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

1. 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 improvement brought by our method compared to the state of the art.

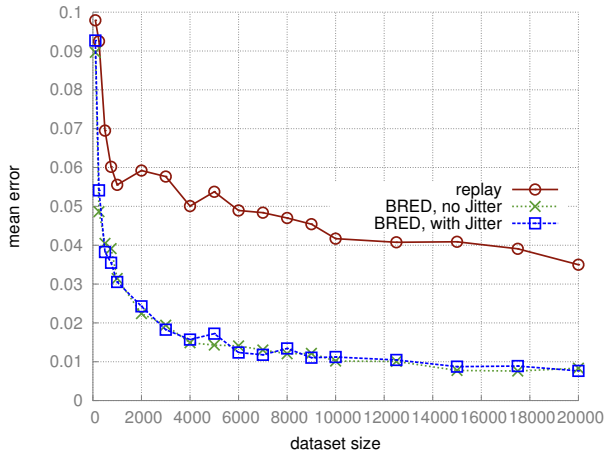

Figure 1. Mean of the absolute value of the difference between the true CTR of a UCB and the estimated one for different methodologies. Conducted on artificial dataset as described in the section 6.1 of the main paper. The lower, the better. Jittering is useless here because UCB does not use the context.

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 $\frac{\widehat{G}_A}{T}$

References

Langford, John, Strehl, Alexander, and Wortman, Jennifer. Exploration scavenging. In *Proceedings of the Inter-*

national 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.