Uplifting Bandits

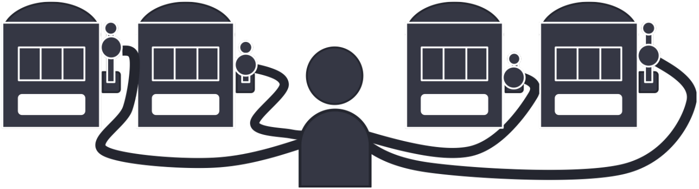
Yu-Guan Hsieh, Shiva Kasiviswanathan, Branislav Kveton

ECML/PKDD'22 Uplift Modeling Workshop

September 19th, 2022

Multi-Armed Bandits

- Learner repeatedly takes actions (pulls arms)
- Learner receives rewards from the chosen actions
- The goal is to maximize the cumulative rewards



Uplift Modeling versus Multi-Armed Bandits

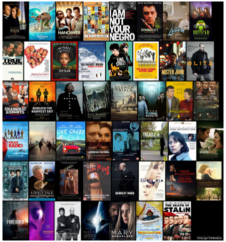
	Uplift Modeling	**Multi-Armed Bandits**
Setup	Offline	Online
Challenges	Confounding bias Model evaluation	Exploration-exploitation trade-off Uncertainty estimates
Advantage	Statistical power	Data efficiency
Objective	Profit maximization / Finding good treatments	

Uplift Modeling versus Multi-Armed Bandits

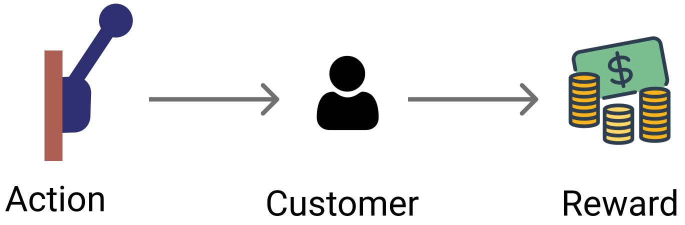
	Uplift Modeling	**Multi-Armed Bandits**
Setup	Offline	Online
Challenges	Confounding bias Model evaluation	Exploration-exploitation trade-off Uncertainty estimates
Advantage	Statistical power	Data efficiency
Objective	Profit maximization / Finding good treatments	

Applications

- Marketing, Online advertisement, Movie Recommendation, Clinical Trials, Portfolio Selections . . .



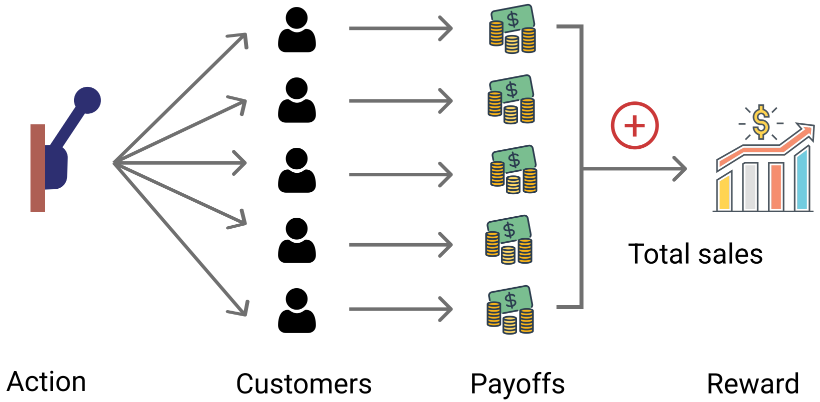
From Multi-Armed Bandits to Uplifting Bandits



