Bayesian Contextual Multi-armed Bandits

Xiaoting Zhao
Joint Work with Peter I. Frazier
School of Operations Research and Information Engineering
Cornell University

October 22, 2012

Outline

Outline

Motivation: Recommender System

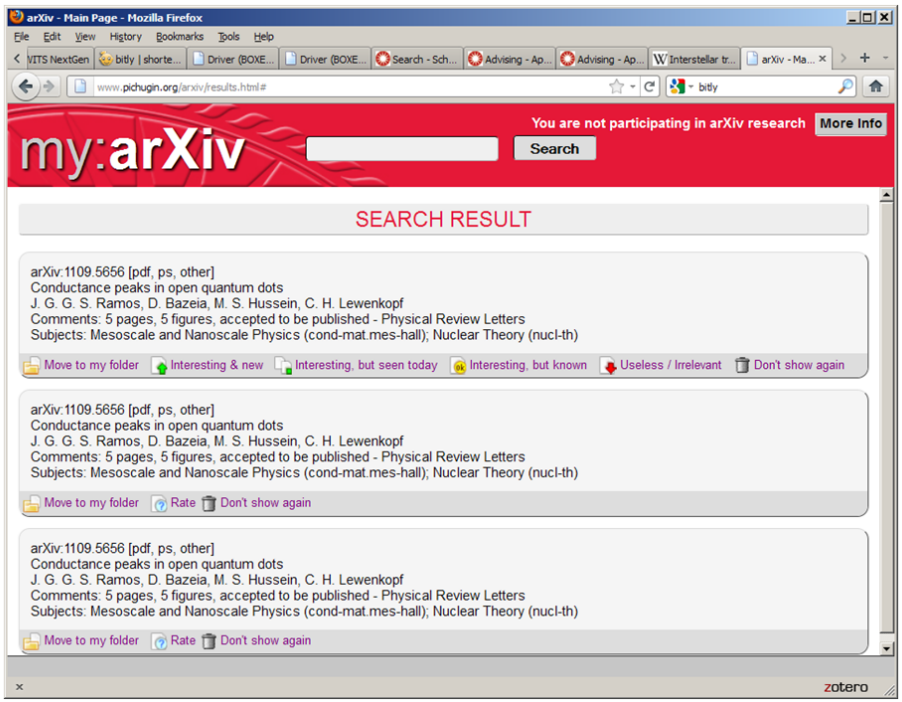
Learning through Exploration in arXiv

- Operate at a large scale:

  -$750,000$ *eprints of scientific articles*

  -1.5 *billions user accesses*

- Learn quickly and react fast
- Exploration vs. Exploitation



Goal: *Interactively combine content and use the observed feedback to improve future content choices.*

Motivation: Optimal Experimental Design

Oil-drilling

- Drill test wells to learn about the potential for oil or natural gas
- Each drilling test can cost millions of dollars
- Computational challenge: size of the search space
- **Goal**: *Choose the best price, temperature, pressure, and other parameters to produce the best results for a business simulator.*

Outline

Modeling: Contextual Multi-armed Bandits

- At each time step $n = 1, 2, ...$
  - Given context $Z_n \in \mathbb{R}^d$, iid from known P_Z
    - query keywords, user features, social media, medical records, etc.
  - Choose an action $X_n \in \{1, ..., K\}$
    - arms/items: articles, advertisement, movies, treatments, etc.
  - Latent item features θ_{X_n}:
    - item features, item relevance, etc.
  - Receive partial rewards, Y_n, with some noise $\epsilon_n \in \mathbb{R}$:
    - CTRs, generated revenue, ratings, relevance score, user engagement
    - Linear model: powerful and analytically tractable

$$Y_n = \theta_{X_n} \cdot Z_n + \epsilon_n \in \mathbb{R}$$

- Maximize the discounted total expected reward

$$\sup_{\pi \in \Pi} E^{\pi} \left[\sum_{n=0}^{\infty} \gamma^n Y_n \right]$$

Modeling: Contextual Multi-armed Bandits

- At each time step $n = 1, 2, ...$
  - Given context $Z_n \in \mathbb{R}^d$, iid from known P_Z
    - query keywords, user features, social media, medical records, etc.
  - Choose an action $X_n \in \{1, ..., K\}$
    - arms/items: articles, advertisement, movies, treatments, etc.
  - Latent item features θ_{X_n}:
    - item features, item relevance, etc.
  - Receive partial rewards, Y_n, with some noise $\epsilon_n \in \mathbb{R}$:
    - CTRs, generated revenue, ratings, relevance score, user engagement
    - Linear model: powerful and analytically tractable

