Structured Bandits and Applications

Exploiting Problem Structure for Better Decision-making under Uncertainty

Candidate: Sharan Vaswani

PhD Defence
University of British Columbia

11th December, 2018

Outline

1. Introduction
2. Influence Maximization
   - IM bandits under the IC model
   - Model-independent IM Bandits
3. Content-based Recommendation
4. Bootstrapping for Bandits
5. Summary

Outline

Motivating Example: Clinical Trials

Figure 1: Clinical trial to infer the "best" drug.

- Do not have complete information about the effectiveness or side-effects of the drugs.
- **Aim**: Infer the "best" drug by running a sequence of trials.

Abstraction to Multi-armed Bandits

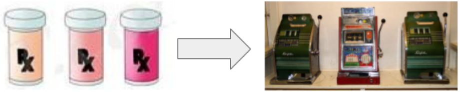


Figure 2: Mapping a clinical trial to the multi-armed bandit framework.

- Each drug choice is mapped to an arm and the drug's effectiveness is mapped to the arm's reward.
- Administering a drug is an action that is equivalent to pulling the corresponding arm.
- The trial goes on for T rounds.

Bandits 101

Algorithm: GENERIC BANDIT FRAMEWORK (K arms, T rounds)

Initialize the expected rewards according to some prior knowledge.
for $t = 1 \to T$ **do**
  SELECT: Use a bandit algorithm to decide which arm(s) to pull.
  OBSERVE: Pull the selected arm(s) and observe the reward and associated feedback.
  UPDATE: Update the estimated reward for the arm(s).

- How do we model the reward of an arm? What is the "best" arm?
- **Stochastic and stationarity assumptions**: The reward for each arm is sampled i.i.d from its underlying stationary distribution. The best arm is the one with the highest expected reward. $\implies$ UPDATE step involves computing the empirical mean of the past observations.
- **Multi-armed bandits assumption**: The reward for each arm is independent of the others.

Bandits 101

- What is the objective function?
- Minimize the expected cumulative regret $\mathbb{E}[R(T)]$. If a^* is the best action in hindsight and a_t is the action chosen at round t, then

$$\mathbb{E}[R(T)] = \sum_{t=1}^{T} [\mathbb{E}[\text{Reward for } a^*] - \mathbb{E}[\text{Reward for } a_t]]$$

- Minimizing $R(T)$ results in a exploration-exploitation trade-off:
  Exploration: Pull an arm to learn more about it.
  Exploitation: Pull the arm that has a higher empirical reward.
- **Common bandit algorithms**: Epoch-Greedy, Optimism under uncertainty, Thompson sampling.

Structured Bandits

- In problems with large number of arms, learning about each arm separately is inefficient.
- Can the rewards for arms depend on each other?
- **Contextual bandits**: Each arm j has a feature vector $\mathbf{x}_j$ and there exists an unknown vector θ^* such that

$$\mathbb{E}[\text{reward for arm } j] = m(\mathbf{x}_j, \theta^*)$$

- **Linear bandits**: The function m is linear $\implies m(\mathbf{x}, \theta) = \langle \mathbf{x}, \theta \rangle$.
- **Combinatorial bandits**: The chosen arms are required to satisfy a combinatorial constraint.

Outline

Influence maximization (IM)

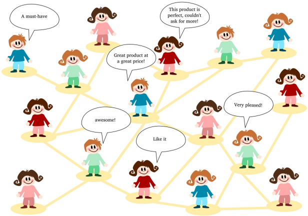


Figure 3: Information diffusion in a social network

- **Underlying principle**: Influence propagates through word-of-mouth in a social network.
- **Idea**: Give discounts to "influential" users who will trigger off word-of-mouth epidemics.
- **Aim**: Find the subset of users (seed or source set) that will result in the maximum number of people becoming aware of the product.