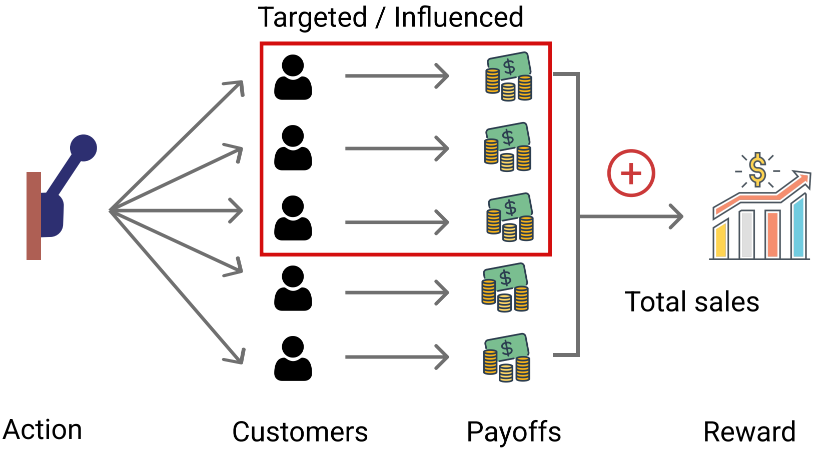
Incorporating uplift: use uplift as reward

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

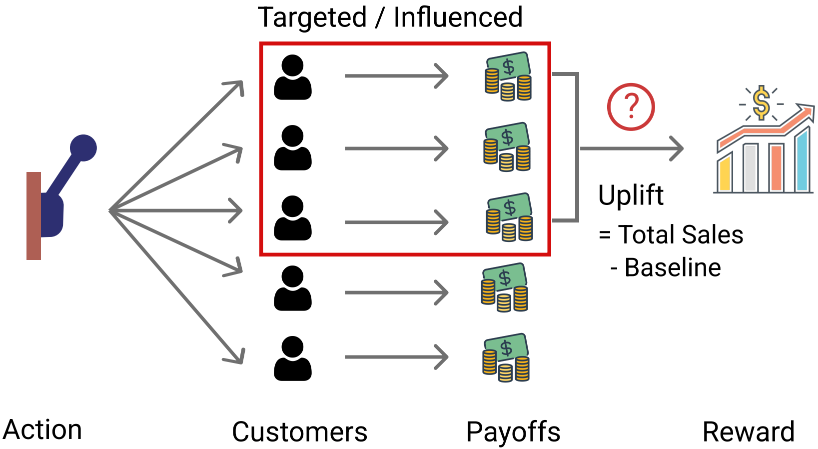
From Multi-Armed Bandits to Uplifting Bandits



From Multi-Armed Bandits to Uplifting Bandits

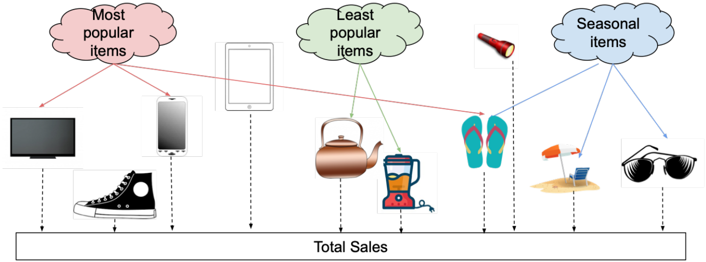


From Multi-Armed Bandits to Uplifting Bandits



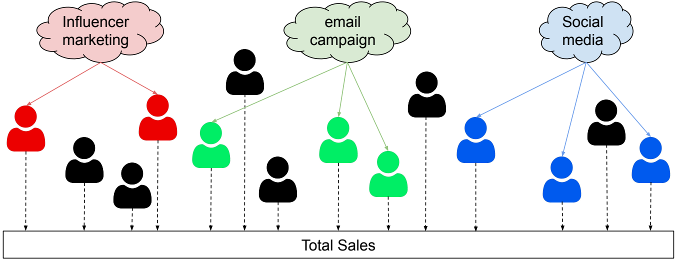
Motivating Example in Product Discount

- Consider different discount groups: most popular, least popular, seasonal
- Different groups contain different products
- The reward is summed over all the products
- We observe how much sales each product brings