$$Y_n = \theta_{X_n} \cdot Z_n + \epsilon_n \in \mathbb{R}$$

- Maximize the discounted total expected reward

$$\sup_{\pi \in \Pi} E^{\pi} \left[\sum_{n=0}^{\infty} \gamma^n Y_n \right]$$

Modeling: Contextual Multi-armed Bandits

- At each time step $n = 1, 2, ...$
  - Given context $Z_n \in \mathbb{R}^d$, iid from known P_Z
    - query keywords, user features, social media, medical records, etc.
  - Choose an action $X_n \in \{1, ..., K\}$
    - arms/items: articles, advertisement, movies, treatments, etc.
  - Latent item features θ_{X_n}:
    - item features, item relevance, etc.
  - Receive partial rewards, Y_n, with some noise $\epsilon_n \in \mathbb{R}$:
    - CTRs, generated revenue, ratings, relevance score, user engagement
    - Linear model: powerful and analytically tractable

$$Y_n = \theta_{X_n} \cdot Z_n + \epsilon_n \in \mathbb{R}$$

- Maximize the discounted total expected reward

$$\sup_{\pi \in \Pi} E^{\pi} \left[\sum_{n=0}^{\infty} \gamma^n Y_n \right]$$

Modeling: Contextual Multi-armed Bandits

- At each time step $n = 1, 2, ...$
  - Given context $Z_n \in \mathbb{R}^d$, iid from known P_Z
    - query keywords, user features, social media, medical records, etc.
  - Choose an action $X_n \in \{1, ..., K\}$
    - arms/items: articles, advertisement, movies, treatments, etc.
  - Latent item features θ_{X_n}:
    - item features, item relevance, etc.
  - Receive partial rewards, Y_n, with some noise $\epsilon_n \in \mathbb{R}$:
    - CTRs, generated revenue, ratings, relevance score, user engagement
    - Linear model: powerful and analytically tractable