Problem formulation

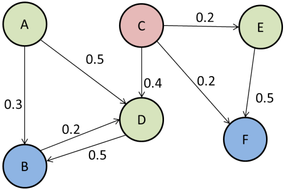


Figure 4: Modelling the social network for IM

- **Input**: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, Influence probabilities $p : \mathcal{E} \rightarrow [0, 1]$, Set of feasible seed sets $\mathcal{C}$, Stochastic diffusion model D.
- **Formal objective**: Find $\mathcal{S}^* \in \mathcal{C}$ that maximizes the expected number of influenced nodes $f(\cdot)$ under the diffusion model D.

$$\mathcal{S}^* \in \arg\max_{\mathcal{S} \in \mathcal{C}} f(\mathcal{S}, p)$$

Practical problems with IM

- × IM is not robust to the influence probabilities p.
- In practice, we do not have knowledge of p and it is difficult to obtain relevant data to learn from.
- × IM is not robust to the choice of the diffusion model.
- In practice, it is not clear how to choose from amongst different plausible diffusion models.
- × Number of parameters to be learned scales with the size of the network.
- In practice, this is not scalable to large real-world networks.
- ◇ **Idea 1**: Perform multiple attempts of IM and learn how to influence through repeated interaction in the bandit framework.
- ◇ **Idea 2**: Reparametrize the problem so that the diffusion process can be learned efficiently.

Mapping IM to Bandits

- **Round** $\leftrightarrow$ IM attempt
- **SELECT** $\leftrightarrow$ Choose a seed set $\mathcal{S}$.
- **OBSERVE** $\leftrightarrow$ Edge/Node semi-bandit feedback from the network.
- **UPDATE** $\leftrightarrow$ Sufficient statistics for estimating the diffusion.
- **Cumulative regret**: If $\mathcal{S}_t$ is the chosen seed set, $\mathbf{w}_t$ summarizes the diffusion in round t and the offline problem can be solved to an approximation factor of $\eta \in (0, 1)$,

$$R^\eta(T) = \sum_{t=1}^{T} \left[f(\mathcal{S}^*, \mathbf{w}_t) - \frac{1}{\eta} f(\mathcal{S}_t, \mathbf{w}_t) \right]$$

Outline

Parametrization

- Assume that the diffusion takes place according to the Independent Cascade (IC) model.
- Possible to obtain edge semi-bandit feedback
  $\implies$ can observe the state of all directed edges (u, v) for which the node u is activated in a diffusion.

◇ Linear parametrization for the influence probability of edge e:

$$p(e) \approx \langle x_e, \theta^* \rangle$$

$x_e \leftrightarrow$ Topological features for edge e
$\theta^* \leftrightarrow$ Unknown parameter mapping x_e to its corresponding $p(e)$.

✓ Casts the IM bandits problem into the linear bandit framework

✓ Number of parameters to be learned is independent of the network size.

Contributions

- ◇ Propose a scalable upper confidence bound-based algorithm.
- ◇ Identify a topology-dependent complexity metric C_* and use it to prove an upper bound on the regret.

Theorem

Assuming that the offline IM problem can be solved to within an η-approximation factor, then

$$\mathbb{E}[R(T)] \leq \widetilde{O}\left(d \cdot C_* \sqrt{m} \cdot \sqrt{T}/(\eta)\right)$$

- ✓ Near-optimal dependence on T, d.
- ✓ First topology-dependent upper bounds on the regret.
- ◇ Experimentally verify the tightness of the theoretical bounds.
- ◇ Show the advantage of linear parametrization on a real dataset.

Outline

Contributions - Parametrization

- Define pairwise reachability probabilities $q_{u,v} = f(\{u\}, v)$ and maximal pairwise reachability as $\widetilde{f}(\mathcal{S}, v, q) = \max_{u \in \mathcal{S}} \; q_{u,v}$.
- Formulate a surrogate objective: $\widetilde{f}(\mathcal{S}, q) = \sum_{v \in \mathcal{V}} \widetilde{f}(\mathcal{S}, v, q)$.

