# Contextual Dependent Click Bandit Algorithm for Web Recommendation

COCOON 2018

---

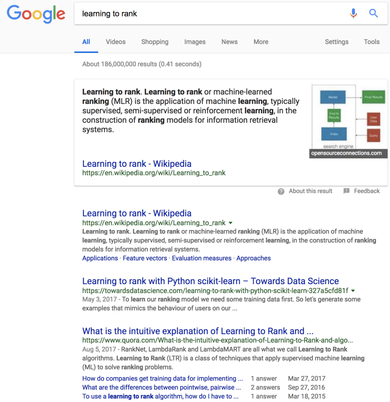
Weiwen Liu[1], Shuai Li[1], and Shengyu Zhang[1,2]

[1]The Chinese University of Hong Kong, Hong Kong
[2]Tencent, Shenzhen, China

# Web Page Recommendation

Given a search query, a `web page recommendation algorithm` recommends a list of related web pages.

- In the online scenario, the learning agent
  - receives user feedback and makes predictions according to previous user behaviors.
  - aims at maintaining a high Click-Through Rate (CTR) over a long period of time



# Exploitation vs Exploration

- Online recommendation involves a fundamental choice:

  **Exploitation:** exploiting the currently confirmed attractive yet suboptimal items

  **Exploration:** exploring uncertain but potentially interesting items which may produce large benefits later

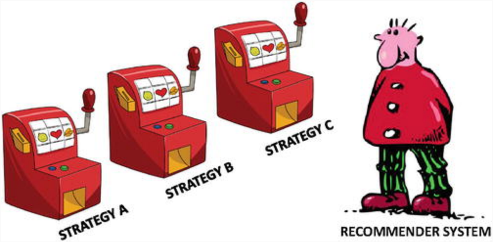
- The best long-term strategy may involve short-term sacrifices
- We need to gather enough information to make the best overall decisions

# Multi-armed Bandit

- At each round $t$, the learning agent selects one arm $i_t$, and receives a reward $R_t(i_t)$.
- Cumulative regret after $n$ rounds is

$$\text{regret} = n\mu^* - \mathbb{E}[\sum_{t=1}^{n} R_t(i_t)].$$

- The objective is to minimize the cumulative regret.



# Upper Confidence Bound (UCB) Algorithm

- Estimate the mean reward $\hat{Q}_t(a)$ of $Q_t(a)$, and the confidence radius $\hat{U}_t(a)$ for each arm $a$, such that $Q(a)$ is upper bounded by

$$Q(a) \le \hat{Q}_t(a) + \hat{U}_t(a)$$

  with high probability.
- The estimation depends on the number of times $N(a)$ that item $a$ has been selected.

  **Small** $N_t(a)$**:** large $\hat{U}_t(a)$ (estimation more uncertain)
  **Large** $N_t(a)$**:** small $\hat{U}_t(a)$ (estimation more accurate)
- UCB algorithm: Select an arm that maximizes the Upper Confidence Bound (UCB)

$$\boxed{a_t = \arg\max_a \underbrace{\hat{Q}_t(a)}_{\text{Exploitation}} + \underbrace{\hat{U}_t(a)}_{\text{Exploration}}.}$$

# Combinatorial Multi-armed Bandit

- Action is combinatorial $\Rightarrow$ generate a ranked list at each time step.
- Semi-bandit feedback $\Rightarrow$ only outcomes of the played arms are observed to the agent.
- Challenges
  - Exponential number of actions $\Rightarrow$ cannot be fully explored.

# Contextual Information

- A lot of available features:
  - user profiles, search keywords, hyperlinks, images, tags, comments, titles, etc.
- A solution to the cold-start problem in the recommender systems

# Dependent Click Model (DCM)

- models multiple clicks.
- Two sets of unknown parameters to be estimated:
    - **attraction probability:** item-dependent parameters
    - **termination probability:** position-dependent parameters

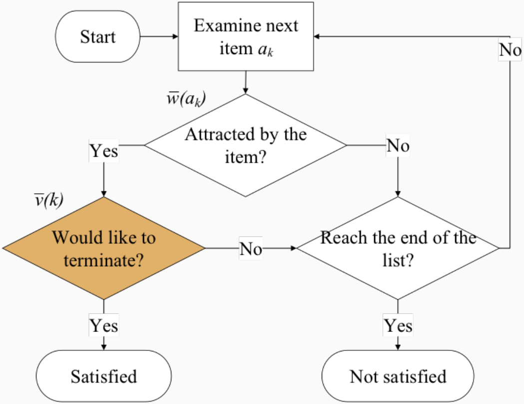



**Figure 1:** Dependent click model

## Problem Formulation

- A set of ground items $E = \{1, \dots, L\}$ and a feasible action set $\Pi_K(E)$.