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

Uplifting Bandits

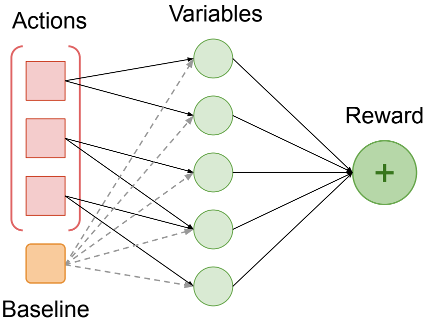
- K actions, T rounds
- m variables, $\mathcal{V} = \{1, \dots, m\}$
- When action a_t is taken, the payoffs of the variables $y_t = (y_t(i))_{i \in \mathcal{V}}$ are drawn from $\mathcal{P}^{a_t}$ (distribution over $\mathbb{R}^m$), and the reward is $r_t = \sum_{i \in \mathcal{V}} y_t(i)$

Key Assumptions

- **Limited Number of Affected Variables.**
  - $\mathcal{P}^0$: Baseline distribution
  - $\mathcal{V}^a$: variables affected by action a; $\mathcal{P}^a$ and $\mathcal{P}^0$ coincide on $\overline{\mathcal{V}^a} \coloneqq \mathcal{V} \setminus \mathcal{V}^a$
  - $L^a = \mathrm{card}(\mathcal{V}^a)$: 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

X is σ-sub-Gaussian if $\mathbb{E}[\exp(\gamma X)] \leq \exp(\sigma^2 \gamma^2 / 2), \forall \gamma$

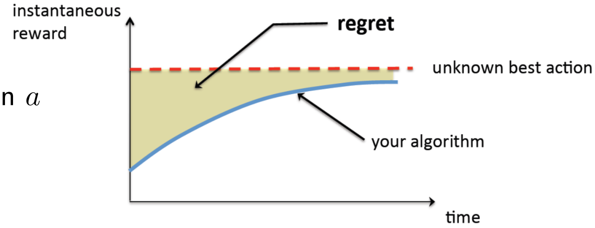
An Example



$K = 3,\ m = 5,\ L^a \equiv 2$

	baseline	arm 1	arm 2	arm 3
var. 1	0.3	0.4	0.3	0.3
var. 2	0.5	0.7	0.5	0.5
var. 3	0	0	0.2	0
var. 4	0.9	0.9	0.7	1
var. 5	0.5	0.5	0.5	0.3
reward	2.2	2.5	2.2	2.1
uplift	-	0.3	0	-0.1

Regret



- $r^a = \mathbb{E}_{r\sim\mathcal{D}^a}[r]$: expected reward of action a
- Optimal action is $a^\star \in \arg\max_{a\in\mathcal{A}} r^a$
- Optimal reward is $r^\star = r^{a^\star} = \max_{a\in\mathcal{A}} r^a$
- Regret compares the expected performance between the learner and the optimal action

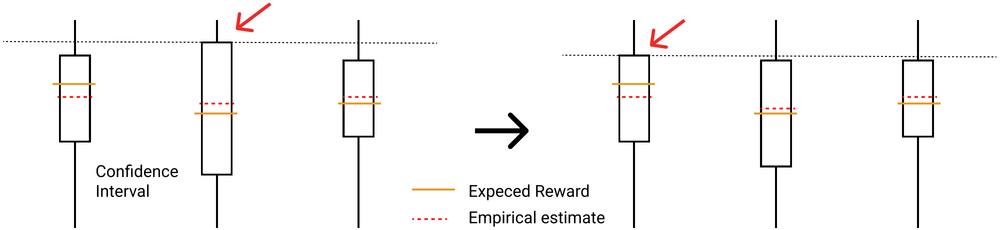
$$\mathrm{Reg}_T = r^\star T - \sum_{t=1}^{T} r^{a_t} = \sum_{a\in\mathcal{A}} \underbrace{\sum_{t=1}^{T} \mathbb{1}\{a_t = a\}}_{N_T^a} \underbrace{(r^\star - r^a)}_{\Delta^a},$$

