Stay With Me: Lifetime Maximization Through Heteroscedastic Linear Bandits With Reneging

Ping-Chun Hsieh[1], Xi Liu[1], Anirban Bhattacharya[2], and P. R. Kumar[1]

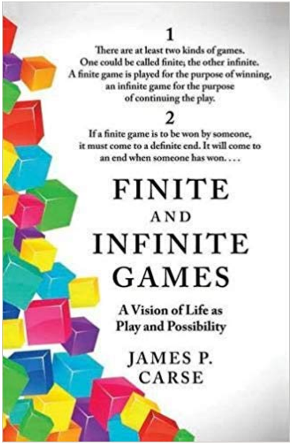
[1]Department of ECE
Texas A&M University

[2] Department of Statistics
Texas A&M University

ICML 2019

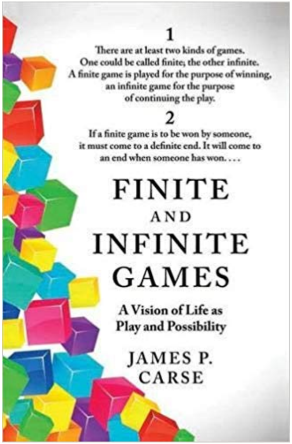
Poster @ Pacific Ballroom # 124

Lifetime Maximization: Continuing The Play



- A finite game is played for the purpose of winning.
- An infinite game is for the purpose of continuing the play.

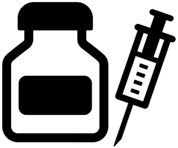
Lifetime Maximization: Continuing The Play



- A finite game is played for the purpose of winning.
- An infinite game is for the purpose of continuing the play.

Lifetime maximization

Why Lifetime Maximization?



Medical treatments

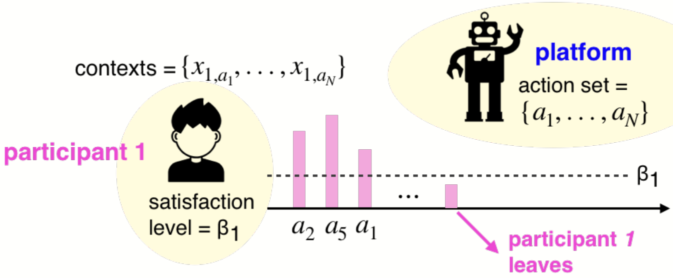
Portfolio selection

Cloud services

Salient features of these applications:

1. Each participant has a satisfaction level.
2. A participant drops if the outcomes are not satisfactory.
3. The outcomes depend heavily on the contextual information of the participant.

Model: Linear Bandits With Reneging



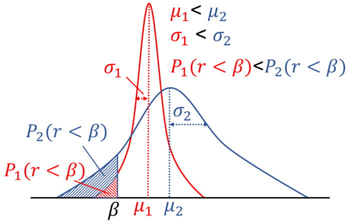
1. $\{x_{t,a}\}_{a \in A}$ are pairwise participant-action contexts (observed by the platform when participant t arrives).
2. Outcome $r_{t,a}$ is conditionally independent given the context and has mean $\theta_*^T x_{t,a}$.
3. Participant t keeps interacting with the platform as long as $r_{t,a} \geq \beta_t$. Otherwise, the participant drops.

Heteroscedastic Outcomes

- Heteroscedasticity: Outcome variations can be wildly different across different participants and actions

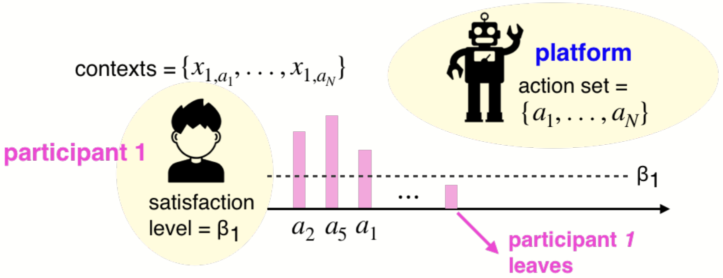
Heteroscedastic Outcomes

- Heteroscedasticity: Outcome variations can be wildly different across different participants and actions
- Example:
  - Two actions, 1 (red) and 2 (blue)
  - Participant satisfaction level $= \beta$



- Heteroscedasticity is widely studied in econometrics, and is usually captured through regression on variance.