# Problem Formulation

- A set of ground items $E = \{1, \ldots, L\}$ and a feasible action set $\Pi_K(E)$.
- At each time step $t$,
  - A set of contextual vectors is given $\{x_{i,t}\}_{i \in E}$ (e.g., user profiles/ keywords).

# Problem Formulation

- A set of ground items $E = \{1, \ldots, L\}$ and a feasible action set $\Pi_K(E)$.
- At each time step $t$,
  - A set of contextual vectors is given $\{x_{i,t}\}_{i \in E}$ (e.g., user profiles/ keywords).
  - The learning agent selects an ordered list $\mathbf{A}_t = (\mathbf{a}_1^t, \ldots, \mathbf{a}_K^t)$ of $K$ distinct items

# Problem Formulation

- A set of ground items $E = \{1, \ldots, L\}$ and a feasible action set $\Pi_K(E)$.
- At each time step $t$,
  - A set of contextual vectors is given $\{x_{i,t}\}_{i \in E}$ (e.g., user profiles/ keywords).
  - The learning agent selects an ordered list $\mathbf{A}_t = (\mathbf{a}_1^t, \ldots, \mathbf{a}_K^t)$ of $K$ distinct items
  - The user checks the list of items one by one from top to bottom. For each item $a$ at position $k$,
    - the user is **attracted** with probability $\bar{w}_t(a) \in [0, 1]$.
    - if attracted (the user clicks the item), the user will **feel satisfied and leave** with probability $\bar{v}_t(k)$.

# Problem Formulation

- A set of ground items $E = \{1, \dots, L\}$ and a feasible action set $\Pi_K(E)$.
- At each time step $t$,
  - A set of contextual vectors is given $\{x_{i,t}\}_{i \in E}$ (e.g., user profiles/ keywords).
  - The learning agent selects an ordered list $\mathbf{A}_t = (\mathbf{a}_1^t, \dots, \mathbf{a}_K^t)$ of $K$ distinct items
  - The user checks the list of items one by one from top to bottom. For each item $a$ at position $k$,
    - the user is **attracted** with probability $\bar{w}_t(a) \in [0, 1]$.
    - if attracted (the user clicks the item), the user will **feel satisfied and leave** with probability $\bar{v}_t(k)$.
  - The feedback is a sequence of $k$ binary click indicators $(\mathbf{w}_1', \dots, \mathbf{w}_K')$.

## Problem Formulation

- The reward of action $A$ is defined by

$$f(A, v, w) = 1 - \prod_{k=1}^{K} (1 - v(k) w(a_k)).$$

## Problem Formulation

- The reward of action $A$ is defined by

$$f(A, v, w) = 1 - \prod_{k=1}^{K} (1 - v(k)w(a_k)).$$

- The cumulative regret in $n$ rounds

$$\mathcal{R}(n) = \mathbb{E}\Big[\sum_{t=1}^{n} \big(f(A_t^*, \bar{v}_t, \bar{w}_t) - f(\mathbf{A}_t, \bar{v}_t, \bar{w}_t)\big)\Big],$$

where $A_t^*$ is the optimal list.

# Problem Formulation

- The reward is not revealed to the learning agent since the leaving position is ambiguous, e.g.,

$$\boxed{0100110000 \Rightarrow \begin{cases} \text{leave at the 6-th item with satisfaction,} \\ \text{finish the list and find nothing satisfied.} \end{cases}}$$

# Problem Formulation

- The attraction weight $\mathbf{w}_t(a)$ satisfies the `generalized linear model` (GLM)

$$\bar{w}_t(a) = \mathbb{E}[\mathbf{w}_t(a) | \mathcal{H}_t] = \mu(\theta_*^\top x_{t,a}),$$

where

  - $\{\mathcal{H}_t\}_{t=1}^n$ represents the history up to time $t$,
  - $\theta_*$ is a fixed but unknown vector $\theta_* \in \mathbb{R}^d$
  - The *inverse link function* $\mu$ is chosen s.t. $0 \leq \mu(\theta_*^\top x_{t,a}) \leq 1$. This GLM admits a wider range of nonlinear distributions such as Gaussian, binomial, Poisson, gamma distributions, etc. In particular, when the feedback is binary or count variables, the logistic or Poisson regression can be used.

