For $t = 1, \ldots, T$:

1. The world produces some context $x \in X$
2. The learner chooses an action $a \in A$
3. The world react with reward $r_a \in [0, 1]$

**Goal:** Learn a good policy for choosing actions given context.

What does learning mean?
Reminder: Contextual Bandit Setting

For $t = 1, \ldots, T$:

1. The world produces some context $x \in X$
2. The learner chooses an action $a \in A$
3. The world reacts with reward $r_a \in [0, 1]$

Goal: Learn a good policy for choosing actions given context.

What does learning mean? Efficiently competing with some large reference class of policies $\Pi = \{\pi : X \rightarrow A\}$:

$$\text{Regret} = \max_{\pi \in \Pi} \text{average}_t (r_{\pi(x)} - r_a)$$
A Rejection Sampling approach

Rejection Sampler (policy \( \pi \), events \((\vec{x}, a, r, p)^T\))

Let \( h = \emptyset \) a history, \( R = 0 \)

For each event \((\vec{x}, a, r, p)\)

1. If \( \pi(h, \vec{x}) = a \)
2. then with probability \( \frac{p_{\min}}{p} \)
   1. \( h \leftarrow h \cup (\vec{x}, a, r) \)
   2. \( R \leftarrow R + r \)

Return \( R/|h| \)
A Rejection Sampling approach

Rejection_Sampler(policy \( \pi \), events \((\vec{x}, a, r, p)\)^T )
Let \( h = \emptyset \) a history, \( R = 0 \)
For each event \((\vec{x}, a, r, p)\)

1. If \( \pi(h, \vec{x}) = a \)
2. then with probability \( \frac{p_{\min}}{p} \)
   1. \( h \leftarrow h \cup (\vec{x}, a, r) \)
   2. \( R \leftarrow R + r \)

Return \( R/|h| \)

Theorem: For all history lengths \( T \), For all nonstationary policy \( \pi \), and all IID worlds \( D \), the probability of a simulated history of length \( T \) = the probability of the same history of length \( T \) in the real world.
A Master Evaluator

Eval(policy $\pi$, events $(\vec{x}, a, r, p)^T$, quantile $\rho$, bound $b$)
Let $h = \emptyset$, $R = 0$, $C = 0$, $Q = \emptyset$, $c = b$
For each event $(\vec{x}, a, r, p)$

1. $R \leftarrow R + c \left( \frac{\pi(a|x,h)}{p}(r - \hat{r}(x, a)) + \sum_{a'} \pi(a'|x, h)\hat{r}(x, a') \right)$
2. $C \leftarrow C + c$
3. $Q \leftarrow Q \cup \left\{ \frac{p}{\pi(a|x,h)} \right\}$
4. With probability $\frac{c\pi(a|x,h)}{p}$:
   1. $h \leftarrow h + (x, a, r)$
   2. $c \leftarrow \min\{b, \rho\text{-th quantile of } Q\}$

Return $R/C$
A Master Evaluator

Eval(policy $\pi$, events $(\vec{x}, a, r, p)^T$, quantile $\rho$, bound $b$)

Let $h = \emptyset$, $R = 0$, $C = 0$, $Q = \emptyset$, $c = b$

For each event $(\vec{x}, a, r, p)$

1. $R \leftarrow R + c \left( \frac{\pi(a|x,h)}{p} (r - \hat{r}(x, a)) + \sum_{a'} \pi(a'|x, h) \hat{r}(x, a') \right)$
2. $C \leftarrow C + c$
3. $Q \leftarrow Q \cup \left\{ \frac{p}{\pi(a|x,h)} \right\}$
4. With probability $\frac{c\pi(a|x,h)}{p}$:
   - $h \leftarrow h + (x, a, r)$
   - $c \leftarrow \min\{b, \rho\text{-th quantile of } Q\}$

Return $R/C$

Incorporates Double Robust + Nonstationary evaluation.

Theorem: Introduces bounded bias + much more efficient.

Empirically, an order of magnitude better for nonstationary eval.
An improved(?) Master Evaluator

Eval(policy $\pi$, events $(\vec{x}, a, r, p)^T$)

Let $h = \emptyset$, $R = 0$

For each event $(\vec{x}, a, r, p)$

1. $R \leftarrow R + \frac{\pi(a|x, h)}{p} (r - \hat{r}(x, a)) + \sum_{a'} \pi(a'|x, h) \hat{r}(x, a')$

2. if $\frac{\pi(a|x, h)}{p} < 1$ With probability $\frac{\pi(a|x, h)}{p}$:
   - $h \leftarrow h + (x, a, r)$ with importance weight 1

3. else
   - $h \leftarrow h + (x, a, r)$ with importance weight $\frac{\pi(a|x, h)}{p}$

Return $R / T$
An improved (?) Master Evaluator

Eval(policy $\pi$, events $(\vec{x}, a, r, p)^T$)

Let $h = \emptyset$, $R = 0$

For each event $(\vec{x}, a, r, p)$

1. $R \leftarrow R + \frac{\pi(a|x, h)}{p}(r - \hat{r}(x, a)) + \sum_{a'} \pi(a'|x, h)\hat{r}(x, a')$

2. if $\frac{\pi(a|x, h)}{p} < 1$ With probability $\frac{\pi(a|x, h)}{p}$:
   1. $h \leftarrow h + (x, a, r)$ with importance weight 1

3. else
   1. $h \leftarrow h + (x, a, r)$ with importance weight $\frac{\pi(a|x, h)}{p}$

Return $R/T$

Does this work in theory? (...it seems to work well in practice)

vw –explore_eval –epsilon 0.05 〈cb_adf_dataset〉
vw –explore_eval –multiplier 0.2 –epsilon 0.05 〈cb_adf_dataset〉
Bibliography