Model: Heteroscedastic Bandits With Reneging



1. $\{x_{t,a}\}_{a \in A}$ are pairwise participant-action contexts (observed by the platform when participant t arrives)
2. Outcome $r_{t,a}$ is conditionally independent given the context and satisfies that $r_{t,a} \sim \mathcal{N}(\theta_*^\top x_{t,a}, f(\phi_*^\top x_{t,a}))$.
3. Participant t keeps interacting with the platform if $r_{t,a} \geq \beta_t$. Otherwise, the participant drops.

Oracle Policy and Regret

- Oracle policy π^{oracle} already knows θ_* and ϕ_*.
- For each participant t, π^{oracle} keeps choosing the action that minimizes reneging probability $\mathbb{P}\{r_{t,a} < \beta_t | x_{t,a}\}$
  - Hence, π^{oracle} is a fixed policy

Oracle Policy and Regret

- Oracle policy π^{oracle} already knows θ_* and ϕ_*.
- For each participant t, π^{oracle} keeps choosing the action that minimizes reneging probability $\mathbb{P}\{r_{t,a} < \beta_t | x_{t,a}\}$
  - Hence, π^{oracle} is a fixed policy
- For T participants, define

$$\begin{aligned}\text{Regret}^{\pi}(T) = &(\text{the total expected lifetime under } \pi^{\text{oracle}}) \\ &- (\text{the total expected lifetime under } \pi)\end{aligned}$$

Proposed Algorithm: HR-UCB

- When participant *t* arrives, obtain estimators $\widehat{\theta}, \widehat{\phi}$ with confidence intervals C_θ, C_ϕ based on past observations.

Proposed Algorithm: HR-UCB

- When participant t arrives, obtain estimators $\widehat{\theta}, \widehat{\phi}$ with confidence intervals C_θ, C_ϕ based on past observations.
- For each action a, construct a UCB index as

$$Q_t^{\text{HR}}(x_{t,a}) = \underbrace{\left[\Phi\left(\frac{\beta_t - \widehat{\theta}^\top x_{t,a}}{\sqrt{f(\widehat{\phi}^\top x_{t,a})}}\right)\right]^{-1}}_{\text{estimated expected lifetime}} + \underbrace{\Delta(C_\theta, C_\phi, x_{t,a})}_{\text{confidence interval for lifetime}} \quad (1)$$

Proposed Algorithm: HR-UCB

- When participant t arrives, obtain estimators $\widehat{\theta}, \widehat{\phi}$ with confidence intervals C_θ, C_ϕ based on past observations.
- For each action a, construct a UCB index as

$$Q_t^{\mathrm{HR}}(x_{t,a}) = \underbrace{\left[\Phi\left(\frac{\beta_t - \widehat{\theta}^\top x_{t,a}}{\sqrt{f(\widehat{\phi}^\top x_{t,a})}}\right)\right]^{-1}}_{\text{estimated expected lifetime}} + \underbrace{\Delta(C_\theta, C_\phi, x_{t,a})}_{\text{confidence interval for lifetime}} \tag{1}$$

- Apply the action $\arg\max_a Q_t^{\mathrm{HR}}(x_{t,a})$.

Proposed Algorithm: HR-UCB

- When participant t arrives, obtain estimators $\widehat{\theta}, \widehat{\phi}$ with confidence intervals C_θ, C_ϕ based on past observations.
- For each action a, construct a UCB index as

$$Q_t^{\mathrm{HR}}(x_{t,a}) = \underbrace{\left[\Phi\left(\frac{\beta_t - \widehat{\theta}^\top x_{t,a}}{\sqrt{f(\widehat{\phi}^\top x_{t,a})}}\right)\right]^{-1}}_{\text{estimated expected lifetime}} + \underbrace{\Delta(C_\theta, C_\phi, x_{t,a})}_{\text{confidence interval for lifetime}} \quad (1)$$

- Apply the action $\arg\max_a Q_t^{\mathrm{HR}}(x_{t,a})$.