[1] Sumeet Katariya et al. "DCM bandits: Learning to rank with multiple clicks". In: *International Conference on Machine Learning*. 2016, pp. 1215–1224.

# Problem Formulation

- The attraction weight $\mathbf{w}_t(a)$ satisfies the `generalized linear model` (GLM)

$$\bar{w}_t(a) = \mathbb{E}[\mathbf{w}_t(a)|\mathcal{H}_t] = \mu(\theta_*^\top x_{t,a}),$$

  where

  - $\{\mathcal{H}_t\}_{t=1}^n$ represents the history up to time $t$,
  - $\theta_*$ is a fixed but unknown vector $\theta_* \in \mathbb{R}^d$
  - The *inverse link function* $\mu$ is chosen s.t. $0 \leq \mu(\theta_*^\top x_{t,a}) \leq 1$. This GLM admits a wider range of nonlinear distributions such as Gaussian, binomial, Poisson, gamma distributions, etc. In particular, when the feedback is binary or count variables, the logistic or Poisson regression can be used.

- we adopt the same assumption as in [Katariya et al., 2016][1] that the order $\pi(\bar{v})$ of $\bar{v} = (\bar{v}(1), \ldots, \bar{v}(K))$ is known to the agent.

[1] Sumeet Katariya et al. "DCM bandits: Learning to rank with multiple clicks". In: *International Conference on Machine Learning*. 2016, pp. 1215–1224.

# Algorithm

- For round $t = 1, \ldots, n$
  - Obtain context $x_{t,a}$ for all $a \in E$.

# Algorithm

- For round $t = 1, \dots, n$
  - Obtain context $x_{t,a}$ for all $a \in E$.
  - By $\mathbf{w}_t(a) = \mu(\theta_*^\top x_{t,a})$, we obtain an estimate $\hat{\theta}_{t-1}$ of $\theta_*$, solved by Maximum Likelihood Estimation (MLE). With high probability,

$$\boxed{\mathbf{w}_t(a) \in \left[\mu(\hat{\theta}_{t-1}^\top x_{t,a}) - \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}, \mu(\hat{\theta}_{t-1}^\top x_{t,a}) + \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}\right]}$$

# Algorithm

- For round $t = 1, \dots, n$
  - Obtain context $x_{t,a}$ for all $a \in E$.
  - By $\mathbf{w}_t(a) = \mu(\theta_*^\top x_{t,a})$, we obtain an estimate $\hat{\theta}_{t-1}$ of $\theta_*$, solved by Maximum Likelihood Estimation (MLE). With high probability,

$$\boxed{\mathbf{w}_t(a) \in \left[\mu(\hat{\theta}_{t-1}^\top x_{t,a}) - \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}, \mu(\hat{\theta}_{t-1}^\top x_{t,a}) + \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}\right]}$$

  - Compute the Upper Confidence Bound (UCB) of each arm $a \in E$,

$$\mathbf{U}_t(a) = \min\{\mu(\hat{\theta}_{t-1}^\top x_{t,a}) + \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}, 1\}.$$

# Algorithm

- For round $t = 1, \ldots, n$
  - Obtain context $x_{t,a}$ for all $a \in E$.
  - By $\mathbf{w}_t(a) = \mu(\theta_*^\top x_{t,a})$, we obtain an estimate $\hat{\theta}_{t-1}$ of $\theta_*$, solved by Maximum Likelihood Estimation (MLE). With high probability,

$$\boxed{\mathbf{w}_t(a) \in \left[\mu(\hat{\theta}_{t-1}^\top x_{t,a}) - \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}, \mu(\hat{\theta}_{t-1}^\top x_{t,a}) + \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}\right]}$$

  - Compute the Upper Confidence Bound (UCB) of each arm $a \in E$,

$$\mathbf{U}_t(a) = \min\{\mu(\hat{\theta}_{t-1}^\top x_{t,a}) + \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}, 1\}.$$

  - Select action

$$\mathbf{A}_t \leftarrow \mathrm{argmax}_{A \in \Pi_K(E)} f(A, \bar{v}_t, \mathbf{U}_t).$$