- Δ^a is the suboptimality gap of a, Δ is minimum non-zero suboptimality gap

Plan

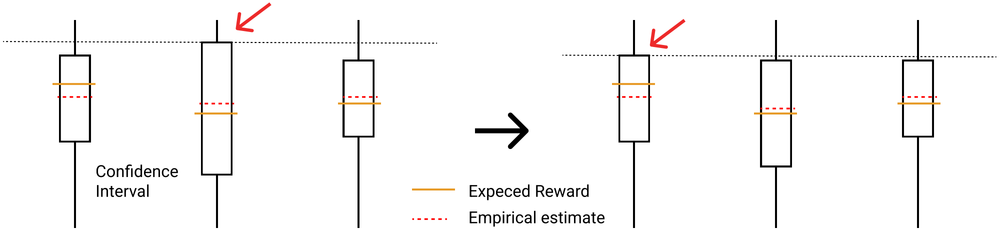
UCB– Optimism in Face of Uncertainty

- Empirical estimate of reward: $\hat{r}_t^a = \sum_{s=1}^{t} r_s \mathbb{1}\{a_s = a\} / \max(1, N_t^a)$
- Width of confidence interval: $c_t^a = \sigma\sqrt{2\log(1/\delta')/N_t^a}$, where σ is the scale of noise
- Take action with the highest upper confidence bound (UCB): $U_t^a = \hat{r}_t^a + c_t^a$



UCB– Why does it Work?

- Choose the seemly best arm for exploitation
- Add the confidence interval c_t^a for exploration: the fewer number of times an arm is pulled, the higher its UCB index
- With enough data, bad arms get pulled less and less frequently



UCB for Uplifting Bandits

The number of time a suboptimal action is taken scales with the noise in its reward

- The reward is $r_t = \sum_{i \in \mathcal{V}} y_t(i)$
- Each $y_t(i)$ is 1-sub-Gaussian (assumption)
- Therefore r_t is m-sub-Gaussian (we do not assume independence)
- The regret is in $\mathcal{O}(Km^2 \log T/\Delta)$ $\longrightarrow$ substantial when m is large

Fixing UCB: Looking at Uplift

- For $a \in \mathcal{A}_0 \coloneqq \mathcal{A} \cup \{0\}$, let $y^a = (y^a(i))_{i\in\mathcal{V}}$ follow distribution $\mathcal{P}^a$
- Define expected payoffs $\mu^a(i) = \mathbb{E}[y^a(i)]$; Baseline payoff vector is $\mu^0 = (\mu^0(i))_{i\in\mathcal{V}}$
- Individual uplift: $\mu^a_{\mathsf{up}}(i) = \mu^a(i) - \mu^0(i)$
- The (total) uplift of an action is

$$r^a_{\mathsf{up}} = \sum_{i\in\mathcal{V}^a} \mu^a_{\mathsf{up}}(i) = \sum_{i\in\mathcal{V}^a} (\mu^a(i) - \mu^0(i)) = \sum_{i\in\mathcal{V}} \mu^a_{\mathsf{up}}(i) = r^a - r^0.$$

$r^0 = \sum_{i\in\mathcal{V}} \mu^0(i)$ is the reward of the baseline

- We can rewrite $\mathrm{Reg}_T = r^\star_{\mathsf{up}} T - \sum_{t=1}^{T} r^{a_t}_{\mathsf{up}}$, $\Delta^a = r^\star_{\mathsf{up}} - r^a_{\mathsf{up}}$, where $r^\star_{\mathsf{up}} = r^{a^\star}_{\mathsf{up}} = \max_{a\in\mathcal{A}} r^a_{\mathsf{up}}$

UpUCB (b)– UCB for Estimating the Uplifts

The learner knows

1. Baseline payoffs $\mu^0 = (\mu^0(i))_{i \in \mathcal{V}}$
2. The sets of affected variables $(\mathcal{V}^a)_{a \in \mathcal{A}}$

- UCB applied to transformed rewards $r'_t = \sum_{i \in \mathcal{V}_{a_t}} (y_t(i) - \mu^0(i))$

  r'_t can be computed thanks to the learner's prior knowledge
- $\mathbb{E}[r'_t] = r^{a_t}_{\mathrm{up}}$, and thus the regret is not modified
- $r'_t = \sum_{i \in \mathcal{V}_{a_t}} (y_t(i) - \mu^0(i))$ is L^{a_t}-sub-Gaussian ; Regret in $\mathcal{O}(KL^2 \log T/\Delta)$

