Incentivizing Bandit Exploration: Recommendations as Instruments

Daniel Ngo, Logan Stapleton, U of Minnesota
Vasilis Syrgkanis, Microsoft Research
Zhiwei Steven Wu Carnegie Mellon

Example: Discriminatory lending

- Loan officer of a bank can offer low interest loan to a business client
- Unobserved client type correlates with both the baseline revenue and the treatment choice (which is biased against small businesses)
- Bank wants to estimate the mean return (treatment effect) of these loans

$$\overbrace{y_t}^{\substack{\text{return}\\\text{(cost – revenue)}}} = \overbrace{\theta}^{\substack{\text{mean}\\\text{return}}} \underbrace{x_t}_{\substack{\text{treatment}\\\text{choice}}} + \overbrace{g(\underbrace{u_t}_{\substack{\text{business type}\\\text{(small / big)}}})}^{\substack{\text{baseline}\\\text{revenue}}} + \overbrace{\epsilon_t}^{\substack{\text{random}\\\text{error}}}$$

$$x_t = \underbrace{f}_{\substack{\text{loan officer's}\\\text{selection}\\\text{function}}}(\underbrace{u_t}_{\substack{\text{business}\\\text{type}}}, \overbrace{z_t}^{\substack{\text{bank's}\\\text{recommendation}}})$$

$$\boxed{x_t \in \{0, 1\}}$$

no loan (control) loan (treatment)

Model: Bank Loan
Repeated Interaction

Recommendations sway some salespeople to offer low-interest loan to small businesses.

- For each round $t \in T$:
- New business comes in & assigned loan officer
- Bank recommends control ($z_t = 0$) or treatment ($z_t = 1$) to loan officer based on the history of returns on past loans
- Loan officer decides to offer loan ($x_t = 1$) or not ($x_t = 0$)
- Profit y_t is observed by the bank
- Estimate treatment effect $\hat{\theta}$ based on observed data

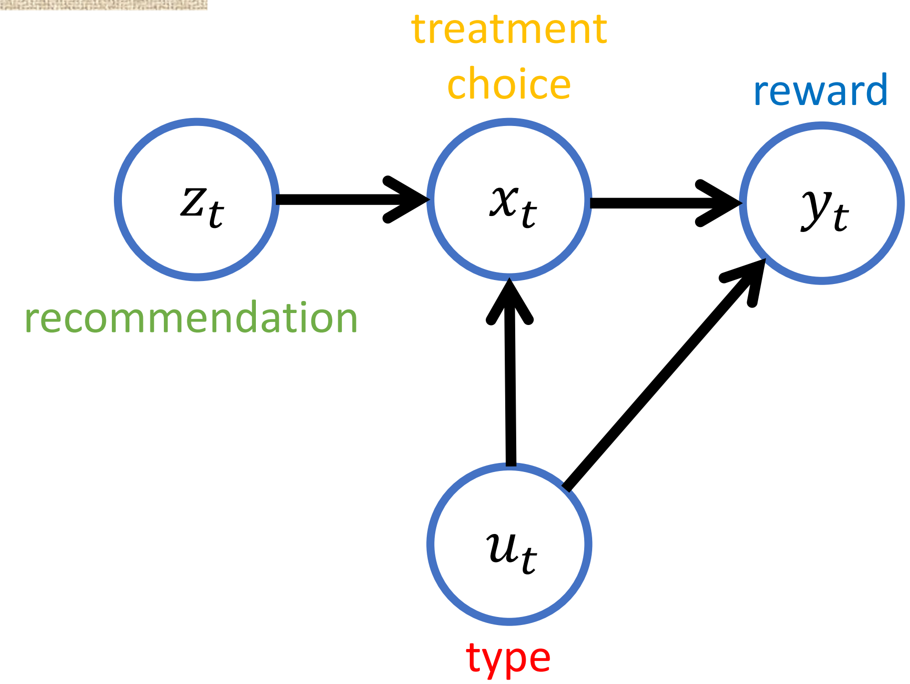
Bank's goal: Maximize net profit over all T rounds, estimate treatment effect

Loan officer's goal: Maximize their own expected profits

Recommendation as Instrument

Recommendations induce variability:

- Bank's recommendations incentivize loan officer to offer loan to small businesses
- Gather new information about loan returns



Instrument Variable Regression:

- Use n samples to estimate θ with $\hat{\theta}_n$:

$$\hat{\theta}_n = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}$$

Main Contributions

Multi-armed bandits algorithm with recommendations that incentivize exploration & act as instruments

Regret bound for our algorithm:

$$R(T) = \underbrace{c}_{\text{prior-dependent constant}} + \tilde{O}(\sqrt{\underbrace{T}_{\text{time horizon}}})$$

Treatment effect estimation error bound after running our algorithm:

$$\underbrace{|\hat{\theta}_n}_{\text{treatment effect estimation}} - \theta| = \tilde{O}\left(\frac{\sqrt{\overbrace{k}^{\text{\# of arms}}}}{\underbrace{n_{\min}}_{\text{least \# of times any arm is recommended}}}\right)$$

Fully incentive-compatible (benchmark)	Our Algorithm
Constant $c > T$ Linear regret	Constant $c < T$ Sublinear regret