Main technical challenges

1. Design estimators $\widehat{\theta}, \widehat{\phi}$ under heteroscedasticity
2. Derive the confidence intervals C_θ, C_ϕ for $\widehat{\theta}, \widehat{\phi}$
3. Convert the C_θ, C_ϕ into the confidence interval of lifetime

Estimators of θ_* and ϕ_* (Challenge 1)

- Generalized least square estimator (Wooldridge, 2015): With any n outcome observations,

$$\widehat{\theta}_n = \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}_n^\top r,$$
$$\widehat{\phi}_n = \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}_n^\top f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}).$$

  - $\boldsymbol{X}_n$ is the matrix of n applied contexts
  - r is the vector of n observed outcomes
  - $\widehat{\varepsilon}(x_{t,a}) = r_{t,a} - \widehat{\theta}_n^\top x_{t,a}$ is the estimated residual with respect to $\widehat{\theta}_n$

Estimators of θ_* and ϕ_* (Challenge 1)

- Generalized least square estimator (Wooldridge, 2015): With any n outcome observations,

$$\widehat{\theta}_n = \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}_n^\top r,$$
$$\widehat{\phi}_n = \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}_n^\top f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}).$$

  - $\boldsymbol{X}_n$ is the matrix of n applied contexts
  - r is the vector of n observed outcomes
  - $\widehat{\varepsilon}(x_{t,a}) = r_{t,a} - \widehat{\theta}_n^\top x_{t,a}$ is the estimated residual with respect to $\widehat{\theta}_n$

- Nice property (Abbasi-Yadkori et al., 2011): Let $\boldsymbol{V}_n = \boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}$. For any $\delta > 0$, with probability at least $1 - \delta$, for all $n \in \mathbb{N}$,

$$||\widehat{\theta}_n - \theta_*||_{\boldsymbol{V}_n} \leq C_\theta(\delta, n) = \mathcal{O}\left(\log(\frac{1}{\delta}) + \log n\right).$$

Main Technical Contributions (Challenges 2 & 3)

Theorem

For any $\delta > 0$, with probability at least $1 - 2\delta$, we have

$$||\widehat{\phi}_n - \phi_*||_{\boldsymbol{V}_n} \leq C_\phi(\delta, n) = \mathcal{O}\Big(\log(\frac{1}{\delta}) + \log n\Big), \quad \forall n \in \mathbb{N}. \tag{2}$$

- The proof is more involved since $\widehat{\phi}_n$ depends on the residual $\widehat{\varepsilon}$

Main Technical Contributions (Challenges 2 & 3)

Theorem

For any $\delta > 0$, with probability at least $1 - 2\delta$, we have

$$||\widehat{\phi}_n - \phi_*||_{\mathbf{V}_n} \leq C_\phi(\delta, n) = \mathcal{O}\Big(\log(\frac{1}{\delta}) + \log n\Big), \quad \forall n \in \mathbb{N}. \tag{2}$$

- The proof is more involved since $\widehat{\phi}_n$ depends on the residual $\widehat{\varepsilon}$

Theorem

$\Delta(C_\theta(n,\delta), C_\phi(n,\delta), x) := \big(k_1 C_\theta(n,\delta) + k_2 C_\phi(n,\delta)\big) \cdot ||x||_{\mathbf{V}_n^{-1}}$ is a confidence interval with respect to lifetime, where k_1, k_2 are constants independent of past history and x.

Main Technical Contributions (Challenges 2 & 3)

Theorem

For any $\delta > 0$, with probability at least $1 - 2\delta$, we have

$$||\widehat{\phi}_n - \phi_*||_{\boldsymbol{V}_n} \leq C_\phi(\delta, n) = \mathcal{O}\Big(\log(\frac{1}{\delta}) + \log n\Big), \quad \forall n \in \mathbb{N}. \qquad (2)$$

- The proof is more involved since $\widehat{\phi}_n$ depends on the residual $\widehat{\varepsilon}$

Theorem

$\Delta(C_\theta(n,\delta), C_\phi(n,\delta), x) := \big(k_1 C_\theta(n,\delta) + k_2 C_\phi(n,\delta)\big) \cdot ||x||_{\boldsymbol{V}_n^{-1}}$ is a confidence interval with respect to lifetime, where k_1, k_2 are constants independent of past history and x.

Theorem

Under the HR-UCB policy, $\text{Regret}(T) = \mathcal{O}\Big(\sqrt{T(\log T)^3}\Big)$.