UpUCB (b)– UCB for Estimating the Uplifts

The learner knows

$$\left.\begin{array}{l} \text{① Baseline payoffs } \mu^0 = (\mu^0(i))_{i\in\mathcal{V}} \\ \text{② The sets of affected variables } (\mathcal{V}^a)_{a\in\mathcal{A}} \end{array}\right\} \text{not always realistic}$$

- UCB applied to transformed rewards $r'_t = \sum_{i\in\mathcal{V}_{a_t}} (y_t(i) - \mu^0(i))$

  r'_t can be computed thanks to the learner's prior knowledge
- $\mathbb{E}[r'_t] = r^{a_t}_{\text{up}}$, and thus the regret is not modified
- $r'_t = \sum_{i\in\mathcal{V}_{a_t}} (y_t(i) - \mu^0(i))$ is L^{a_t}-sub-Gaussian ; Regret in $\mathcal{O}(KL^2 \log T/\Delta)$

Overview of Our Results

Algorithm	UCB	UpUCB (b)	UpUCB	UpUCB-nAff	UpUCB-iLift
Affected variables known	No	Yes	Yes	No	No
Baseline payoffs known	No	Yes	No	No	No
Regret Bound	$\frac{Km^2}{\Delta}$	$\frac{KL^2}{\Delta}$		$\frac{KL^2}{\Delta}$	$\frac{K\,\mathrm{clip}(\Delta/\Delta_{\mathsf{up}}, L, m)^2}{\Delta}$

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
- Δ_{up}: a lower bound on individual uplift

Lower Bounds– Justifying the Assumptions

Let π be a *consistent* algorithm the is provided the knowledge about $\mathcal{P}^0$ and $(\mathcal{V}^a)_{a\in\mathcal{A}}$. If any of the follow holds

1. All actions affect all variables
2. Only the reward is observed (but not individual payoffs of the variables)
3. π does not use any information about $(\mathcal{V}^a)_{a\in\mathcal{A}}$

Then the regret of π is $\boxed{\Omega(Km^2 \log T/\Delta)}$

UpUCB– When Baseline is Unknown

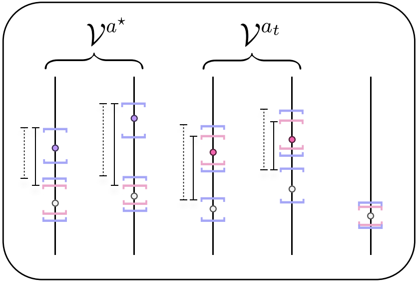
- Estimate the baseline from the rounds that i is not affected (i.e., $i \notin \mathcal{V}^a$)

$$N_t^0(i) = \sum_{s=1}^{t} \mathbb{1}\{i \notin \mathcal{V}^{a_s}\}, \quad c_t^0(i) = \sqrt{\frac{2\log(1/\delta')}{N_t^0(i)}}, \quad \hat{\mu}_t^0(i) = \frac{\sum_{s=1}^{t} y_s(i)\,\mathbb{1}\{i \notin \mathcal{V}^{a_s}\}}{\max(1, N_t^0(i))}.$$

- Define UCB indices $U_t^a = \hat{\mu}_{t-1}^a + c_{t-1}^a$ and $U_t^0(i) = \hat{\mu}_{t-1}^0(i) + c_{t-1}^0(i)$
- Pull arm with highest uplifting index $\boxed{\tau_t^a = \sum_{i \in \mathcal{V}^a} (U_t^a(i) - U_t^0(i))}$ [not optimistic]

UpUCB– Why does it work?