$$Y_n = \theta_{X_n} \cdot Z_n + \epsilon_n \in \mathbb{R}$$

- Maximize the discounted total expected reward

$$\sup_{\pi \in \Pi} E^{\pi} \left[\sum_{n=0}^{\infty} \gamma^n Y_n \right]$$

Modeling: Contextual Multi-armed Bandits

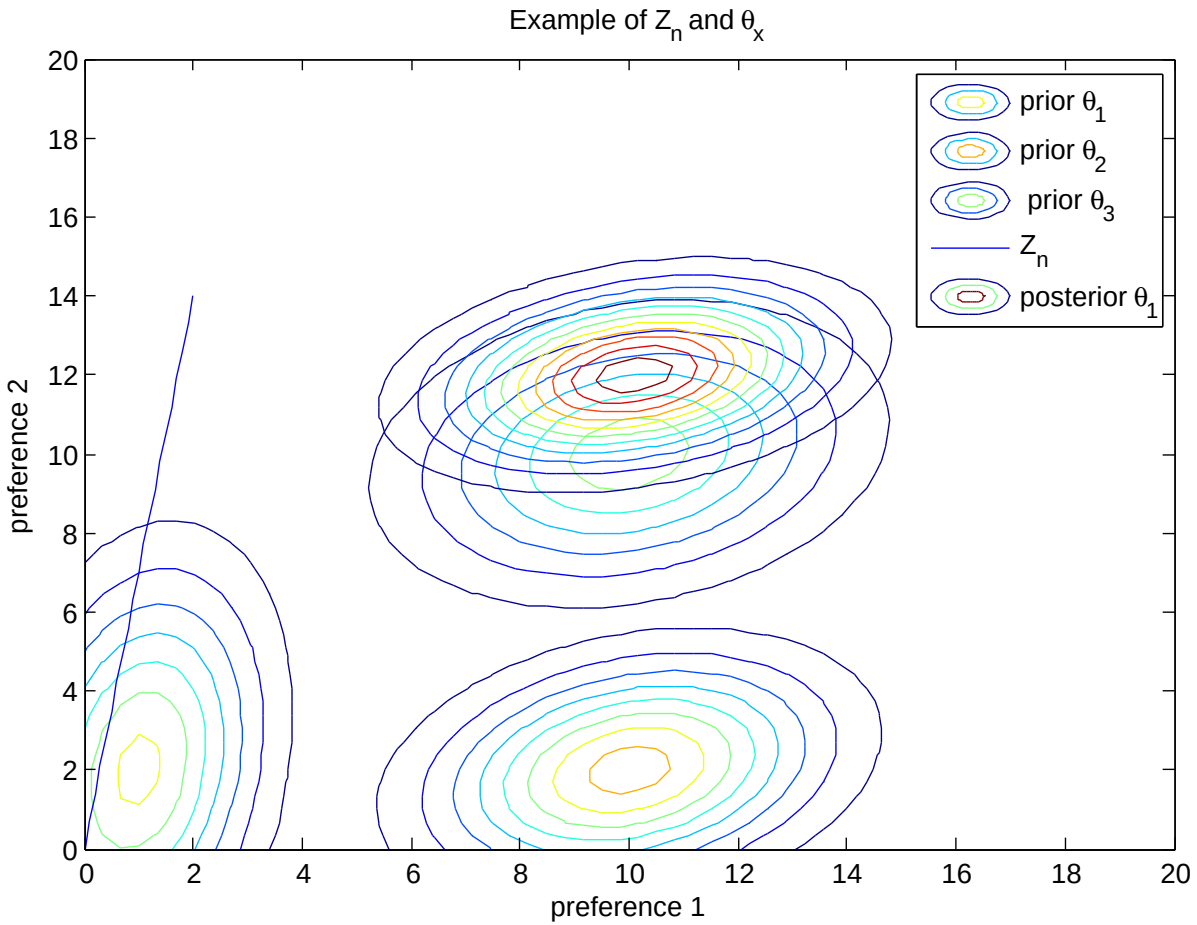
- At each time step $n = 1, 2, ...$
  - Given context $Z_n \in \mathbb{R}^d$, iid from known P_Z
    - query keywords, user features, social media, medical records, etc.
  - Choose an action $X_n \in \{1, ..., K\}$
    - arms/items: articles, advertisement, movies, treatments, etc.
  - Latent item features θ_{X_n}:
    - item features, item relevance, etc.
  - Receive partial rewards, Y_n, with some noise $\epsilon_n \in \mathbb{R}$:
    - CTRs, generated revenue, ratings, relevance score, user engagement
    - Linear model: powerful and analytically tractable

$$Y_n = \theta_{X_n} \cdot Z_n + \epsilon_n \in \mathbb{R}$$

- Maximize the discounted total expected reward

$$\sup_{\pi \in \Pi} E^{\pi} \left[\sum_{n=0}^{\infty} \gamma^n Y_n \right]$$

Intuitive Example



Literature Review

- Non-Bayesian Methods
  - The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits (Langford and Zhang 2007)
  - A Contextual-Bandit Approach to Personalized New Article Recommendation (Langford, et al., 2010)
  - An optimal high probability algorithm for the contextual bandit problem (Beygelzimer et al., 2010)
  - Efficient Optimal Learning for Contextual Bandits (Dudik et al., 2010)
  - PAC-Bayesian Analysis of Contextual Bandits (Seldin et al., 2011)
- Bayesian Approach
  - Explore/Exploit Schemes for Web Content Optimization (Agarwal, Chen, and Elango 2009)
  - Collaborative Topic Modeling for Recommending Scientific Articles (Wang and Blei 2011)

Modeling: Contextual Multi-armed Bandits

- Prior distribution on θ_x:

$$\theta_x \sim N(\mu_{x_0}, \Sigma_{x_0})$$

- Observation

$$Y_n = \theta_{X_n} \cdot Z_n + \epsilon_n \in \mathbb{R}$$

  with noise $\epsilon_n \sim N(0, \sigma^2)$

- Decision X_n at time n is a function of
  - Historical observation: $H_{n-1} = (X_m, Y_m, Z_m : 1 \leq m \leq n-1)$
  - New context: Z_n

