Off-Policy Confidence Sequences

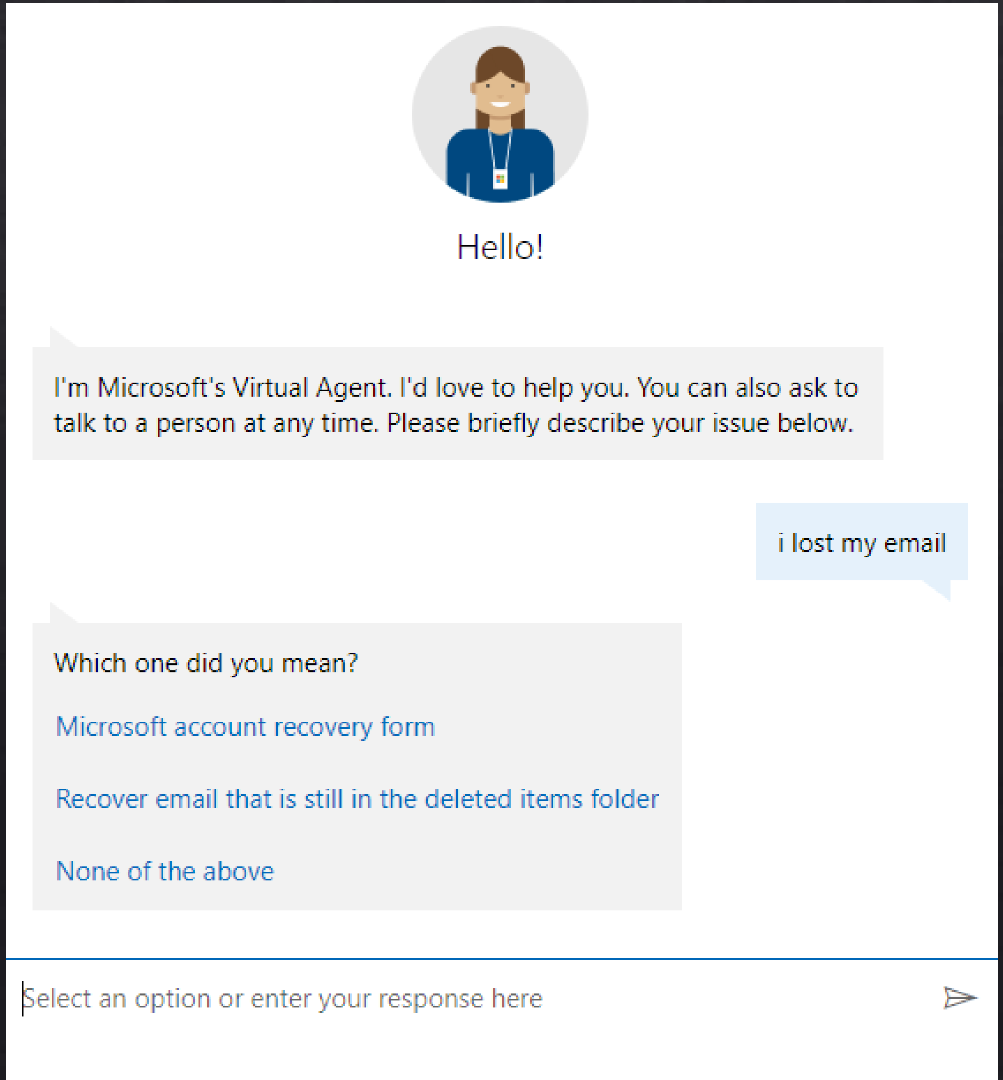
Nikos Karampatziakis (Microsoft)

Paul Mineiro (Microsoft)

Aaditya Ramdas (CMU)

Setup

- We have
  - a system acting according to policy h.
  - **contextual bandit** data $(x_i, a_i, r_i, h(a_i|x_i))$
- We want to know if a new policy π is better than h



Off-Policy Evaluation

- Have: $x \sim D$, $a \sim h(x)$, $r \sim R(x, a)$. Want to estimate $V(\pi) \coloneqq E_{\substack{x \sim D \\ a \sim \pi(x) \\ r \sim R(x,a)}}[r]$
- IPS estimator: $\hat{V}^{IPS}(\pi) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(a_i|x_i)}{h(a_i|x_i)} r_i \coloneqq \frac{1}{N} \sum_{i=1}^{N} w_i r_i$
- Can have large variance

Confidence Intervals and Sequences

- Given a **fixed** dataset S of iid contextual bandit data of size N
- A $(1-\alpha)$ **Confidence Interval** is a set $C = C(S, N)$ such that
$$\Pr(V(\pi) \notin C) \leq \alpha$$
- Confidence Intervals **lack adaptivity** over time:
  - Need to choose N upfront.
  - Can't early stop experiment.
  - Data cannot be reused without correction.
- **Confidence Sequences** are sequences of sets C_t such that
$$\Pr(\exists \boldsymbol{t}\colon V(\pi) \notin C_t) \leq \alpha$$

Off-Policy Confidence Sequences

- Our Off-Policy Confidence Sequences are constructed by observing that the quantity

$$K_t(v) = \prod_{i=1}^{t} \left(1 + \lambda_{1,i}(w_i - 1) + \lambda_{2,i}(w_i r_i - v)\right)$$

is a **non-negative martingale** iff $v = V(\pi)$, under the following constraints:

