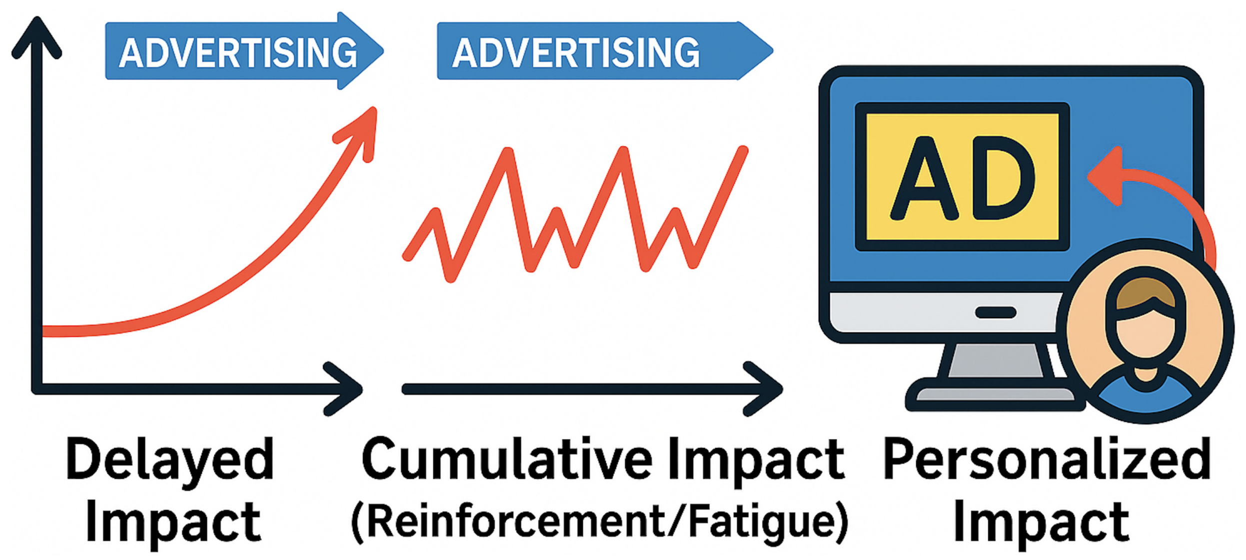
Learning Personalized Ad Impact via Contextual Reinforcement Learning under Delayed Rewards

Yuwei Cheng[1] Zifeng Zhao[2] Haifeng Xu [3]

[1]Department of Statistics, University of Chicago [2]Mendoza College of Business, University of Notre Dame [3]Department of Computer Science, University of Chicago

Motivation: Jointly Model Key Ads Impacts



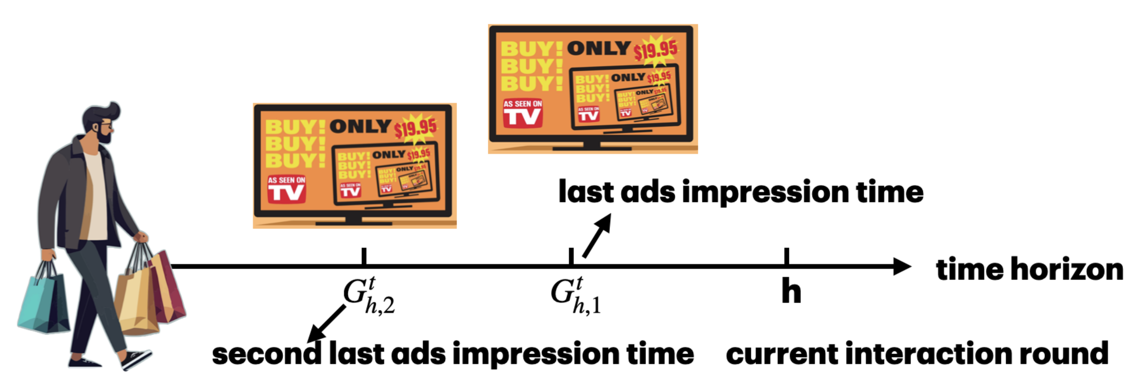
Contextual MDP with Delayed Rewards

Observation: The expected product conversion rate μ_h^t at round h for customer t with context $\mathbf{x}_t$ follows

$$\mu_h^t = \begin{cases} d_{S_{h,1}^t} \mathbf{x}_t \top \boldsymbol{\theta}_{S_{h,2}^t} & \text{if } \boldsymbol{o}_h^t = 0 \\ \mathbf{x}_t^\top \boldsymbol{\theta}_{S_{h,1}^t} & \text{otherwise.} \end{cases}$$