Optimality Equation

Goal: find the optimal policy π^* s.t.

$$V^* = \sup_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{n=1}^{\infty} \gamma^n Z_n \cdot \theta_{X_n} \right],$$

where a policy π is a sequence of $\pi = (\pi_n : n = 1, 2, ...)$ with

$$\pi_n : \left(\{1, ..., K\} \times \mathbb{R} \times \mathbb{R}^d \right)^{n-1} \times \mathbb{R}^d \to \{1, ..., K\},$$

$$X_n = \pi_n(H_{n-1}, Z_n)$$

Outline

Simple Case: $Z_n = ones(d), d = 1$

Classical Multi-armed Bandits

- A row of slot machines to play
- Unknown expected payoff

$$V^* = \sup_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{n=1}^{\infty} \gamma^n R_n \right]$$

Gittins Index Policy (Gittins 1974)

- Assign index, $\nu_i(x_i)$, for each state of each item i

$$\nu_i(x_i) = \sup_{\tau > 0} \frac{< \sum_{n=0}^{\tau-1} \gamma^n R[X(n)] >_{X(0)=i}}{< \sum_{n=0}^{\tau-1} \gamma^n >}$$

- Index Policy: simple and provide the best tradeoff

$$i^* = \arg\max_i \nu_i(x_i)$$

Simple Case: $Z_n = ones(d), d = 1$

Classical Multi-armed Bandits

- A row of slot machines to play
- Unknown expected payoff

$$V^* = \sup_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{n=1}^{\infty} \gamma^n R_n \right]$$

Gittins Index Policy (Gittins 1974)

- Assign index, $\nu_i(x_i)$, for each state of each item i

$$\nu_i(x_i) = \sup_{\tau > 0} \frac{< \sum_{n=0}^{\tau-1} \gamma^n R[X(n)] >_{X(0)=i}}{< \sum_{n=0}^{\tau-1} \gamma^n >}$$

- Index Policy: simple and provide the best tradeoff

$$i^* = \arg\max_i \nu_i(x_i)$$

General Case: Contextual Multi-armed Bandits

$$V^* = \sup_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{n=1}^{\infty} \gamma^n Z_n \cdot \theta_{X_n} \right]$$

Challenges

- Random context, Z_n
  - Gittins index policy fails to generalize
- Curses of dimensionality
  - State space
  - Action space
- Partial feedback revealed

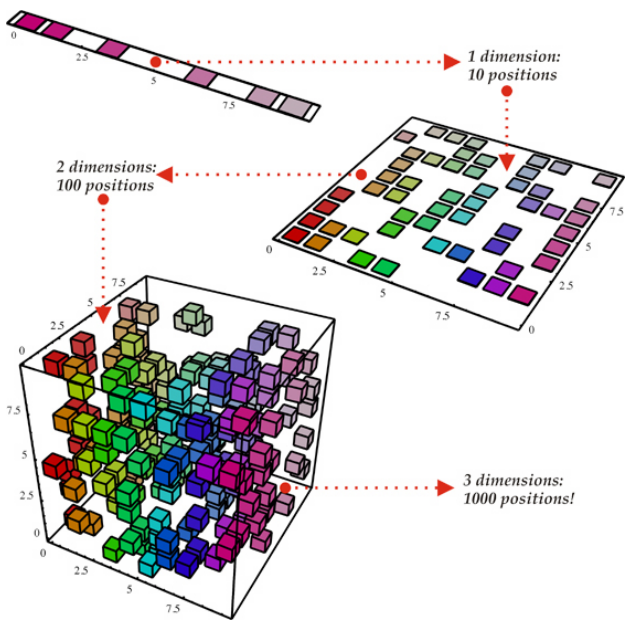
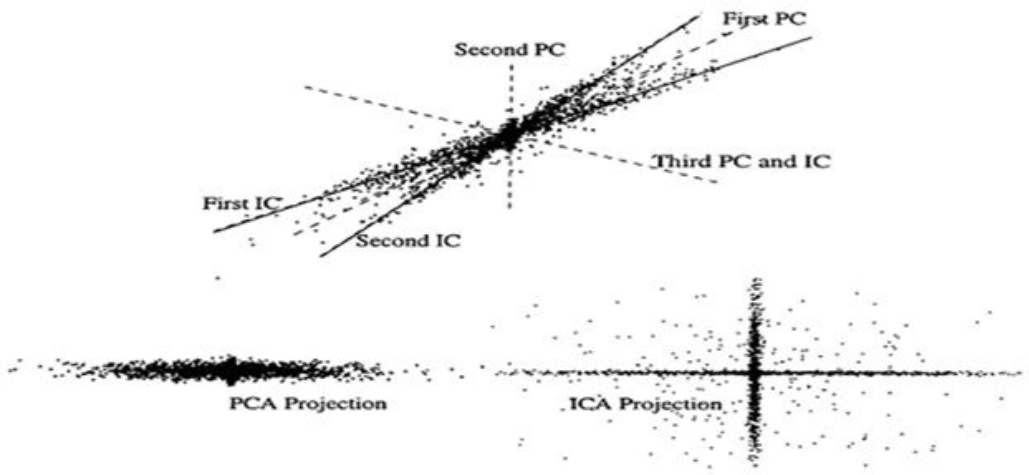


