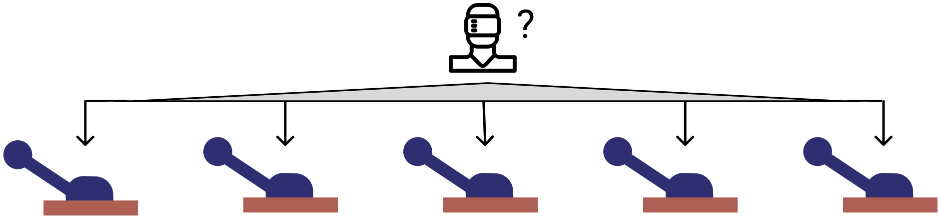
Uplifting Bandits

Yu-Guan Hsieh, Shiva Kasiviswanathan, Branislav Kveton

NeurIPS

December, 2022

Multi-Armed Bandits: A Simple Model for Online Decision Making



- Learner repeatedly takes actions (pulls arms)
- Learner receives rewards from the chosen actions
- The goal is to maximize the cumulative rewards
- Applications: Marketing, Online advertisement, Clinical trials, Portfolio selections, etc

Uplift Modeling versus Multi-Armed Bandits

	Uplift Modeling	**Multi-Armed Bandits**
Setup	Offline	Online
Challenges	Confounding bias Model evaluation	Exploration-exploitation trade-off Uncertainty estimates
Advantage	Statistical power	Data efficiency
Objective	Profit maximization / Finding good treatments	

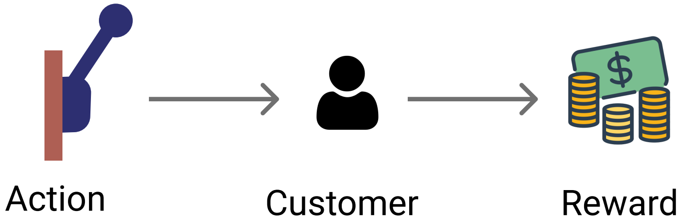
Uplift Modeling versus Multi-Armed Bandits

	Uplift Modeling	**Multi-Armed Bandits**
Setup	Offline	Online
Challenges	Confounding bias Model evaluation	Exploration-exploitation trade-off Uncertainty estimates
Advantage	Statistical power	Data efficiency
Objective	Profit maximization / Finding good treatments	

Uplift Modeling versus Multi-Armed Bandits

	Uplift Modeling	**Multi-Armed Bandits**
Setup	Offline	Online
Challenges	Confounding bias Model evaluation	Exploration-exploitation trade-off Uncertainty estimates
Advantage	Statistical power	Data efficiency
Objective	Profit maximization / Finding good treatments	

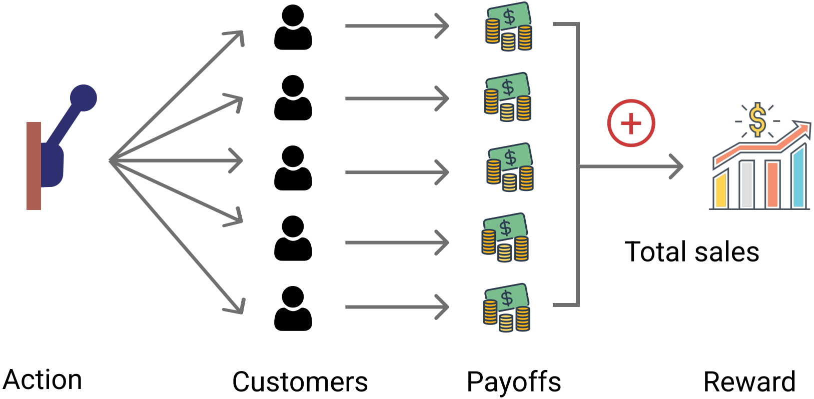
From Multi-Armed Bandits to Uplifting Bandits



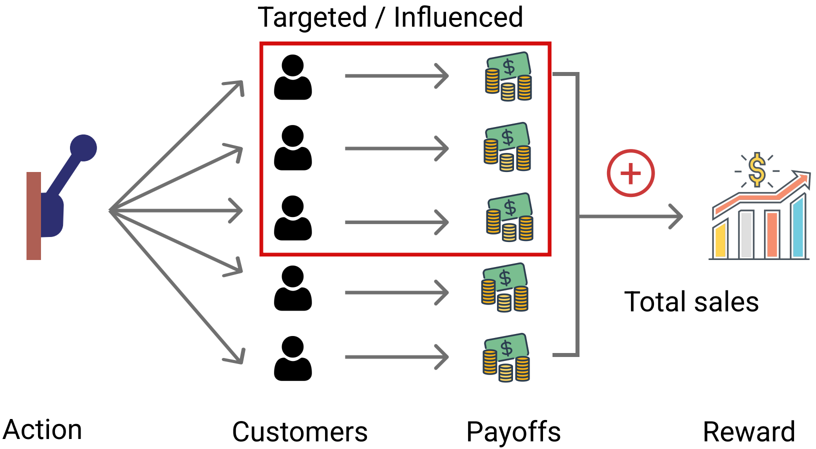
Incorporating uplift: use uplift as reward

- Take costs of actions into account
- Simply subtracting a baseline can lead to better performance in practice because the model is never perfect