The observed product conversion $\mathbf{y}_h^t \sim \text{Poi}(\mu_h^t)$. $d_{S_{h,1}^t}$: delayed advertisements impact. $\boldsymbol{o}_h^t = 1$ if the bid is won and $\boldsymbol{o}_h^t = 0$ otherwise.

State: $S_h^t = [S_{h,1}^t, S_{h,2}^t]$ with $S_{h,1}^t = h - G_{h,1}^t$ and $S_{h,2}^t = G_{h,1}^t - G_{h,2}^t$



Transition: For a given bid amount $\mathbf{a}_h^t$, $\mathbb{P}(\boldsymbol{o}_h^t = 1) = \mathcal{F}_h(\mathbf{a}_h^t, \mathbf{x}_t)$, where $\mathcal{F}_h : [0, \boldsymbol{B}_\mathcal{A}] \times \mathcal{X} \to [0, 1]$ represents CDF of the Highest Other Bids (HOB).

Lognormal Distribution of HOB: $\log(m_h^t) \sim \mathcal{N}(\langle \mathbf{x}_t, \boldsymbol{\beta}_h \rangle, \sigma_h^2)$.

Reward: $R_h^t(S_h^t, \mathbf{a}_h^t, \mathbf{x}_t) = d_{S_{h,1}^t} \langle \boldsymbol{\theta}_{S_{h,2}^t}, \mathbf{x}_t \rangle (1 - \mathcal{F}_h(\mathbf{a}_h^t, \mathbf{x}_t)) + (\langle \boldsymbol{\theta}_{S_{h,1}^t} \mathbf{x}_t \rangle - p_h(\mathbf{a}_h^t, \mathbf{x}_t)) \mathcal{F}_h(\mathbf{a}_h^t, \mathbf{x}_t)$. $p_h(\mathbf{a}_h^t, \mathbf{x}_t)$ is the expected second highest payment.