Figure: Graph inserted from Yoshua Bengio's homepage

Main Ideas of the Heuristic Policy

- Decomposition of the space into independent d sub-space



[1]

- Inspiration from Whittle's work on Restless Multi-armed Bandits
  - Relax the constraint on non-operating arms and average activity
  - Apply Lagrangian Relaxation

[1] Dimensionality Reduction Techniques for Face Recognition, Sharath et al., 2011

Upper Bound: Heuristic Policy

- Decompose to d sub-problems
  - $\phi_1, ...\phi_d$ be an orthonormal basis in $\mathbb{R}^d$, i.e., $\phi_j \in \mathbb{R}^d$

$$\phi_j \cdot \phi_{j'} = \begin{cases} 1 & \text{if } j = j', \\ 0 & \text{if } j \neq j'. \end{cases}$$

  - Arriving context Z_n as

$$Z_n = \sum_{i=1}^{d} Z_{ni}\phi_i \text{ with } Z_{ni} = \phi_i \cdot Z_n$$

- Find the optimal policy for each subproblem: *a 2-dim state-space DP*

$$\sup_{\pi' \in \Pi'} \mathbb{E}^{\pi'} \left[\sum_{n=1}^{\infty} \gamma^n Z_{ni}\phi_{in} \cdot \theta_{X_{ni}} \;\middle|\; \phi_{i'} \cdot \theta_j \quad \forall i' \neq i, \forall X_{ni} \in \{1, ..., K\} \right]$$

Proof of the Upper Bound

$$\sup_{\pi\in\Pi} \mathbb{E}_P^\pi \left[\sum_{n=1}^{\infty} \gamma^n Y_n\right] \leq \sup_{\pi\in\Pi'} \sum_{i=1}^{d} \mathbb{E}_Q^\pi \left[\sum_{n=1}^{\infty} \gamma^n Z_{ni}\phi_i \cdot \theta_{X_{ni}}\right]$$

$$\leq \sum_{i=1}^{d} \sup_{\pi\in\Pi'} \mathbb{E}_Q^\pi \left[\sum_{n=1}^{\infty} \gamma^n Z_{ni}\phi_i \cdot \theta_{X_{ni}}\right]$$

$$\leq \sum_{i=1}^{d} \mathbb{E}\left[\sup_{\pi\in\Pi'} \mathbb{E}_Q^\pi \left[\sum_{n=1}^{\infty} \gamma^n Z_{ni}\phi_{in} \cdot \theta_x \mid \phi_{i'} \cdot \theta_j \quad \forall i' \neq i, \forall x\right]\right].$$

Solving each sub-problem

$$V_i(\cdot) = \sup_{\pi' \in \Pi'} \mathbb{E}^{\pi'} \left[\sum_{n=1}^{\infty} \gamma^n Z_{ni} \phi_{in} \cdot \theta_{X_{ni}} \mid \phi_{i'} \cdot \theta_j \quad \forall i' \neq i, \forall j \right]$$

- **Gittins construction**
  - Non-operating arms stay static
  - But Z_n could evolve even for passive arms
- **Restless Multi-arm Bandits (Whittle 1988)**
  - Both active and passive arms could evolve
  - Whittle index policy: optimal under a relaxed constraint for average activity