✓ **Model independence**: Depends only on the state after the diffusion has occurred and not on the nature of the diffusion process.

✓ **Optimization**: Function $\widetilde{f}(\mathcal{S}, q)$ is monotone and submodular in $\mathcal{S}$ regardless of the diffusion model.

✓ **Guaranteed approximation**: If the original objective $f(\mathcal{S})$ is monotone and submodular in $\mathcal{S}$, then the surrogate approximation factor $\rho \in [1/K, 1]$.

Contributions - Formulation

- ◇ Propose pairwise reachability feedback: Can observe whether each node $v \in \mathcal{V}$ was influenced by each source node $u \in \mathcal{S}$.
- ◇ Linear parametrization of pairwise reachability probabilities:

$$q_{u,v} \approx \langle x_v, \theta_u^* \rangle$$

$x_v \in\leftrightarrow$ Topological features for the node v.
$\theta_u^* \leftrightarrow$ Learnable parameter modelling the influence of node u.

- ✓ Casts model-independent IM bandits as n independent linear-bandit problems.
- ✓ Amount of feedback ($O(K \cdot n)$) is of the same order as the number of parameters ($O(d \cdot n)$) to be learned.

Contributions - Analysis

◇ Propose an upper confidence bound-based algorithm for which the regret can be bounded as follows:

Theorem

Assuming that the offline problem can be solved to within an η-approximation factor, then

$$\mathbb{E}[R(T)] \leq \widetilde{O}(d \cdot n^2 \cdot \sqrt{T}/(\eta\rho))$$

✓ Near-optimal dependence on T, d.

✓ First upper bounds for model-independent IM bandits.

Contributions - Experimental Results

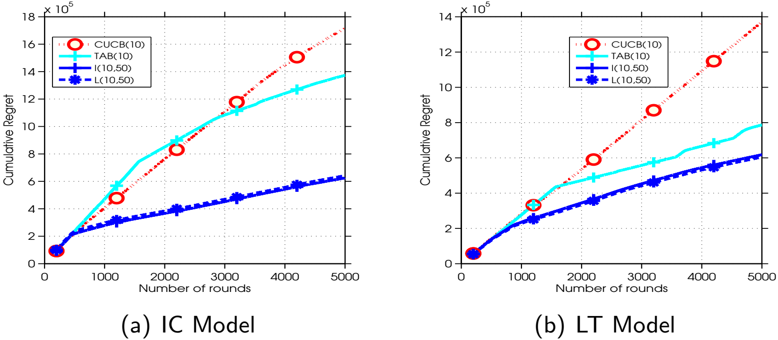


Figure 5: Comparing DILinUCB and CUCB on the Facebook subgraph with $K = 10$.

Outline

Problem formulation

- **Setup**: Newly established recommender system without any user meta-data or rating information. Have access to the content for the items to be recommended.
- **Common solution**: Model the recommendation problem as a contextual bandit for each user. Learn the users' preferences simultaneously while making recommendations.
- **Additional structure**: Users of the recommender system are part of an existing social network. E.g: Facebook, Quora.

◇ **Idea**: Exploit homophily between connected users using Laplacian regularization. Share information between users to learn their preferences faster.

Figure 6: Content-based recommendation with a user-user network.