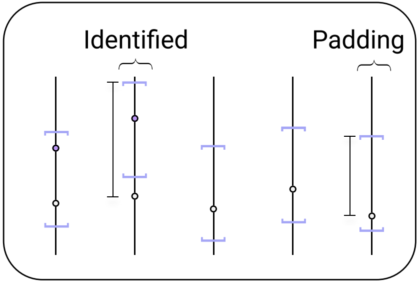
- When action a is taken, we learn about the baseline payoffs of $i \notin \mathcal{V}^a$
- An arm that has not been pulled many times has large $U_t^a(i)$ and small $U_t^0(i)$ for $i \in \mathcal{V}^a$
  $\rightarrow$ implying large uplifting index τ_t^a



- If suboptimal a action is taken
  - τ_t^a decreases, since all $U_t^a(i)$ for $i \in \mathcal{V}^a$ do
  - $\tau_t^{a^\star}$ increases, since $U_t^0(i)$ decrease for any i affected by $a^\star$ but not a

UpUCB-nAff (b)– Known Baseline and Known L

- L upper bound on number of affected variables
- Construct $\tau_t^a = \sum_{i \in \widehat{\mathcal{V}}_t^a \cup \mathcal{L}_t^a} \rho_t^a(i)$ in two steps



1. Identification of affected

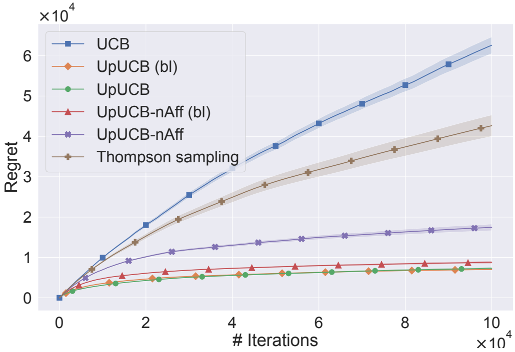
$$\widehat{\mathcal{V}}_t^a = \{i \in \mathcal{V} : \mu^0(i) \notin \mathcal{C}_t^a(i)\}$$

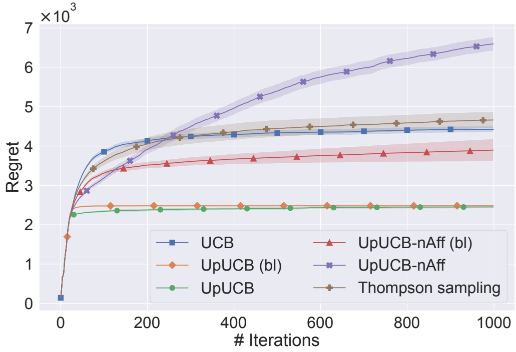
2. Optimistic padding
$[\rho_t^a(i) = \hat{\mu}_{t-1}^a(i) + c_{t-1}^a - \mu^0(i))]$

$$\mathcal{L}_t^a = \underset{\substack{\mathcal{L} \subseteq \mathcal{V} \setminus \widehat{\mathcal{V}}_t^a \\ \mathrm{card}(\mathcal{L}) = L_t^a}}{\arg\max} \sum_{i \in \mathcal{L}} \rho_t^a(i) \quad \text{where} \quad L_t^a = \max(0, L - \mathrm{card}(\widehat{\mathcal{V}}_t^a))$$

Plan

Experiments





Synthetic data
Gaussian noises
$K = 10, m = 100, L^a \equiv 10, \Delta \sim 0.2$

Constructed with Criteo Uplift Modeling Dataset
Bernoulli noises, independent across variables
$K = 20, m = 10^5, L = 12654, \Delta \sim 30$

UCB and Thomposon sampling with Gaussian Prior only use the rewards

Conclusion

- Introduce uplifting bandits to formally capture the benefit of estimating uplift in the bandit setup
- Provide optimal regrets bounds using variants of UCB
- Contextual extension are also discussed in our work:
  Associate each variable with a feature vector $x_t(i) \in \mathbb{R}$

Perspectives– Uplift modeling and causal inference

- From an uplift viewpoint: Can we make use of more complex uplift modeling approach in the procedure? (Need for accounting the uncertainty)
- From a causal viewpoint: Can the method be generalized? View abstractly, the reward is generated from an underlying causal mechanism and each action only affects a small number of the involving variables.
- Use of (confounded) offline data for warm-up