Whittle Index Policy Relaxation

$$W_{ij}^{*} = \sup_{\pi''' \in \Pi'''} \mathbb{E}^{\pi'''} \left[\sum_{n=1}^{\infty} \sum_{\forall x} \gamma^{n} Z_{ni} \theta_{x} U_{nix} \mid \phi_{i'} \cdot \theta_{j} \quad \forall i' \neq i, x = 1, ..., K \right]$$

subject to:

$$U_{nix} \in \{0, 1\},$$

$$\mathbb{E}^{\pi} \left[\sum_{n=1}^{\infty} \gamma^{n} \left(\sum_{x} U_{nix} \right) \frac{Z_{ni}}{\mathbb{E}[Z_i]} \right] \leq \sum_{n=1}^{\infty} \gamma^{n} = \frac{1}{1-\gamma}$$

Lagrangian Relaxation

Relax the constraint with a Lagrange multiplier ν:

$$L\big(\nu, (\mu_0, \sigma_0^2, Z_{ni})\big) = \sum_x \sup_\pi \mathbb{E}^\pi \left[\sum_{n=1}^{\infty} \gamma^n Z_{ni} \left[\theta_x - \frac{\nu}{\mathbb{E}[Z_i]} \right] U_{nix} \right] + \frac{\nu}{1-\gamma}$$

subject to:

$$U_{nix} \in \{0, 1\}$$

Note: ν is a subsidy for resting an arm.

Weak Duality

- Dynamics of the state space $(\mu_{n+1,i,j}, \sigma^2_{n+1,i,j}, Z_{n+1,i})$

$$\begin{cases} (\mu_{nij}, \sigma^2_{nij}, Z_{n+1,i}) & \text{if } U_{nij} = 0, \\ \left(\mu_{nij} + \tilde{\sigma}(\sigma^2_{nij}) S_n, \ \left(\sigma^{-2}_{nij} + Z^2_{n,i}/\lambda^2\right)^{-1}, \ Z_{n+1,i}\right) & \text{if } U_{nij} = 1, \end{cases}$$

  where $S_n \sim \mathcal{N}(0,1)$ and $\tilde{\sigma}(s^2) = \sqrt{s^2 - (s^{-2} + (\lambda/Z_{ni})^{-2})^{-1}}$.

- Use Bisection algorithm to find ν^*

$$L(\mu_0, \sigma^2_0, Z_{ni})^* = \inf_{\nu \geq 0} L\big(\nu, (\mu_0, \sigma^2_0, Z_{ni})\big)$$

- Lagrangian relaxation optimality
  - Upper bound for the restless multi-armed bandit

$$W^* \leq L^*.$$

Weak Duality

- Dynamics of the state space $(\mu_{n+1,i,j}, \sigma^2_{n+1,i,j}, Z_{n+1,i})$

$$\begin{cases} (\mu_{nij}, \sigma^2_{nij}, Z_{n+1,i}) & \text{if } U_{nij} = 0, \\ \left(\mu_{nij} + \tilde{\sigma}(\sigma^2_{nij}) S_n,\ \left(\sigma^{-2}_{nij} + Z^2_{n,i}/\lambda^2\right)^{-1},\ Z_{n+1,i}\right) & \text{if } U_{nij} = 1, \end{cases}$$

  where $S_n \sim \mathcal{N}(0,1)$ and $\tilde{\sigma}(s^2) = \sqrt{s^2 - (s^{-2} + (\lambda/Z_{ni})^{-2})^{-1}}$.

- Use Bisection algorithm to find ν^*

$$L(\mu_0, \sigma^2_0, Z_{ni})^* = \inf_{\nu \geq 0} L\big(\nu, (\mu_0, \sigma^2_0, Z_{ni})\big)$$

- Lagrangian relaxation optimality
  - Upper bound for the restless multi-armed bandit

$$W^* \leq L^*.$$

Weak Duality

- Dynamics of the state space $(\mu_{n+1,i,j}, \sigma^2_{n+1,i,j}, Z_{n+1,i})$

$$\begin{cases} (\mu_{nij}, \sigma^2_{nij}, Z_{n+1,i}) & \text{if } U_{nij} = 0, \\ \left(\mu_{nij} + \tilde{\sigma}(\sigma^2_{nij}) S_n, \ \left(\sigma^{-2}_{nij} + Z^2_{n,i}/\lambda^2\right)^{-1}, \ Z_{n+1,i}\right) & \text{if } U_{nij} = 1, \end{cases}$$

  where $S_n \sim \mathcal{N}(0,1)$ and $\tilde{\sigma}(s^2) = \sqrt{s^2 - (s^{-2} + (\lambda/Z_{ni})^{-2})^{-1}}$.

