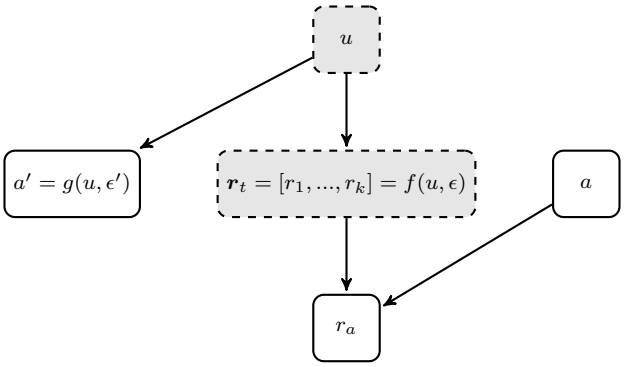
# 1 Contextual bandit related problems

The confounded causal bandit problem, as introduced by [Bareinboim et al., 2015], is illustrated in figure 1. At each timestep $t$, $u_t$, $\boldsymbol{r_t}$ and $a'_t$ are generated stochastically. Only $a'$ is observed. The algorithm then selects an action $a_t \in \{1, ..., k\}$ and observes the corresponding reward $r_{a_t}$. $a' \in \{1, ..., k\}$ corresponds to the action that would have been selected if action choice were not decided by the algorithm. This corresponds exactly to an instance of the standard stochastic contextual bandit problem, with context $x_t = a'_t$. See the following definitions from [Langford and Zhang, 2008]

Figure 1: Causal model for confounded bandits




**Definition 1** (Contextual bandit problem)**.** In a contextual bandits problem, there is a distribution $P$ over $(x, r_1, ..., r_k)$, where $x$ is context, $a \in \{1...k\}$ is one of the $k$ arms to be pulled,and $r_a \in [0, 1]$ is the reward for arm $a$. The problem is a repeated game: on each round, a sample $(x, r_1, ..., r_k)$ is drawn from $P$, the context $x$ is announced, and then for precisely one arm a chosen by the player, its reward $r_a$ is revealed.

**Definition 2** (Contextual bandit algorithm)**.** A contextual bandit algorithm $\mathcal{B}$ determines an arm $a \in \{1...k\}$ to pull at each timestep $t$, based on the previous observation sequence $(x_1, a_1, r_{a,1})$, and the current context $x_t$.

**Definition 3** (Regret)**.** Assume we have a hypothosis set $\mathcal{H}$ consisting of hypotheses $h : \mathcal{X} \to \{1, ..., k\}$ The regret of an algorithm $\mathcal{B}$ with respect to $h$ is

$$\Delta R(h, \mathcal{B}, T) = T\mathbb{E}\left[r_{h(x)}\right] - \mathbb{E}\left[\sum_{t=1}^{T} r_{\mathcal{B}(x),t}\right]$$

and the (psudo) regret of $\mathcal{B}$ with respect to the hypothesis space $\mathcal{H}$ is

$$\Delta R(\mathcal{H}, \mathcal{B}, T) = \sup_{h \in \mathcal{H}} \Delta R(h, \mathcal{B}, T)$$

Note - that the definition of a contextual bandit problem does not require that the context is a direct cause of the outcome.

# 2 Questions

Two questions arose in our discussion of this problem.

## 2.1 Are noisy features more problematic in the bandit setting

Intuitively adding features that are noisy (ie only weakly related to the target) is more problematic in the bandit setting than in the batch learning setting because in batch learning, we can apply feature selection or regularization and tune it via cross-validation. It is unclear how to do this (or model/parameter selection in general) in the contextual bandit setting.

Some more concrete version of the question.

- Assume we are in the confounded causal bandit setting and $Y_t \in \{0, 1\}$ Suppose we make no assumptions about $f$ track the rewards separately for each value of $a'_t$. Under what conditions (on $g$, $f$ and $T$) does including the context have lower regret than applying a standard bandit algorithm.

- What happens if we add context that is entirely uncorrelated with the target. Ie a contextual bandit where the context $\boldsymbol{x} = [x_1, ...x_m]$ but $Y = f(x_1...x_d, \epsilon)$ where $d < m$. How does the regret depend on $m$ and $d$ (with or witout assumptions such as linearity on f). This paper may be relevant [Abbasi-Yadkori et al., 2012].
- How can we quantify dependence without reference to a model? KL divergence between $P(X, Y)$ and $P(Y)$ averaged over $X$? max over possible kernelized fits (as in kernelized conditional independence testing)? For linear models, all captured in the inverse correlation matrix.

### 2.2 How should we represent the context

In the confounded bandit problem, $x$ is discrete with $k$ unique values. We could:

1. Treat it as a categorical feature
2. Treat each value of $x$ as a task and apply a multi-task algorithm
3. Treat it as a structured output prediction problem (how)?

Its hard to find a quantitative definition of multi-task learning. One variant involves parameterising models for each task in terms of some weights and putting a joint constraint over the weights of all tasks **??**. This is equivalent to treating the tasks as a categorical variable and regularizing for some choices of constraint, and regularizer (given conditions on the loss-function).

The feature-mappings for structured output prediction involve the label - so at first pass this seems different -but I haven't properly understood how this approach works and how the problem can be represented as a structured output prediction problem.

## References


[Abbasi-Yadkori et al., 2012] Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. (2012). Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits. In *AISTATS*, volume XX, pages 1–9.

[Bareinboim et al., 2015] Bareinboim, E., Forney, A., and Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350.

[Langford and Zhang, 2008] Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824.