Perspectives– Multi-armed bandits

- Misspecified model: Small impact on $\overline{\mathcal{V}^a}$
- Contexts, possibility of taking multiple actions in one round
- Use of other algorithms: Thompson sampling, information directed sampling
- Dealing with non-stationarity and the adversarial setup

Perspectives– Multi-armed bandits

- Misspecified model: Small impact on $\overline{\mathcal{V}^a}$
- Contexts, possibility of taking multiple actions in one round
- Use of other algorithms: Thompson sampling, information directed sampling
- Dealing with non-stationarity and the adversarial setup

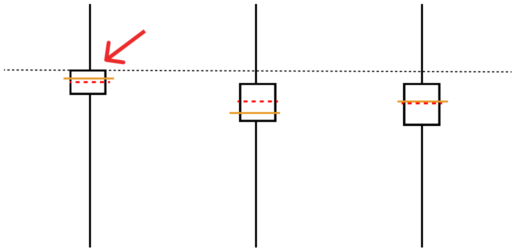
– H., Kasiviswanathan, S. P., & Kveton, B. (2022). Uplifting Bandits. Accepted at NeurIPS 2022.

Thank you for your attention

Analyses

UCB Analysis in a Nutshell

- Assume all expected rewards lie in their confidence intervals
- Then a suboptimal action is not taken anymore if $2c_t^a < \Delta^a$
- This shows $N_T^a \leq \dfrac{8\sigma^2 \log(1/\delta')}{(\Delta^a)^2} + 1$
- Conclude with $\mathrm{Reg}_T = \sum_{a \in \mathcal{A}} N_t^a \Delta^a$



The number of time a suboptimal action is taken scales with the noise in its reward

UpUCB– Regret

- The regret is in $\mathcal{O}(KL^2 \log T/\Delta)$ because
  1. Only $\mathcal{O}(L)$ variables are involved in the estimates
  2. The confidence intervals of the baseline payoffs are small
- If action a is taken at round t then

$$\sum_{i \in \mathcal{V}^a} U_t^a(i) + \sum_{i \in \mathcal{V}^{a^\star} \setminus \mathcal{V}^a} U_t^0(i) \geq \sum_{i \in \mathcal{V}^{a^\star}} U_t^{a^\star}(i) + \sum_{i \in \mathcal{V}^a \setminus \mathcal{V}^{a^\star}} U_t^0(i).$$

similar structure as UCB

UpUCB-nAff (b)– Regret

The regret is in $\mathcal{O}(KL^2 \log T/\Delta)$ because

1. Only $\mathcal{O}(L)$ variables are involved in the estimates
2. If a variable is identified, it's like in UpUCB (b)
3. If a variable is not identified, $\hat{\mu}^a_{t-1}(i)$ and $\mu^0(i)$ are close, so $\rho^a_t(i)$ is small

Indeed, if $\widehat{\mathcal{V}}^a_t \subseteq \mathcal{V}^a$, then

$$\tau^a_t = \sum_{i \in \widehat{\mathcal{V}}^a_t} \rho^a_t(i) + \sum_{i \in \mathcal{L}^a_t} \rho^a_t(i) = \underbrace{\sum_{i \in \mathcal{V}^a} \rho^a_t(i)}_{\text{UpUCB (b)}} + \underbrace{\sum_{i \in \mathcal{L}^a_t \setminus \mathcal{V}^a} \rho^a_t(i) - \sum_{i \in \mathcal{V}^a \setminus \widehat{\mathcal{V}}^a_t} \rho^a_t(i)}_{\text{small}}.$$

UpUCB-nAff– Unknown Baseline and Known L

- Unclear how baseline can be estimated in this case
- Key observation: The payoffs of any action can be a baseline because μ^a and $\mu^{a'}$ only differ on $\mathcal{V}^a \cup \mathcal{V}^{a'}$, and $\mathrm{card}(\mathcal{V}^a \cup \mathcal{V}^{a'}) \leq 2L$
- Take the payoffs of an action as baseline at each round