# Algorithm

- For round $t = 1, \ldots, n$
  - Obtain context $x_{t,a}$ for all $a \in E$.
  - By $\mathbf{w}_t(a) = \mu(\theta_*^\top x_{t,a})$, we obtain an estimate $\hat{\theta}_{t-1}$ of $\theta_*$, solved by Maximum Likelihood Estimation (MLE). With high probability,

$$\boxed{\mathbf{w}_t(a) \in \left[\mu(\hat{\theta}_{t-1}^\top x_{t,a}) - \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}, \mu(\hat{\theta}_{t-1}^\top x_{t,a}) + \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}\right]}$$

  - Compute the Upper Confidence Bound (UCB) of each arm $a \in E$,

$$\mathbf{U}_t(a) = \min\{\mu(\hat{\theta}_{t-1}^\top x_{t,a}) + \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}, 1\}.$$

  - Select action

$$\mathbf{A}_t \leftarrow \operatorname{argmax}_{A \in \Pi_K(E)} f(A, \bar{v}_t, \mathbf{U}_t).$$

  - Play $\mathbf{A}_t$ and observe the last click position $\mathcal{C}_t$, and the click sequence $\mathbf{w}_t(\mathbf{a}_k^t)$, $k \in [\mathcal{C}_t]$.

# Algorithm

- For round $t = 1, \ldots, n$
  - Obtain context $x_{t,a}$ for all $a \in E$.
  - By $\mathbf{w}_t(a) = \mu(\theta_*^\top x_{t,a})$, we obtain an estimate $\hat{\theta}_{t-1}$ of $\theta_*$, solved by Maximum Likelihood Estimation (MLE). With high probability,

$$\boxed{\mathbf{w}_t(a) \in \left[\mu(\hat{\theta}_{t-1}^\top x_{t,a}) - \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}, \mu(\hat{\theta}_{t-1}^\top x_{t,a}) + \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}\right]}$$

  - Compute the Upper Confidence Bound (UCB) of each arm $a \in E$,

$$\mathbf{U}_t(a) = \min\{\mu(\hat{\theta}_{t-1}^\top x_{t,a}) + \rho(t-1)\|x_{t,a}\|_{\mathbf{V}_{t-1}^{-1}}, 1\}.$$

  - Select action

$$\mathbf{A}_t \leftarrow \mathrm{argmax}_{A \in \Pi_K(E)} f(A, \bar{v}_t, \mathbf{U}_t).$$

  - Play $\mathbf{A}_t$ and observe the last click position $\mathcal{C}_t$, and the click sequence $\mathbf{w}_t(\mathbf{a}_k^t)$, $k \in [\mathcal{C}_t]$.
  - Update $\mathbf{V}_t \leftarrow \mathbf{V}_{t-1} + \sum_{k=1}^{\mathcal{C}_t} x_{t,\mathbf{a}_k^t} x_{t,\mathbf{a}_k^t}^\top$

# Algorithm

**Lemma [Li et al., 2017][2]**

For any $\delta \in [1/n, 1)$, with probability at least $1-\delta$, for all $1 \le t \le n$, we have

$$\|\hat{\theta}_t - \theta_*\|_{\mathbf{V}_t} \le \frac{\sigma}{c_\mu}\sqrt{\frac{d}{2}\log(1 + t/(\lambda d)) + \log(1/\delta)}.$$

[2] Lihong Li, Yu Lu, and Dengyong Zhou. "Provable Optimal Algorithms for Generalized Linear Contextual Bandits". In: *Proceedings of The 34rd International Conference on Machine Learning* (2017).

# Result

**Theorem**

For $n \geq 1$, and the reward function $f(A, v, w) = 1 - \prod_{k=1}^{K}(1 - v(k)w(a_k))$, the cumulative regret $\mathcal{R}(n)$ has a bound of $\tilde{O}(d\sqrt{n})$

$$\mathcal{R}(n) \leq \frac{4dKp_v k_\mu \sigma}{c_\mu}\sqrt{nK \log\left(\frac{1 + n/(\lambda d)}{\delta}\right) \log(1 + Kn/(\lambda d))}.$$

- independent of the size of the ground item set $L$.
- improves the previous regret bound in [Filippi et al., 2010][3] by a $\sqrt{\log(n)}$ term.

[3] Sarah Filippi et al. "Parametric bandits: The generalized linear case". In: *Advances in Neural Information Processing Systems*. 2010, pp. 586–594.

