Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques Supplementary material

Olivier Nicol OLI.NICOL@GMAIL.COM
Jérémie Mary JEREMIE.MARY@INRIA.FR
Philippe Preux PHILIPPE.PREUX@UNIV-LILLE3.FR
University of Lille / LIFL (CNRS) & INRIA Lille Nord Europe, 59650 Villeneuve d'Ascq, France

Abstract

These notes contain supplementary material for the paper entitled "Improving offline evaluation of contextual bandit algorithms" submitted to ICML 2014. Mainly we detail the modifications that were made following the cycle 1 reviews and provide an implementation of the state-of-the-art replay method using our notations just for the record. Note that the main paper is entirely self contained.

1. Detail of the modifications

Compared to our submission for the first cycle, the paper was rewritten to address the concerns raised in the reviews. There were four major ones:

1. An overstating title (*Offline evaluation of recommender systems*).
2. A lack of motivation for this work.
3. A too long introductory part (first three sections of the previous version).
4. Not enough details about the experiments.

The title was changed to *Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques*. Although the main application of this work is still recommendation, it disappears from the title so that it reflects more our contributions. We rewrote completely the first sections with two purposes in mind: (i) describing the problem and our contribution with more efficiency and clarity; (ii) motivating the problem further. To motivate our work further we:

1. motivated the use of contextual bandits for some specific yet widely spread recommendation applications (section 1).
2. emphasized that the estimation of the distribution of the CTR that bootstrapping enables is a very desirable thing in practice.
3. explained the main limitation of the previous works that we called time acceleration (first part of section 3).
4. argued that in real life, acquiring more data was not a solution to this issue and that being more data-efficient in our simple setting was in fact a way to solve the real problem. (second part of section 3)

The theoretical and empirical parts were not changed a lot. The remarks were made more concise, an estimator quality assessment ξ was introduced for more theoretical accuracy (it is typical in bootstrapping theory (Kleiner et al., 2012; Efron, 1979; Horowitz, 2001)) and the notations were lightened when possible. The proof of the second theorem was included in the main paper for it is very important to understand why our approach works. Finally the experiments remained unchanged but more details were added to allow repeatability of the results.

2. Miscellaneous

The detailed implementation of *replay* using our notations is given in algorithm 1. Note that apart from notations, no modification are made. Figure 1 is the same experiment as in section 6.1 but with a non-contextual algorithm UCB. Although the improvement compared to the state of the art is significant, it was not included in the main paper for lack of space. The figure about LinUCB (figure 2) that we did include in the main paper is more informative as it exhibits both the importance of Jittering and the improve-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

ment brought by our method compared to the state of the art.

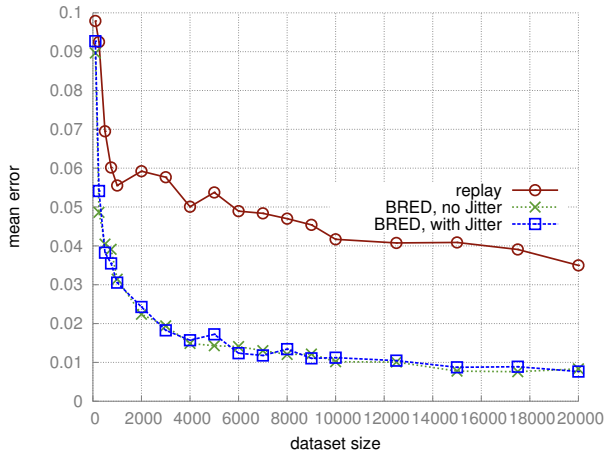

Figure 1. Mean of the absolute value of the difference between the true CTR of a UCB and the estimated one for different methodologies. Conducted on artificial dataset as described in the section 6.1 of the main paper. The lower, the better. Jittering is useless here because UCB does not use the context.

References

Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. ISSN 00905364. doi: 10.2307/2958830.

Horowitz, Joel L. The bootstrap. *Handbook of econometrics*, 5:3159–3228, 2001.

Kleiner, Ariel, Talwalkar, Ameet, Sarkar, Purnamrita, and Jordan, Michael. The big data bootstrap. In Langford, John and Pineau, Joelle (eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pp. 1759–1766, New York, NY, USA, July 2012. Omnipress. ISBN 978-1-4503-1285-1.

Langford, John, Strehl, Alexander, and Wortman, Jennifer. Exploration scavenging. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 528–535, 2008.

Li, Lihong, Chu, Wei, Langford, John, and Wang, Xuanhui. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In King, Irwin, Nejdl, Wolfgang, and Li, Hang (eds.), *Proc. Web Search and Data Mining (WSDM)*, pp. 297–306. ACM, 2011. ISBN 978-1-4503-0493-1.

Algorithm 1 *Replay method* (Langford et al., 2008; Li et al., 2011).

Remark: for the sake of the precision of the specification of the algorithm, we use a history h which is the list of triplets (x, a, r) that have yet been used to estimate the performance of the algorithm A. The goal is to avoid hiding internal information maintenance in A; a real implementation may be significantly different for the sake of efficiency, by learning incrementally.

Input:

- A contextual bandit algorithm A
- A set S of L triplets (x, a, r)

Output: An estimate of g_A

$h \leftarrow \emptyset$
$\widehat{G}_A \leftarrow 0$
$T \leftarrow 0$
for $t \in \{1..L\}$ **do**
  Get the t-th element (x, a, r) of S
  $\pi \leftarrow A(h)$
  if $\pi(x) = a$ **then**
    add (x, a, r) to h
    $\widehat{G}_A \leftarrow \widehat{G}_A + r$
    $T \leftarrow T + 1$
  end if
end for
return $\frac{\widehat{G}_A}{T}$