UpUCB-nAff– Regret

The regret is in $\mathcal{O}(KL^2 \log T/\Delta)$ because

1. Only $\mathcal{O}(L)$ variables are involved in the estimates
2. The confidence intervals of the chosen action b_t is small as b_t is an action that has been taken the most number of times

Lower Bound on Individual Uplift

- $\Delta_{\mathsf{up}} > 0$ such that for all $a \in \mathcal{A}$ and $i \in \mathcal{V}^a$, $|\mu^a(i) - \mu^0(i)| \geq \Delta_{\mathsf{up}}$
- If we know baseline and Δ_{up}, we know how many times we need to take an action to find all the affected variables
- By combining UCB with this idea, we get a regret in $K\,\mathrm{clip}(\Delta/\Delta_{\mathsf{up}}, L, m)^2/\Delta$

Pseudo Code

UpUCB (b)– UCB for Estimating the Uplifts

Algorithm UpUCB (b)

1: **Input:** Error probability δ', Baseline payoffs μ^0, Sets of affected variables $\{\mathcal{V}^a : a \in \mathcal{A}\}$
2: **Initialization:** Take each action once
3: **for** $t = K+1, \ldots, T$ **do**
4: **for** $a \in \mathcal{A}$ **do**
5: Compute empirical estimate $\hat{\mu}_t^a(i) = \sum_{s=1}^{t} y_s(i)\,\mathbb{1}\{a_s = a\}/\max(1, N_t^a)$
6: Compute widths of confidence interval $c_t^a = \sqrt{2\log(1/\delta')/N_t^a}$
7: Compute uplifting index $\tau_t^a \leftarrow \sum_{i \in \mathcal{V}^a} (\hat{\mu}_{t-1}^a(i) + c_{t-1}^a - \mu^0(i))$
8: Select action $a_t \in \arg\max_{a \in \mathcal{A}} \tau_t^a$

UpUCB– When Baseline is Unknown

Algorithm UpUCB

Input: Error probability δ', the sets of variables each action affects $\{\mathcal{V}^a : a \in \mathcal{A}\}$
Initialization: Take each action once
for $t = K + 1, \ldots, T$ **do**
    Compute the UCB indices
    For $a \in \mathcal{A}$, set $\tau_t^a \leftarrow \sum_{i \in \mathcal{V}^a} (U_t^a(i) - U_t^0(i))$
    Select action $a_t \in \arg\max_{a \in \mathcal{A}} \tau_t^a$

UpUCB-nAff

Algorithm UpUCB-nAff (Input: δ' and L; Initialization: take each action once)

for $t = K+1, \ldots, T$ **do**
  Choose $b_t \in \arg\max_{a \in \mathcal{A}} N_{t-1}^{a}$
  Compute UCBs and confidence intervals
  for $a \in \mathcal{A}$ **do**
    Set $\widehat{\mathcal{V}}_t^a \leftarrow \{i \in \mathcal{V} : \mathcal{C}_t^a(i) \cap \mathcal{C}_t^{b_t}(i) = \varnothing\}$
    For $i \in \mathcal{V}$, compute $\rho_t^a(i) \leftarrow U_t^a(i) - U_t^{b_t}(i)$
    Set $\mathcal{L}_t^a \leftarrow \operatorname*{arg\,max}_{\substack{\mathcal{L} \subseteq \mathcal{V} \setminus \widehat{\mathcal{V}}_t^a \\ \operatorname{card}(\mathcal{L}) \le L_t^a}} \sum_{i \in \mathcal{L}} \rho_t^a(i)$, where $L_t^a \leftarrow \max(0, 2L - \operatorname{card}(\widehat{\mathcal{V}}_t^a))$
    Compute uplifting index $\tau_t^a \leftarrow \sum_{i \in \widehat{\mathcal{V}}_t^a \cup \mathcal{L}_t^a} \rho_t^a(i)$
  Select action $a_t \in \arg\max_{a \in \mathcal{A}} \tau_t^a$