- **SELECT** $\leftrightarrow$ Choose item j_t to recommend to the target user i_t.
- **OBSERVE** $\leftrightarrow$ Rating r_{i_t,j_t}.
- **UPDATE** $\leftrightarrow$ Preference vector estimate $\theta_{i,t}$ for user i at round t.
- Linear reward model: $\mathbb{E}[r_{i,j}] = \langle \theta_i^*, \mathbf{x}_j \rangle$
  $\mathbf{x} \leftrightarrow$ item content information; $\theta^* \leftrightarrow$ "true" preference vector.

$$\mathbb{E}[R(T)] = \sum_{t=1}^{T} \left[\max_{\mathbf{x}_j \in \mathcal{C}_t} \langle \theta_{i_t}^*, \mathbf{x}_j \rangle - \langle \theta_{i_t}^*, \mathbf{x}_{j_t,t} \rangle \right].$$

Contributions

Estimate users' preferences by solving:

$$\theta_t = \arg\min_{\theta} \left[\sum_{i=1}^{n} \sum_{k \in \mathcal{M}_{i,t}} (\langle \theta_i, \mathbf{x}_k \rangle - r_{i,k})^2 + \lambda \langle \theta, (L \otimes I_d) \theta \rangle \right],$$

× Previous approach requires $O(d^2 n^2)$ memory and computation.

◇ **Idea**: Interpret it as MAP estimation in a Gaussian Markov Random Field (GMRF) under the generative model:

$$r_{i,j} \sim \mathcal{N}(\langle \theta_i, \mathbf{x}_j \rangle, \sigma^2), \quad \theta \sim \mathcal{N}(0, (\lambda L \otimes I_d)^{-1}).$$

✓ Posterior $= \mathcal{N}(\theta_t, \Sigma_t^{-1})$; Σ_t is a block diagonal + sparse matrix
$\Longrightarrow$ Require $O\left(\kappa(nd^2 + md)\right)$ memory and computation.

Contributions

- ◇ Use the connection to GMRF and sampling by perturbation in order to design an efficient Thompson sampling algorithm.
- ◇ Prove an upper bound on the regret for Thompson sampling:

Theorem

With probability $1 - \delta$,

$$\mathbb{E}[R(T)] = \widetilde{O}\left(\frac{dn\sqrt{T}}{\sqrt{\lambda}}\sqrt{\log\left(\frac{3\,\mathrm{Tr}(L^{-1})}{n} + \frac{\mathrm{Tr}(L^{-1})T}{\lambda d n^2 \sigma^2}\right)}\right)$$

- ◇ Prove an analogous regret bound for Epoch-Greedy.
- ✓ Near-optimal dependence on T, dependence on the graph connectivity.
- ◇ Experimental comparison showing that using graph information leads to lower regret.

Outline

Motivation

- Complex non-linear functions are necessary for modelling structured data such as images or text. Need to resolve the exploration-exploitation trade-off for these complicated models.

× Can construct only approximate confidence sets in the non-linear setting
 $\Longrightarrow$ bad empirical performance of UCB-like algorithms.

× No closed form posteriors for non-linear models
 $\Longrightarrow$ need computationally-expensive approximate sampling techniques for Thompson sampling.

× Typically use ε-Greedy in practice, but it is sensitive to hyper-parameter tuning.

◇ **Idea**: Use bootstrapping to incorporate complex models in the bandit framework.

General bootstrapping algorithm for bandits

Algorithm: Bootstrapping for contextual bandits

Input: K arms, Model class m
Initialize history: $\forall j \in [K]$, $\mathcal{D}_j = \{\}$
for $t = 1$ **to** T **do**
  Observe context vector $\mathbf{x}_t$
  For all j, compute the bootstrap sample $\widetilde{\theta}_j$
  Select arm: $j_t = \arg\max_{j \in [K]} m(\mathbf{x}_t, \widetilde{\theta}_j)$
  Observe reward r_t
  Update history: $\mathcal{D}_{j_t} = \mathcal{D}_{j_t} \cup \{\mathbf{x}_t, r_t\}$

- **Computing a bootstrap sample**:
  - Formulate a bootstrapping log-likelihood function $\widetilde{\mathcal{L}}(\theta, Z)$ such that $\mathbb{E}_Z \left[\widetilde{\mathcal{L}}(\theta, Z)\right] = \mathcal{L}(\theta)$.
  - Given $Z = z$, generate a bootstrap sample: $\widetilde{\theta} \in \arg\max_\theta \widetilde{\mathcal{L}}(\theta, z)$.