Learning Goal: Regret Minimization

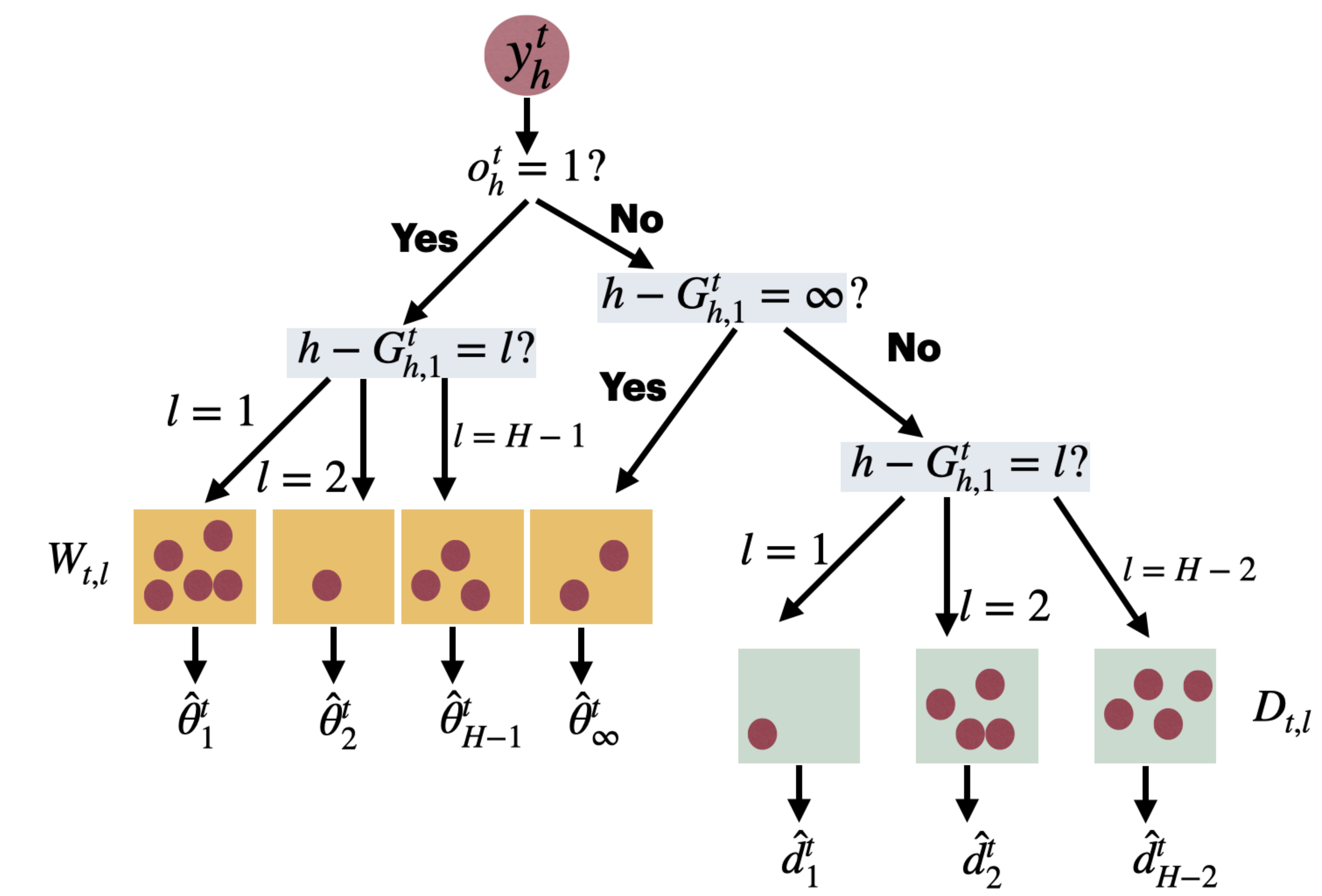
$\text{Reg}_T := \sum_{t=1}^T \text{OPT}(\Theta, \mathbf{x}_t, \mathcal{F}^{\mathbf{x}_t}) - \mathbb{E}_{\pi_1, \ldots, \pi_T \sim \mathcal{G}} \left[\sum_{t=1}^T R(\pi_t; \mathbf{x}_t, \Theta, \mathcal{F}^{\mathbf{x}_t}) \right]$.

$R(\pi_t; \mathbf{x}_t, \Theta, \mathcal{F}^{\mathbf{x}_t}) = \mathbb{E}_{S_h^t \sim \mathcal{P}^{\mathbf{x}_t}} [\sum_{h=1}^H R_h^t(S_h^t, \pi_t(S_h^t, \mathbf{x}_t), \mathbf{x}_t)]$.

Definition of Two Datasets: $\mathbf{W}_{t,l} = \{h | S_{h,1}^t = l, \boldsymbol{o}_h^t = 1\}$ and $\mathbf{D}_{t,l} = \{h | S_{h,1}^t = l, \boldsymbol{o}_h^t = 0\}$, for $l \in \mathcal{H} \setminus [-\infty]$. We define $\mathbf{W}_{t,-\infty} = \{h | S_{t,1}^t = \infty, \boldsymbol{o}_h^t = 0\}$ for the estimation of $\theta_{-\infty}$. The collection of observations across the first t customers are $\mathbf{W}_l^t = \{\mathbf{W}_{s,l}\}_{s=1}^t$ and $\mathbf{D}_l^t = \{\mathbf{D}_{s,l}\}_{s=1}^t$.

Efficient Estimation via Data Splitting (Purifying)

Figure 1. Dataset construction illustration. In essence, the data used to estimate each parameter θ_l depends solely on θ_l itself, not on other parameters, which enables us to isolate and control estimation error using purified data.



