Exploration Scavenging in comparison to other off-policy estimators.

Exploration Scavenging is an off-policy estimator for contextual bandits. Many off-policy estimators for contextual bandits exist, such as, IPS [Rosenbaum 1983], SNIPS [Swaminathan 2015], DR [Dudik 2014], Direct Matching, and Exploration Scavenging [Langford 2008]. These Off-Policy estimators have different operating conditions and also different bias/variance.

Table 1: Various Off-Policy Evaluators

Estimator	Needs Non-Ctx Logging Policy	Requires Propensity	Is Unbiased	Variance or Conc. Bound
ES	Yes	No	Yes	$\sum_a \sqrt{\frac{\log(kT/\delta)}{T_a}}$
IPS	No	Yes	Yes	TODO
SNIPS	No	Yes	No	TODO
DR	No	Yes	Yes	TODO

Exploration Scavenging Assume we are given a dataset $\mathcal{D}$ containing triples of (x = feature, a = action, r = reward) generated by following some logging policy, then the exploration scavenging policy paper's theorem 2 tells us that For any distribution D over (x, r) and any exploration policy π such that 1) Let the number of times each action is chosen during T trials be T_a. T_a is a random variable. The first condition is that $P(T_a > 0) = 1$. 2) π chooses a_t independent of x_t, in other words, the logging policy is non-contextual, but the logging policy can also be non-stationary! Then the ES OPE – defined below – is an unbiased estimator for the value of the new policy h

$$\mathbb{E}_{\{(x,r)\sim D\}}\Big(\sum_{t=1}^{T} \frac{r_{t,a}\mathbb{I}(h(x)=a)}{T_{a_t}}\Big)$$

The crucial requirements in the proof are that

1. $\mathbb{I}(h(x_t)=1)$ is independent of time.
2. $\mathbb{I}(h(x_t)=1)$ is independent of $r_{t,a}$.
3. $r_{t,a}$ is independent of time.

Comparing the different variance of different estimators TODO
It will be good to know that how much does IPS or SNIPS buy us in the regime where ES can work?
IPS is analyzed in general in http://alekhagarwal.net/bandits_and_rl/off_policy.pdf. I'll like to know how does IPS perform given a non-stationary non-contextual logging policy? What is the value of additionally knowing exact propensities of an evolving logging policy? The SNIPS and DR are analyzed in their respective papers. How well do they perform?

References

[Rosenbaum 1983] "The central role of the propensity score in observational studies for causal effects." Rosenbaum, Paul R. and Rubin, Donald B. Biometrika (1983)

[Langford 2008] "Exploration scavenging." Langford, John and Strehl, Alexander and Wortman, Jennifer. ICML (2008)

[Swaminathan 2015] "The self-normalized estimator for counterfactual learning." Swaminathan, Adith and Joachims, Thorsten. Neurips (2015)

[Dudik 2014] "Doubly robust policy evaluation and optimization." Dudík, Miroslav and Erhan, Dumitru and Langford, John and Li, Lihong. Statistical Science (2014)