Motivation for batched setting A common critique from the reviewers is about the relevance of the batched bandit setting we consider. Our most important point on this is that the batched setting is very common in applications, e.g., in many mobile health [4-6] and online education problems [7,8] multiple users use apps / take courses simultaneously, so a batch corresponds to the number of unique users the bandit algorithm acts on at once. The batched setting is even common in online advertising because it is impractical to update the bandit after every action if many users visit the site simultaneously [1-3]. We provide a highly-abridged list of papers on real experiments run in a batched setting:

[1] doi:10.1145/2645710.2645732 (RecSys'14)
[2] doi:10.1145/2505515.2514700 (CIKM'13)
[3] doi:10.1287/mksc.2016.1023 (Mark Sci'17)
[4] doi:10.2196/jmir.7994 (JMIR'17)
[5] doi:10.1145/3381007 (ACM'20)
[6] doi:10.1093/abm/kay067 (Ann Behav Med'19)
[7] doi:10.5281/zenodo.3554749 (JEDM'19)
[8] doi:10.1073/pnas.1921417117 (PNAS'20)

In many such experimental settings T cannot be arbitrarily adjusted, e.g., in online education, courses generally cannot be made arbitrarily long, and clinical trials often run for a standard amount of time that depends on the domain science (e.g. the length of mobile health studies is a function of the scientific community's belief in how long it should take for users to form a habit). On the other hand, the number of users can in principle grow as large as funding allows, and thus for analyzing data from an experiment it is quite standard to consider asymptotics as the number of users grows.

Non-stationarity is very common in many of the problems settings [1,5,7]. In response to reviewer 3's comments, the non-stationarity is aligned with batches since batches correspond to actions selected in the same time period.

Comments specific to reviewer 5 We thank you for your constructive comments and references; we apologize for the omission and will certainly correct it in the revision. In particular, we will not claim a distinction between best arm identification and obtaining a confidence interval for the margin, and will instead explain how the two are very much connected, especially since the intervals obtained from the uniform bounds from the best arm identification literature are particularly relevant when the experiment ends at a data-dependent time. We will reference the main works in this field (and we thank you for giving us a starter list to build on) and also include a comparison in our simulation section. We were able to obtain one preliminary comparison in time for this rebuttal: in the stationary setting under Thompson Sampling ($n = 25$, $T = 25$, 0.1 clipping), the 95% confidence interval using BOLS has average width 0.405, while confidence intervals constructed with the anytime self-normalized martingale bound used in the regret bounds of many bandit algorithms (doi:10.5555/2986459.2986717) are more than double the width at 0.943.

Comments specific to reviewer 1 Regarding your comment on n and T needing to be comparable, we can interpret this in two ways—we will address both. First, to use BOLS for approximate inference, n and T can be comparable—in fact, in our experiments we often set $n = 25$ and $T = 25$; however n must be sufficiently large for a reliable asymptotic approximation. The second interpretation is that n and T must be comparable in order to run a bandit algorithm that minimizes regret. It is correct that the batched bandit algorithms we consider cannot be optimal with respect to the standard oracle (non-batched, stationary environment). However, given the realities of the problem settings bandits are used for, it can make sense to consider other oracles. Perechet et al. consider the oracle to be the best algorithm when the number of batches is limited and the batch sizes can be adjusted adaptively. For the problems we consider it is not realistic to adjust the batch sizes adaptively, e.g., if there are a fixed number of users in the study across all time points, or a certain number of users who visit a website in some time period. In our problem setting, it makes sense to choose an oracle that is optimal in the batched setting in which the batch sizes cannot be adjusted adaptively.

Comments specific to reviewer 4 To address your first concern, we now discuss why the adaptivity in the batched setting doesn't disappear even as $n \to \infty$. As we proved, when $\Delta = 0$, under common bandit algorithms the π_t does not concentrate and will always depends on whether we happened to receive greater rewards on average from arm 0 or arm 1 in the previous batches. The asymptotic dependence between the π_t and the rewards of the previous batch is what leads the OLS estimator to be asymptotically non-normal when $\Delta = 0$. Further, as we illustrate in Figure 2, when Δ is near-zero the distribution of the OLS estimator is not well approximated by a normal distribution.

To address your second concern, of course we cannot calculate the regret, since the optimal arm is unknown in real life experiments. One could propose to compare the total sum of rewards for each arm after the study is over. However, since the action selection probabilities do not concentrate when $\Delta = 0$, this means that even if the bandit happens to select one arm much more often (thus greater sum of rewards for that arm), it doesn't mean that arm is necessarily better than the other arm. For example, for ϵ-greedy, if $\Delta = 0$, across many identical experiments with different random seeds, half of the experiments will choose arm 1 more often and half the experiments will choose arm 0 more often. If we look at the difference in the average reward between arms then this is exactly the OLS estimator for the margin. We proved that this OLS estimator is asymptotically non-normal when $\Delta = 0$, so we cannot use a normal approximation to obtain valid confidence intervals for this estimator.

Comments specific to reviewer 2 (a) We agree that investigating the power is interesting, but beyond the scope of this paper. As π_t gets closer to 0.5, the power increases, but the regret increases; trading these two objectives off is non-trivial. (b) We will mention the issue of bias in the revision. (c) We allow the clipping rate to decay in Theorem 3.