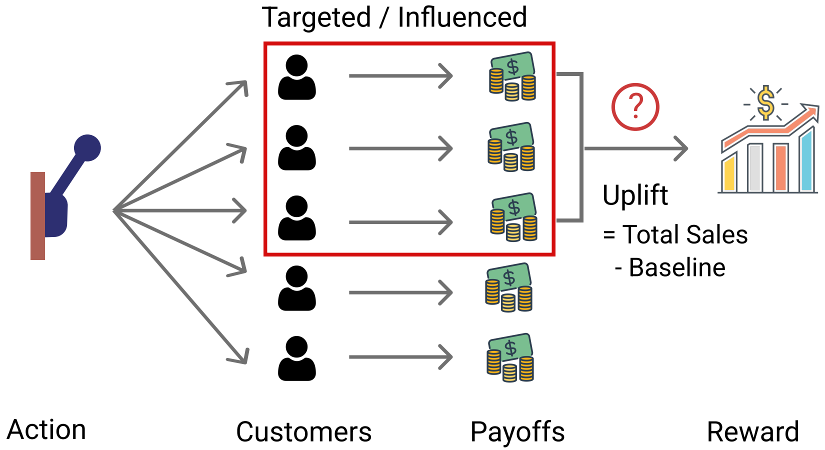
From Multi-Armed Bandits to Uplifting Bandits



From Multi-Armed Bandits to Uplifting Bandits

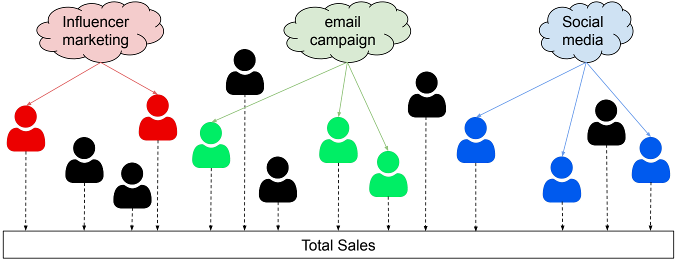


From Multi-Armed Bandits to Uplifting Bandits



Motivating Example in Online Marketing

- Marketing strategies: email campaign, influencer marketing, social media
- Different customers are sensitive to different strategies
- The reward is summed over all the customers
- We observe how much each customer spends



Formulation

Stochastic Bandits

- K actions: $\mathcal{A} = \{1, \ldots, K\}$
- T rounds: $[T] = \{1, \ldots, T\}$
- When action a_t is taken, the reward r_t is drawn from $\mathcal{D}^a$ (distribution over $\mathbb{R}$)

Uplifting Bandits

- K actions, T rounds
- m variables, $\mathcal{V} = \{1, \ldots, m\}$
- When action a_t is taken, the payoffs of the variables $y_t = (y_t(i))_{i \in \mathcal{V}}$ are drawn from $\mathcal{P}^{a_t}$ (distribution over $\mathbb{R}^m$), and the reward is $r_t = \sum_{i \in \mathcal{V}} y_t(i)$

Key Assumptions

- **Limited Number of Affected Variables** $L \ll m$.
  - $\mathcal{P}^0$: Baseline distribution $\mathcal{P}^a$ and $\mathcal{P}^0$ coincide
  - L^a: number of variables affected by action a
  - L: upper bound on number of affected variables, i.e., $L \geq \max_{a \in \mathcal{A}} L^a$

Key Assumptions

- **Limited Number of Affected Variables** $L \ll m$.
  - $\mathcal{P}^0$: Baseline distribution $\mathcal{P}^a$ and $\mathcal{P}^0$ coincide
  - L^a: number of variables affected by action a
  - L: upper bound on number of affected variables, i.e., $L \geq \max_{a \in \mathcal{A}} L^a$
- **Observability of Individual Payoff.** All of $(y_t(i))_{i \in \mathcal{V}}$ is observed

Key Assumptions

- **Limited Number of Affected Variables** $L \ll m$.
  - $\mathcal{P}^0$: Baseline distribution $\mathcal{P}^a$ and $\mathcal{P}^0$ coincide
  - L^a: number of variables affected by action a
  - L: upper bound on number of affected variables, i.e., $L \geq \max_{a \in \mathcal{A}} L^a$
- **Observability of Individual Payoff.** All of $(y_t(i))_{i \in \mathcal{V}}$ is observed
- Assumptions on payoff noise. 1-sub-Gaussian

Overview of Our Results

Regret compares an algorithm against the algorithm that constantly takes the best action

Algorithm	UCB	UpUCB (b)	UpUCB	UpUCB-nAff
Affected variables known	No	Yes	Yes	No
Baseline payoffs known	No	Yes	No	No
Regret Bound	$\frac{Km^2}{\Delta}$	$\frac{KL^2}{\Delta}$		$\frac{KL^2}{\Delta}$

Key takeaway: **focusing on the uplift gives much smaller regret**

- K: number of actions
- m: number of variables
- L: upper bound on number of affected variables
- Δ: minimum non-zero suboptimality gap

What You Can Expect From Our Paper

- Introduce uplifting bandits to formally capture the benefit of estimating uplift in the bandit setup

What You Can Expect From Our Paper

- Introduce uplifting bandits to formally capture the benefit of estimating uplift in the bandit setup
- Provide regrets upper bounds using variants of UCB

What You Can Expect From Our Paper

- Introduce uplifting bandits to formally capture the benefit of estimating uplift in the bandit setup
- Provide regrets upper bounds using variants of UCB
- Matching lower bounds that justify the necessity of the assumptions

What You Can Expect From Our Paper

- Introduce uplifting bandits to formally capture the benefit of estimating uplift in the bandit setup
- Provide regrets upper bounds using variants of UCB
- Matching lower bounds that justify the necessity of the assumptions
- Discussion on contextual extension