# Proof (sketch)

- Reduce the problem to the cascading bandit problem

$$\begin{aligned}\mathbb{E}[\mathcal{R}_t|\mathcal{H}_t] &= f(A_t^*, \bar{v}_t, \bar{w}_t) - f(\mathbf{A}_t, \bar{v}_t, \bar{w}_t)\\ &\le \sum_{k=1}^{K} \bar{v}_t(k)\bar{w}_t(a_k^*) - \sum_{k=1}^{K} \bar{v}_t(k)\bar{w}_t(\mathbf{a}_k^t)\\ &\qquad\qquad\qquad\text{(by definition of } A_t^* \text{ and } f)\\ &= \sum_{i=1}^{K} (\bar{v}_t(i) - \bar{v}_t(i+1)) \sum_{k=1}^{i} (\bar{w}_t(a_k^*) - \bar{w}_t(\mathbf{a}_k^t)),\end{aligned}$$

where $\bar{v}_t(K+1) = 0$.

# Proof (sketch)

- Use the following Lemma to bound the cascade difference,

**Lemma**

Let $t \geq 1$ and $\mathbf{A}_t = (\mathbf{a}_1^t, ..., \mathbf{a}_i^t)$, $i \in [K]$, we have:

$$\sum_{k=1}^{i}(\mu(\theta_*^\top x_{t,a_k^*}) - \mu(\theta_*^\top x_{t,\mathbf{a}_k^t})) \leq 2\sum_{k=1}^{i}\rho(t-1)\|x_{t,\mathbf{a}_k^t}\|_{\mathbf{V}_{t-1}^{-1}},$$

where $\rho(t) = \frac{k_\mu \sigma}{c_\mu}\sqrt{\frac{d}{2}\log(1+t/(\lambda d)) + \log(1/\delta)}$.

# Proof (sketch)

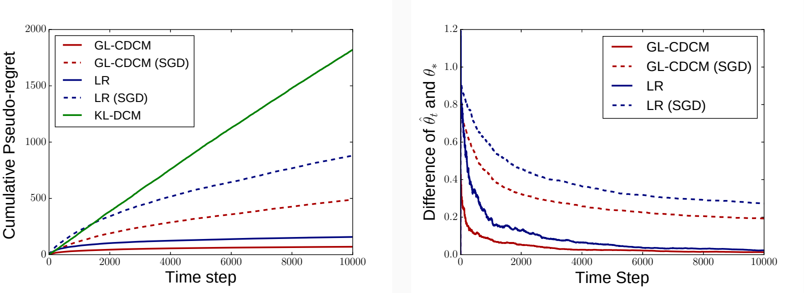
- The norm term can be further bounded by

**Lemma**

If $\lambda \geq K$, then

$$\sum_{s=1}^{t} \sum_{k=1}^{i} \left\| x_{s,\mathbf{a}_k^s} \right\|_{V_t^{-1}}^2 \leq 2d \log(1 + \frac{Kt}{\lambda d}).$$

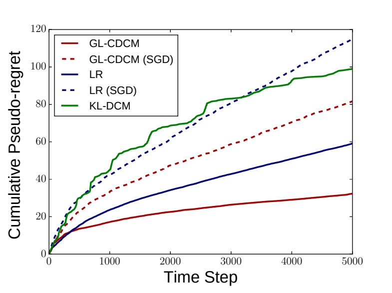
# Experiment: Synthetic Data



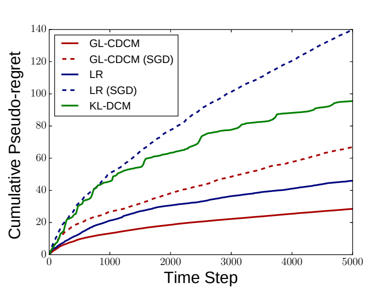

**(a)** Cumulative regrets. **(b)** Difference between $\theta_*$ and $\hat{\theta}_t$.

**Figure 2:** Synthetic data, select $K = 5$ items out of $L = 200$ items, dimension $d = 100$.

# Experiment: Real-world Data




**(a)** with only item features




**(b)** with both item and user features

**Figure 3:** Yandex Dataset: select $K = 10$ items out of $L = 100$ items, dimension $d = 200$.

# Conclusion

- Present a bandit algorithm for web page recommendation that automatically balances the exploration and exploitation.
- Incorporate contextual information in DCM bandit.
- Prove a regret bound of $\tilde{O}(d\sqrt{n})$ for the algorithm.