Contributions

- We provide the first public dataset with accurately logged propensities from a production interactive system with recorded user feedback:
  - The dataset was collected at Criteo.
  - The dataset enables research into the problem of Batch Learning from Bandit Feedback (BLBF).
- We propose new sanity checks and evaluation methodologies when running BLBF experiments.
- We provide a standardized test-bed that implements our workflow and benchmark several counterfactual learning algorithms in a sample BLBF task.

Motivation

![BLBF algorithm](image)

Figure: BLBF algorithm.

This dataset and test-bed will hopefully enable research into:
- New training objectives, learning algorithms, and regularization mechanisms;
- Improved model selection procedures (analogous to cross-validation);
- Effective and tractable policy classes \( \pi \in \Pi \) for the specified task \( x \rightarrow y \); and
- Algorithms that can scale to massive amounts of data.

Dataset

The logging policy \( \pi \) stochastically selects products to construct a banner by first computing non-negative scores \( f_p \) for all candidate products \( p \in P \), and using:

\[
P(\text{slot1} = p) = \frac{f_p}{\sum_{p' \in P} f_{p'}} \quad P(\text{slot2} = p' \mid \text{slot1} = p) = \frac{f_{p'}}{\sum_{p' \in P} f_{p'}}
\]

The propensity of a chosen banner \( \{p_1, p_2, \ldots\} \) is \( P(\text{slot1} = p_1) \times P(\text{slot2} = p_2 \mid \text{slot1} = p_1) \times \ldots \) and our dataset was logged as follows:

```
example $(\text{ex10})$: $(\text{itemID})$ $(\text{buahID})$ $(\text{wasAdClicked})$ $(\text{propensity})$ $(\text{nbSlots})$
$($\text{itemCandidate1}$) $(\text{displayFormat})$ $(\text{adType1})$ $(\text{adType2})$ $(\text{adType3})$ $(\text{adType4})$
$($\text{wasProductClicked}$) $(\text{exid})$ $(\text{ex10})$ $(\text{productFeat1})$ $(\text{productFeat2})$ $(\text{productFeat3})$ $(\text{productFeat4})$
```

Download our dataset at:

http://www.cs.cornell.edu/~adith/Criteo/index.html

Statistics

Sub-sampling to limit dataset size. Accounted for in the statistics and subsequent evaluation in our code.

<table>
<thead>
<tr>
<th>#Slots</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Impressions</td>
<td>( \frac{2.13 \pm 0.07}{N} )</td>
<td>( \frac{3.35 \pm 0.07}{N} )</td>
<td>( \frac{2.21 \pm 0.07}{N} )</td>
<td>( \frac{0.92 \pm 0.07}{N} )</td>
<td>( \frac{2.93 \pm 0.07}{N} )</td>
<td>( \frac{1.40 \pm 0.07}{N} )</td>
</tr>
<tr>
<td>Avg(InvPropensity)</td>
<td>11.96</td>
<td>12.92</td>
<td>13.76</td>
<td>14.27</td>
<td>15.23</td>
<td>15.86</td>
</tr>
<tr>
<td>Max(InvPropensity)</td>
<td>5.36</td>
<td>5.38</td>
<td>5.38</td>
<td>5.38</td>
<td>5.38</td>
<td>5.40</td>
</tr>
</tbody>
</table>

Table. Number of impressions and propensity statistics for slices of traffic with k-slot banners with \( 1 \leq k \leq 6 \). Estimated sample size \( \hat{N} \) corrects for 10% sub-sampling of non-clicked impressions.

Consequences:
- Don’t rely on a single point estimate (like IPS), but report multiple estimates.
- Confidence intervals can mislead (esp. when \( k \geq 4 \)).

Benchmark Learning Algorithms

- Slice of traffic can enable logged contextual bandit learning: 1-slot filling task.
  - Regression to predict CTR of candidates. Pick best estimated CTR;
  - Off-policy learning method like DRO or POEM.

Results for 1-slot task

<table>
<thead>
<tr>
<th>Approach</th>
<th>Test set estimates</th>
<th>( \hat{R}(\pi) \times 10^4 )</th>
<th>( \hat{C}(\pi) \times 10^4 )</th>
<th>( \hat{C}(\pi) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random ( \pi )</td>
<td>44.67</td>
<td>45.46</td>
<td>0.98</td>
<td>0.91</td>
</tr>
<tr>
<td>Regression</td>
<td>53.50</td>
<td>53.50</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>IPS</td>
<td>48.35</td>
<td>48.16</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td>DRO</td>
<td>54.12</td>
<td>54.12</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>POEM</td>
<td>57.35</td>
<td>57.35</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table. Test set performance of policies learnt using different counterfactual learning baselines. Errors bars are 99% confidence intervals under a normal distribution. Confidence interval for SNIPS is constructed using the delta method.

Where \( \hat{C}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \pi(x_i)}{\sum_{p \in P} \pi(x_i) f_p} \) and \( \hat{R}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \pi(x_i)}{\sum_{p \in P} \pi(x_i) f_p} \).

Grand BLBF challenges

- Size of the action space: Increase the size of the action space.
- Feedback granularity: Use per item feedback.
- Contextualization: We can learn a separate model for each banner type or learn a contextualized model across multiple banner types.

We hope you find this first public user impressions dataset with logged propensities useful for your research.

References