Bootstrapping for Bandits

- ✓ Requires only point estimates instead of characterizing the entire posterior distribution.
- ✓ Performance is not sensitive to hyper-parameter tuning.

- × Popular non-parametric bootstrapping (NPB) procedure has no theoretical guarantee even in the simple Bernoulli or Gaussian bandit setting.
- × Uses ensembling and other heuristics to approximate the bootstrapping procedure that requires tuning additional hyper-parameters.

Contributions - Analysis

- Prove that the NPB procedure can be provably inefficient in the Bernoulli MAB setting.

Theorem

For any $\gamma \in (0,1)$ *and any* $T \geq \exp\left[\frac{2}{\gamma}\exp\left(\frac{80}{\gamma}\right)\right]$*, non-parametric bootstrapping can result in*

$$\mathbb{E}[R(T)] > \frac{T^{1-\gamma}}{32} = \Omega(T^{1-\gamma}).$$

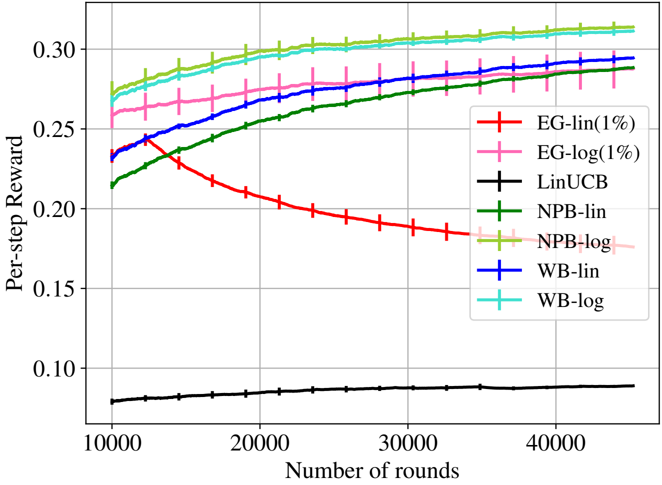
- Prove that NPB with appropriate forced exploration (done in practice) can result in sub-linear though sub-optimal $O(T^{2/3})$ regret.

Contributions - Algorithms

- ◇ Propose weighted bootstrapping (WB) that involves a random weighted transformation of the rewards.
- For Bernoulli rewards, WB involves
  - Generate exponential weights: $\forall i \in \mathcal{D}$, $w_i \sim Exp(1)$.
  - Transform labels: $y_i :\to w_i \cdot y_i$ and $(1 - y_i) :\to w_i \cdot (1 - y_i)$.

  $\implies$ Bootstrapping log-likelihood: $\widetilde{\mathcal{L}}(\theta) = \sum_{i \in \mathcal{D}_j} w_i \cdot \ell_i(\theta)$
- ✓ Easy and computationally efficient to implement.
- ✓ Results in near-optimal regret bounds in the Bernoulli and Gaussian MAB setting.

Contributions - Experimental results



(a) Adult

Outline

Summary

- **Chapter 2** [**V**KWGLS, ICML'17], [WKV**V**, NIPS'17]: Mapped the influence maximization problem to the linear bandit framework.
- **Chapter 3** [**V**LS, AISTATS'17]: Mapped content-based recommendation in the presence of a network to a graph-based contextual bandit framework.
- **Chapter 4** [**V**KWRSY, Under submission'18]: Investigated bootstrapping to model complex non-linear functions in the bandits framework.
- **Other work not included in this thesis**:
  - Fast and Faster Convergence of SGD for Over-Parametrized Models and an Accelerated Perceptron [**V**BS, Under submission'18]
  - Combining Bayesian Optimization and Lipschitz Optimization [A**V**S, Under submission'18]