Two Stage Maximum Likelihood Estimator

Transition Parameters: $\hat{\boldsymbol{\beta}}_h^t : \left(\sum_{s=1}^t \mathbf{x}_s \mathbf{x}_s^\top + \lambda \mathbb{I}_d\right)^{-1} \left(\sum_{s=1}^t \mathbf{x}_s \log(m_h^s)\right)$;

$\hat{\sigma}_h^t : \sqrt{\frac{1}{t} \sum_{s=1}^t \left(\log(m_h^s) - \mathbf{x}_s^\top \hat{\boldsymbol{\beta}}_h^s\right)^2}$. **Cumulative Ads Impact:** $\hat{\boldsymbol{\theta}}_l^t = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \{\frac{1}{2} \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_l^{t-1}\|_{\mathbf{V}_l^t}^2 + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_l^{t-1})^\top \nabla \tilde{l}_t(\hat{\boldsymbol{\theta}}_l^{t-1})\}$. $\mathbf{V}_l^t = \mathbf{V}_l^{t-1} + \frac{1}{2} \mathbf{x}_t \mathbf{x}_t^\top$. $\mathbf{V}_l^0 = \mathbf{I}_d$ and $\nabla \tilde{l}_t(\hat{\boldsymbol{\theta}}_l^{t-1}) = (-\tilde{\mathbf{y}}_h^t + \mathbf{x}_t^\top \hat{\boldsymbol{\theta}}_l^{t-1}) \mathbf{x}_t$, where $\tilde{\mathbf{y}}_h^t = \mathbf{y}_h^t \mathbb{I}_{\|\mathbf{x}_t\|_{(\mathbf{V}_l^t)^{-1}} |\mathbf{y}_h^t| \leq \Gamma}$.

Maximizes log likelihood $\mathcal{L}\left(\mathbf{y}, d_l; \hat{\boldsymbol{\theta}}_l^s\right) = \sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \mathbf{y}_h^s \log\left(d_l \mathbf{x}_s^\top \hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s\right) - d_l \mathbf{x}_s^\top \hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s$ by plugging sequential first stage estimates $\hat{\boldsymbol{\theta}}_l^s$ we get the two stage estimator: **Delayed Impacts:** $\hat{d}_l^t = \frac{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \mathbf{y}_h^s}{\sum_{s=1}^t \sum_{h \in \mathbf{D}_{s,l}} \langle \hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s, \mathbf{x}_s \rangle}$.

Confidence Region for Delayed Impact $\hat{d}_l^t$

There exists positive constants $\boldsymbol{b}, \boldsymbol{B}_x, \boldsymbol{B}_\theta, \boldsymbol{B}_d, \boldsymbol{B}_{\boldsymbol{\beta}}, \bar{\sigma}$ such that $\|\mathbf{x}_t\|_2 \leq \boldsymbol{B}_x$, and $\boldsymbol{\theta}_l^\top \mathbf{x}_t \geq \boldsymbol{b}, \forall h \in [H], \forall t \in [T], \forall l \in \mathcal{H}$, $\|\boldsymbol{\theta}_l\|_2 \leq \boldsymbol{B}_\theta, \forall l \in \mathcal{H}$, $d_l \in [0, \boldsymbol{B}_d], \forall l \in \mathcal{H}_1$, and $\|\boldsymbol{\beta}_h\|_2 \leq \boldsymbol{B}_{\boldsymbol{\beta}}$, $\sigma_h \leq \bar{\sigma}, \forall h \in [H]$.

input datasets $\mathbf{D}_l^t, \{\mathbf{x}_s\}_{s=1}^t, \{\hat{\boldsymbol{\theta}}_{S_{h,2}^s}^s\}_{s=1}^t$ $\boldsymbol{b}, \boldsymbol{B}_x, \boldsymbol{B}_\theta, \boldsymbol{B}_d, d, T, H, \delta, \gamma$