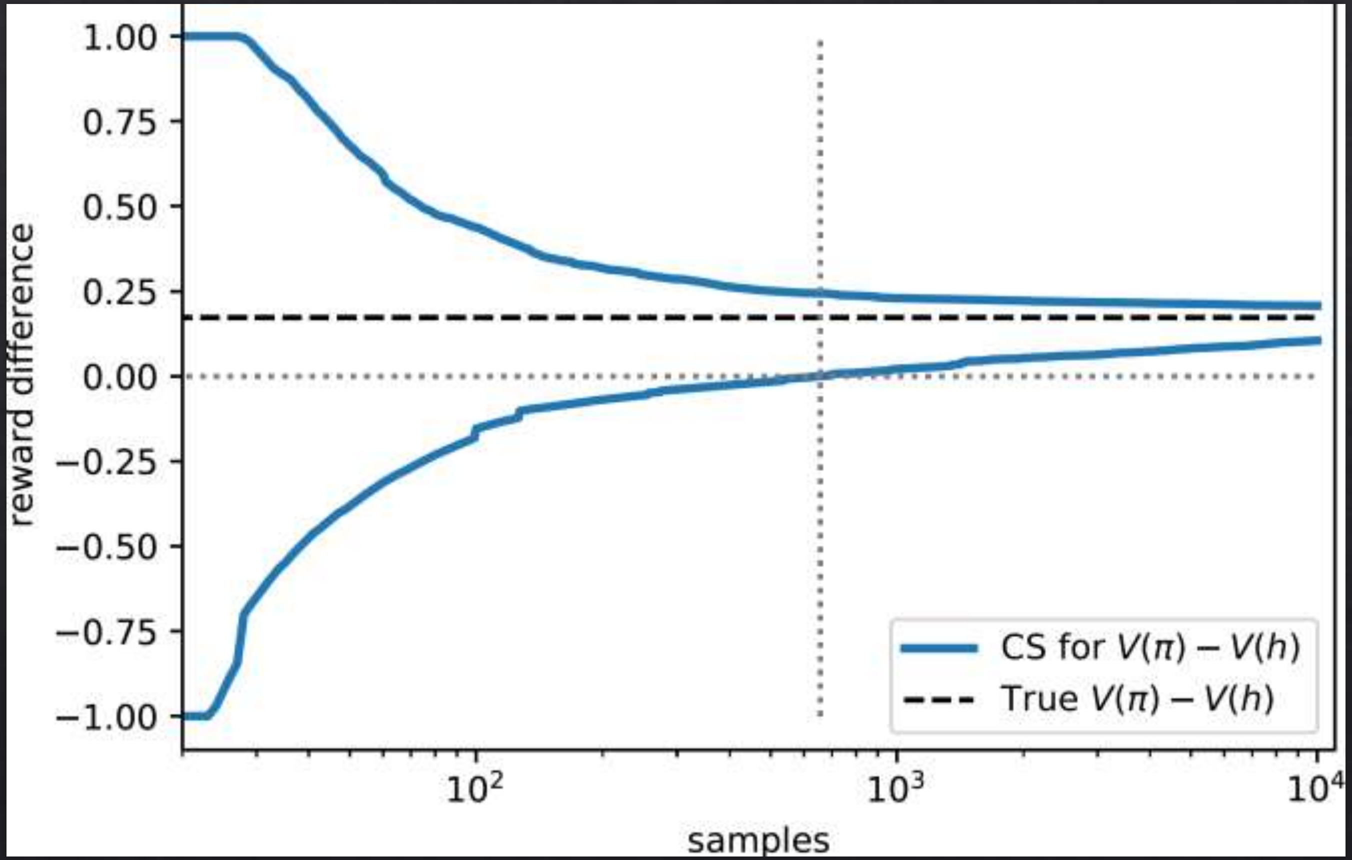
  - $\lambda_{1,i}, \lambda_{2,i}$ are numbers **chosen online**
  - Each factor is non-negative

- **Theorem**: The set $\left\{v: K_t(v) \leq \frac{1}{\alpha}\right\}$ is a $1 - \alpha$ CS
- Good ways of choosing $\lambda_{1,i}, \lambda_{2,i}$ will lead to a small set

Our techniques

- View $K_t(v)$ as the **wealth** of a skeptic betting against the hypothesis $V(\pi) = v$
- $\lambda_{1,i}, \lambda_{2,i}$ are related to size and direction (e.g. $v > V(\pi)$) of skeptic's **bets**.
- Choose bets to maximize wealth.
- To do this efficiently for all v we use:
  - Bets derived by optimizing wealth lower bound
  - A "hedging" technique and common bets for all v
  - A tight relaxation of $\left\{v: K_t(v) \leq \frac{1}{\alpha}\right\}$
- We also show how to incorporate a **reward predictor** to reduce variance

Example Experiment



More experiments in the paper assessing coverage, width, timings, design decisions (ablations) and the effect of a reward predictor.

Conclusions

- Confidence Sequences let you **monitor** your experiment
- We derived Confidence Sequences for off-policy evaluation via a **betting** view
  - Simple to implement https://github.com/n17s/mope
  - Computationally efficient
  - Empirically tight