Large-scale Validation of Counterfactual Learning Methods: A Test-Bed

Damien Lefortier[1,2], Adith Swaminathan[3], Xiaotao Gu[4], Thorsten Joachims[3], and Maarten de Rijke[2]

[1] Facebook
[2] University of Amsterdam
[3] Cornell University, Ithaca, NY
[4] Tsinghua University, Beijing, China

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

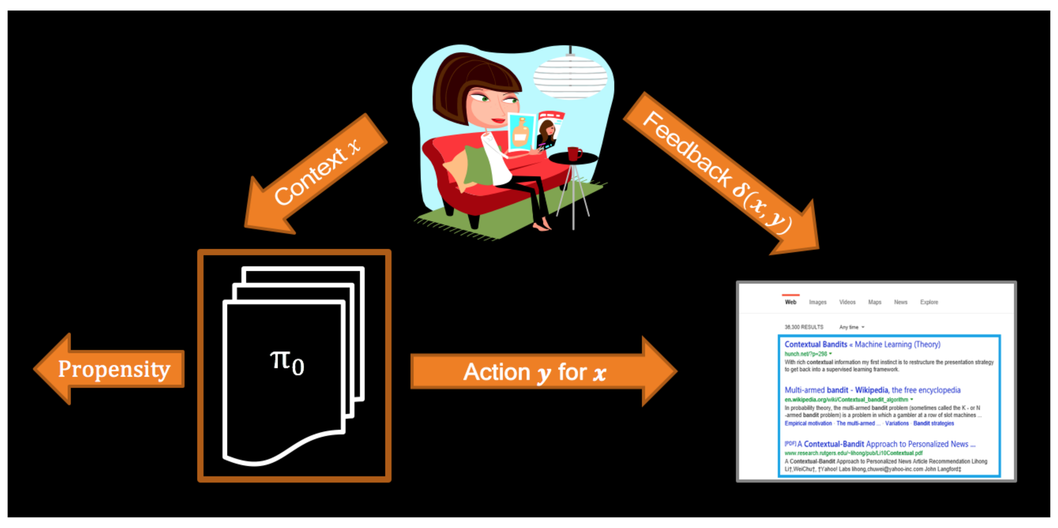

Figure: BLBF algorithm.

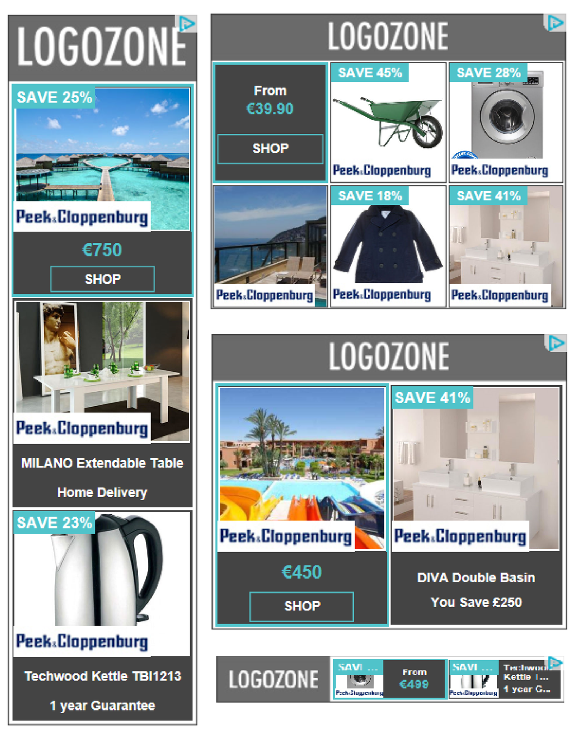

Figure: Concrete example: banner-filling task at Criteo.

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 $x \mapsto y$; and
- Algorithms that can scale to massive amounts of data.

Dataset

The logging policy π_0 stochastically selects products to construct a banner by first computing non-negative scores f_p for all candidate products $p \in P_c$, and using:

$$P(slot1 = p) = \frac{f_p}{\Sigma_{\{p' \in P_c\}} f_{p'}} \qquad P(slot2 = p' \mid slot1 = p) = \frac{f_{p'}}{\Sigma_{\{p^\dagger \in P_c \wedge p^\dagger \neq p\}} f_{p^\dagger}}, \quad \ldots$$

The propensity of a chosen banner ad $\langle p_1, p_2, \ldots \rangle$ is $P(slot1 = p_1) * P(slot2 = p_2 \mid slot1 = p_1) * \ldots$. and our dataset was logged as follows:

```
 example ${exID}:  ${hashID} ${wasAdClicked} ${propensity} ${nbSlots}
${nbCandidates} ${displayFeat1}:${v_1} ...
 ${wasProduct1Clicked} exid:${exID} ${productFeat1_1}:${v1_1} ...
 ...
 ${wasProductMClicked} exid:${exID} ${productFeatM_1}:${vM_1} ...
```

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.

#Slots	1	2	3	4	5	6
#Impressions	$2.13e+07$	$3.55e+07$	$2.27e+07$	$6.92e+06$	$2.95e+06$	$1.40e+07$
$\hat{N}$	$2.03e+08$	$3.39e+08$	$2.15e+08$	$6.14e+07$	$2.65e+07$	$1.30e+08$
Avg(InvPropensity)	11.96	$3.29e+02$	$1.87e+04$	$2.29e+06$	$2.62e+07$	$3.51e+09$
Max(InvPropensity)	$5.36e+05$	$3.38e+08$	$3.23e+10$	$9.78e+12$	$2.03e+12$	$2.34e+15$

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 >= 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

Approach	Test set estimates $\hat{R}(\pi_\epsilon) \times 10^4$	$\hat{R}(\pi_\epsilon) \times 10^4 / \hat{C}(\pi_\epsilon)$	$\hat{C}(\pi_\epsilon)$
Random	44.676 ± 2.112	45.446 ± 0.001	0.983 ± 0.021
π_0	53.540 ± 0.224	53.540 ± 0.000	1.000 ± 0.000
Regression	48.353 ± 3.253	48.162 ± 0.001	1.004 ± 0.041
IPS	54.125 ± 2.517	53.672 ± 0.001	1.008 ± 0.016
DRO	57.356 ± 14.008	57.086 ± 0.005	1.005 ± 0.025
POEM	58.040 ± 3.407	57.480 ± 0.001	1.010 ± 0.018

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 $\hat{C}(\pi) = \frac{1}{\hat{N}} \Sigma_{i=1}^{N} \frac{\pi(y_i|x_i)}{q_i} \frac{1\{o_i=1\}}{\Pr(O=1|\delta_i)}$ and $\hat{R}(\pi) = \frac{1}{\hat{N}} \Sigma_{i=1}^{N} \delta_i \frac{\pi(y_i|x_i)}{q_i} \frac{1\{o_i=1\}}{\Pr(O=1|\delta_i)}$.

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

[1] Counterfactual reasoning and learning systems: the example of computational advertising. L. Bottou et al. JMLR 2013.

[2] Batch learning from logged bandit feedback through counterfactual risk minimization. A. Swaminathan et al. JMLR 2015.

[3] Doubly Robust Policy Evaluation and Learning. M. Dudik et al. ICML 2011.

[4] Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms. L. Li et al. WSDM 2011.

[5] The self-normalized estimator for counterfactual learning. A. Swaminathan et al. NIPS 2015.