1: Update the two-stage estimator $\hat{d}_l^t$ and

$$\mathcal{D}_l^t = \left\{ |\hat{d}_l^t - d_l| \leq \frac{4H\boldsymbol{B}_d \sqrt{d \log(1 + \frac{T}{2d})\gamma + \sqrt{2e\boldsymbol{B}_d \boldsymbol{B}_x \boldsymbol{B}_\theta \log(2/\delta)}}}{\boldsymbol{b}\sqrt{N_{t,l}}} \right\}$$

output $(\hat{d}_l^t, \mathcal{D}_l^t)$

Algo: Contextual RL via Delayed Poisson Reward

input $d, T, H, \boldsymbol{b}, \boldsymbol{B}_x, \boldsymbol{B}_\theta, \boldsymbol{B}_d, \boldsymbol{B}_\mathcal{A}, \delta, \gamma, \Gamma, \underline{n_l}$

for $t = 1$ to T **do**
  Obtain the context $\mathbf{x}_t$ for the arriving customer t
  if $t \leq (H+1)\underline{n_l}$ **then**
    /* Exploration */
    Compute $l = \lfloor \frac{t}{\underline{n_l}} \rfloor$.
    if $l = H$, set $\mathbf{a}_h^t = 0, \forall h \in [H]$; **else** set $\mathbf{a}_1^t = \mathbf{a}_{l+1}^t = \boldsymbol{B}_\mathcal{A}$, and $\mathbf{a}_h^t = 0$ for $\forall h \neq 1, l+1$
    for $h = 1$ to H **do** observe $\mathbf{y}_h^t$, m_h^t; then update $\hat{\boldsymbol{\beta}}_h^t$ and $\hat{\sigma}_h^t$
  else
    /* Exploitation */
    Update $\pi_t = \arg\max_\pi \max_{\tilde{\theta} \in \mathcal{C}_{t-1}} R(\pi; \tilde{\theta}, \hat{\mathcal{F}}_{t-1}^{\mathbf{x}_t})$
    for $h = 1$ to H **do**
      Observe S_h^t and take action $\mathbf{a}_h^t = \pi_t(S_h^t, \mathbf{x}_t)$
      Observe $\boldsymbol{o}_h^t$, m_h^t, and $\mathbf{y}_h^t$ and update $\hat{\boldsymbol{\beta}}_h^t$ and $\hat{\sigma}_h^t$
    end for
  end if
  /* Estimation */
  for $l \in \mathcal{H}$ **do**
    Update dataset $\mathbf{W}_l^t, \mathbf{D}_l^t$
    if $\mathbf{W}_{t,l}$ is nonempty, compute $\hat{\boldsymbol{\theta}}_l^t$ and $\mathcal{C}_l^t$; **else** set $\hat{\boldsymbol{\theta}}_l^t = \hat{\boldsymbol{\theta}}_l^{t-1}$ and $\mathcal{C}_l^t = \mathcal{C}_l^{t-1}$
    if $\mathbf{D}_{t,l}$ is nonempty, compute $\hat{d}_l^t$ and $\mathcal{D}_l^t$; **else** set $\hat{d}_l^t = \hat{d}_l^{t-1}$ and $\mathcal{D}_l^t = \mathcal{D}_l^{t-1}$
  end for
  Set $\mathcal{C}_t = \{\Theta \mid \{\boldsymbol{\theta}_l \in \mathcal{C}_l^t\} \cap \{d_l \in \mathcal{D}_l^t\}, \forall l \in \mathcal{H}\}$
end for

Main Result: Algo Achieves Near Optimal Regret

Theorem: Reg_T is $\mathcal{O}(dH^2 \sqrt{T \log(\frac{TH}{\delta})} \log(1 + \frac{T}{2d}))$, for any $\delta \geq \frac{6}{T^3}$, with probability at least $1 - \delta$.

Experiment Results:

Table 1. Average cumulative regret (± 0.5 standard deviation) and fitted regret order (T^α).

t	Algorithm 1	Aggressive Bid	Random Bid	Passive Bid
500	1770 [1174, 2379]	228 [121, 336]	1920 [1791, 2049]	4630 [3391, 5869]
5000	8465 [5585, 11345]	2373 [1243, 3502]	19285 [17873, 20697]	46179 [33850, 58508]
10000	8484 [5606, 11363]	4741 [2477, 7005]	38399 [35573, 41226]	92468 [67735, 117201]
15000	8505 [5628, 11382]	7142 [3724, 10561]	57651 [53452, 61849]	138652 [101576, 175729]
20000	**8525 [5649, 11401]**	9568 [4991, 14146]	76945 [71390, 82500]	184873 [135414, 234332]
T^α	0.37	1.0	1.0	1.0