- Use Bisection algorithm to find ν^*

$$L(\mu_0, \sigma^2_0, Z_{ni})^* = \inf_{\nu \geq 0} L\big(\nu, (\mu_0, \sigma^2_0, Z_{ni})\big)$$

- Lagrangian relaxation optimality
  - Upper bound for the restless multi-armed bandit

$$W^* \leq L^*.$$

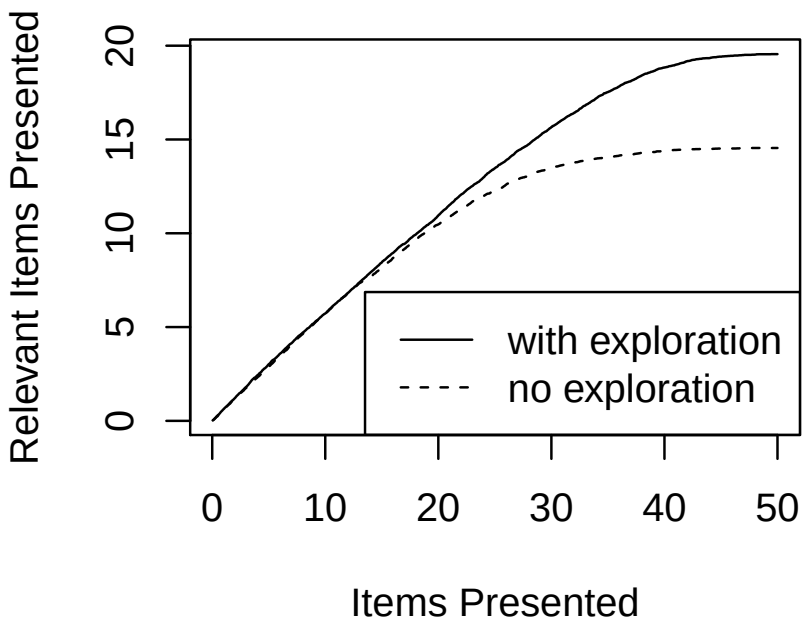
Numerical Results:
Bayesian exploration vs. exploitation on arXiv data

- *Exploration:* topic model w/ a Bayesian multi-armed bandit analysis

$$j^*_{\text{explore}} = \arg\max_j \mathbb{E}[\hat{Z}_i^T \theta_j] + \nu(\text{Var}[\hat{Z}_i^T \theta_j], \gamma_i, \sigma)$$

- *Exploitation:* recommend documents with the largest posterior mean

$$j^*_{\text{exploit}} = \arg\max_j \mathbb{E}[\hat{Z}_i^T \theta_j]$$



Outline

Conclusion: Contextual Multi-armed Bandits

- Many real life applications
  - A dynamic system with incomplete, large and noisy data
  - Optimal learning toward personalized matching
- Solving generalized context multi-arm Bandit
  - Exploration/learning vs. Exploitation/performance
  - Relax the problem to ease computation
  - Whittle index policy to derive an upper bound
- Future Direction
  - Implement the exploration algorithm in arXiv
  - Compare with Heuristic Policy for its closeness to optimality

References

- Agarwal, D. and Chen, B. (2009). Regression-based latent factor models.
- Dayanik, S., Powell, W., Yamazaki, K. (2008). Index policies for discounted bandit problems with availability constraints.
- Gittins, J.C. and Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments.
- Salakhutdinov, R. and Mnih, A. (2008). Probabilistic matrix factorization.
- Wang, C. and Blei, D. (2011). Collaborative topic modeling for recommmending scientific articles.
- Whittle, P. (1980). Multi-armed bandits and the Gittins index.
- Whittle, P. (1988). Restless bandits: Activity allocation in a changing world.

Thank You!