We thank our four reviewers for their careful reading and their useful feedback. We provide answers below to the different questions raised by each reviewer. Numbered citations correspond to the bibliography of the paper.

 # Reviewer 1 Regarding the initialization step, we will clarify that each arm is drawn once at the beginning of the algorithm in order to avoid empty history. Then, regarding the experiment section, we think that the frequentist experiments in the main text are interesting to check the theory and show that the logarithmic regret is achieved in a reasonable time. However, if the space limitation allows us to add the Bayesian experiments in the main text we would gladly do so in the camera-ready version of the paper (Table 7 and Table 8).

 # Reviewer 2 Regarding the application of RB-SDA to Gaussian bandits with known variance, the variance parameter σ is actually *not* needed as an input to the algorithm. Note however that our analysis only covers the case in which all arms belong to the same one-dimensional exponential family, i.e. in the Gaussian case it means that all arms should have the *same* variance. If the variances were to be different, a variant of the duels should be introduced, with some normalization by the variance for each arm, in spirit of what is proposed in the SSTC algorithm [2].

 # Reviewer 3 First, we agree that providing a characterization of the distributions for which RB-SDA has logarithmic regret is an important future work. So far we have indeed only identified a sufficient condition in Theorem 1. We will emphasize this in our conclusion. Finding a bandit algorithm with logarithmic regret for as many distributions as possible is indeed a very interesting question. If by UCB you mean UCB1 of [5], this algorithm indeed provides regret guarantees for a large set of distributions: *sub-gaussian* distributions, and in particular *bounded* distributions. For the latter, we will add in the paper that RB-SDA can be turned to an algorithm with logarithmic regret for any bounded distribution by using the standard binarization trick proposed by [11] for Thompson Sampling. Moreover, in this work our focus is not only on logarithmic regret, or order-optimal regret (reasonable gap-dependent regret bound, as UCB1) but on asymptotically optimal regret (match the regret prescribed by the available lower bound). We will clarify this in the introduction, by providing the expression of the Lai and Robbins and Burnetas and Katehakis lower bounds.

Regarding the novelty with respect to the SSMC algorithm, we believe that deterministic history-dependent versus randomized history-independent sub-sampling is a significant difference. Indeed the philosophy of the algorithms are different: SSMC tries to disadvantage the leader by picking the smallest sub-sample (which requires a scan over all possible rolling means), while SDA uses only one sub-sample per duel, and the exploration is ensured by the diversity of the sub-samples encountered. We think that this idea may be more promising for extensions to problems in which 1) the leader may change a lot (e.g non stationary bandits), or 2) the duels are based on more costly estimators than the empirical mean (e.g. risk averse bandits), in which the computation of a "rolling" estimator will be computationally demanding in practice. One can also note that in our experiments randomized algorithms (RB-SDA, WR-SDA) tend to perform better than deterministic ones (SSMC and LB-SDA). Then, on the technical side, the novel elements of proof we developed for the analysis of randomized algorithms (such as Lemma 4.2) could be interesting for future work in this direction. Finally, we believe that understanding when we can get rid of forced exploration is conceptually important for multi-armed bandit algorithms. Indeed in more sophisticated settings such as structured bandits (see e.g. the OSSB algorithm of Combes et al. NeurIPS 2017), forced exploration can be quite harmful empirically.

The question of finding an alternative algorithm that is not round-based is very interesting. It is indeed a natural idea to choose a challenger at random instead of competing with all arms. However, we believe that this approach may not work when the leader is defined as the most sampled arm as in our setting. The intuition is the following: if the optimal arm is not competing in each round, then a sub-optimal leader with a sufficiently large lead could stay leader forever simply by beating other (possibly numerous) sub-optimal arms, even if it loses all its duels against the optimal arm. Still, we will check this intuition in practice as it may be interesting for future work.

Regarding the experimental section, we conducted all of our experiments with a larger set of algorithms including UCB1 and kl-UCB. However, we chose not to include these algorithms in the figures and tables as their regret was significantly worse than those of the algorithms we decided to represent (at least twice worse for UCB1 for instance). For this reason, we chose Thompson Sampling as the benchmark for "standard" bandit algorithms. Finally, the fact that the lower bound is much higher than the regret of good algorithms is always surprising, but it has already been observed frequently in the bandit literature, see e.g. [6,10].

 # Reviewer 4 Thanks for your suggestions of improvements for the notation and presentation of the paper. Regarding the complexity of the algorithms, the question is a bit difficult because some algorithms (SSMC, LB-SDA) can be made efficient in rounds in which the leader doesn't change and no challenger arm has been pulled. Hence the problem is to estimate the computational cost of these algorithms in the other case, whose number of occurrences is problem-dependent. We will include a more extensive discussion on these issues in our revision.