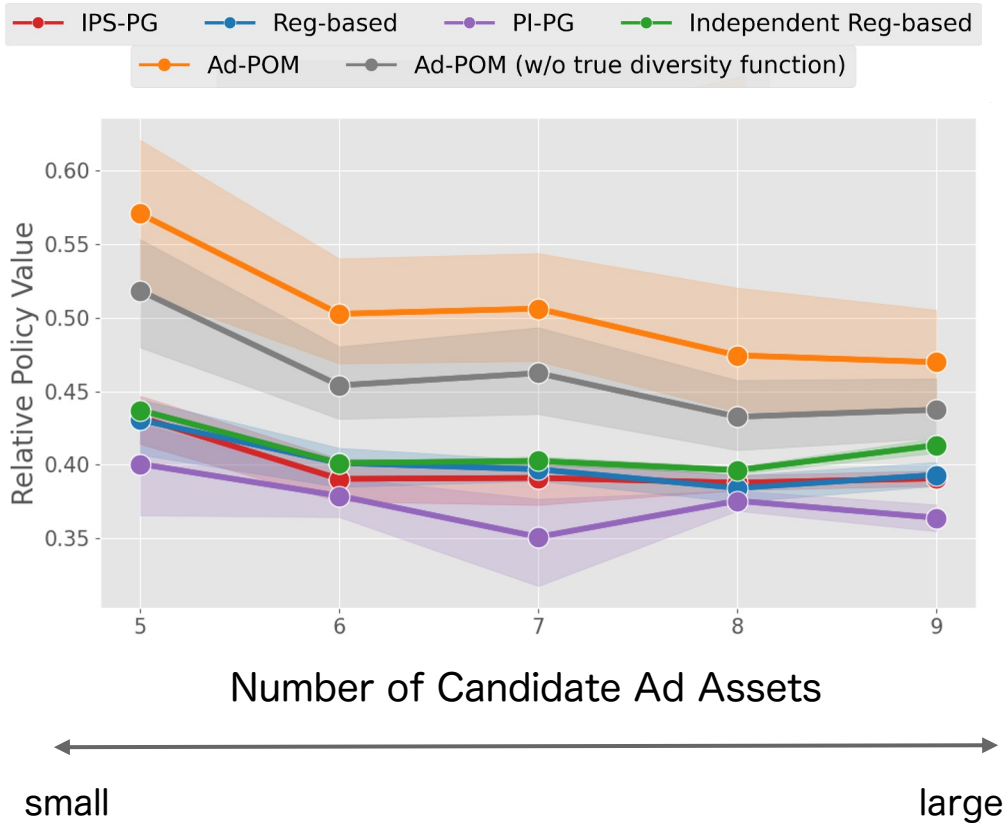
Experiment

Experiment with Varying Degrees of Interaction



- Exact approach (IPS-PG, Reg-based)
  - suffer from large action space
- Independent approach (PI-PG, Independent Reg-based)
  - high performance when 0.0
  - gradually decreasing as the interaction effect increases
- Ad-POM, Ad-POM (w/o ....)
  - always learns high-performing policies
  - not need true structure of q(x,a)

Experiment with Varying # of Ad Assets



- Exact approach (IPS-PG, Reg-based)
  - suffer from large action space
- Independent approach (PI-PG, Independent Reg-based)
  - low performance due to ignoring the interaction effect
- Proposed method (Ad-POM)
  - large variance reduction due to meta information importance weight

Conclusion

- We study ad asset portfolio optimization (Ad-POP) for the first time
- Existing methods have several limitations
  - Exact approach suffers from high variance
  - Independent approach suffers from high bias due to ignoring interaction effects
- We proposed the Ad-POM algorithm
  - decomposition of a policy into 1st- and 2nd-policies
  - can avoid the bias and variance issues of the existing methods
- Experiments show that Ad-POM can learn effective policies with challenging situation where the baselines fail