We greatly appreciate the reviewers' constructive feedback and thank them for their time and interest in our manuscript.

Clarification on assumptions underlying the model (Reviewer 2). We consider an infinite population with two types of arms, each characterized by a *unique* mean reward. We assume that the rewards (not just mean rewards) are bounded in $[0, 1]$. Algorithm 1 relies on said assumptions alone. However, Algorithm 2 additionally requires knowledge of the difference between the maximal and minimal elements of the reward *support*. This information informs the calibration of the $(\theta_m)_{m\geq 1}$ sequence appearing as an input to Algorithm 2. Assumption 1 (line 102, main text) fixes said difference at 1. Thus, examples of permissible reward distributions include Bernoulli(0.5), Beta$(2, 3)$, Uniform on $[0, 1]$, etc. However, Uniform on $[0, 0.9]$ is not a permissible reward distribution. We remark that comparisons of Assumption 1 with conventional assumptions in infinite-armed bandit literature that pertain to regularity-like conditions in the neighborhood of 1, are not directly relevant. The latter concern the *reservoir distribution* of mean rewards while Assumption 1 concerns the reward distributions directly. We intend to make this distinction clearer in the revision.

Remarks on lower bound and phase transition phenomenon (Reviewers 2,3). The multiplicative factor of $\alpha(1-\alpha)$ is indeed a proof artifact and can be improved upon. Under a mild assumption on α, we have been able to eliminate this factor altogether, resulting in an asymptotic scaling that is on par with the finite-armed problem. Indeed, this reconciles with point (i) (line 275, main text). The modified proof assumes $\alpha \in (0, 1)$ fixed and invariant w.r.t. the horizon, and is not amenable to the boundary cases of $\alpha \in \{0, 1\}$. Thus, quite remarkably, a phase transition occurs in the lower bound from linear to logarithmic at $\alpha = 0$ while another from logarithmic to 0 occurs at $\alpha = 1$. We appreciate the reviewers' feedback on this aspect and intend to incorporate the necessary corrections in our revision. The intent behind restriction to consistent policies is to facilitate a direct comparison with the classical Lai and Robbins proof technique for the finite-armed bandit problem. Indeed, the assumption of asymptotic consistency is restrictive, but more generic policy classes are unwieldy for lower bound proofs due to reasons stemming from the combinatorial nature of our problem.

Clarifications on Algorithm 1 (Reviewers 1,2). The algorithm proceeds by "selecting" two arms at random and "pulling" each arm m times. If the separation between the empirical mean rewards is large enough, the algorithm commits to the empirically better arm; else discards "arm 2," randomly selects a new arm in its place, labels it "2" and pulls it m times. The process is repeated thereafter. Throughout the algorithm's lifetime, arm 1 stays fixed while the label "2" potentially recirculates between different arms. In the upper bound of Theorem 2, the $o(1)$ term is independent of (α, Δ) and can be bounded above by a true absolute constant (no dependence on free parameters of the algorithm).

Clarifications on Algorithm 2 (Reviewers 1,2). The algorithm expends all of its sampling effort on a given heterogeneous consideration set of arms only *in expectation*, not with probability 1. In fact, the probability of discarding a heterogeneous consideration set and reinitializing the algorithm is bounded away from 0 at all times. This leads to a multiplicatively larger regret with β^{-1} as the inflation factor, as opposed to merely an additive loss. β is a lower bound on the probability of never discarding a heterogeneous consideration set and depends on the reward distributions alone. In short, the regret is inflated by β^{-1} due to exploration of *new* arms happening throughout the algorithm's lifetime.

Remarks on β appearing in the upper bound of Theorem 3 (Reviewer 2). The behavior of β vis-à-vis Δ is hard to characterize mathematically. However, we empirically observe that β scales with Δ linearly on "well-separated" instances. While absent presently, we intend to include this observation in the revision. The implication is that the regret scales as $1/\Delta^2$ on well-separated instances, as opposed to the classical $1/\Delta$ scaling achievable in finite-armed bandits. In the small Δ regime, however, the precise characterization of the rate at which β vanishes is mathematically challenging and remains an open problem. Although this precludes quantification of the minimax regret and thus a comparison to the infinite-armed problem with a rich (infinite) set of types (a well-studied problem), our paper is focused on instance-dependent bounds alone. We intend to include a relevant discussion on this matter in the revision.

Improvements to Algorithm 2 (Reviewers 2,4). The reviewers are referred to points (iii) and (iv) in the "Miscellaneous remarks" section (line 275, main text) concerning potential improvements to Algorithm 2. Similar improvements to the extension of Algorithm 2 to K types are also possible. We reemphasize that the consideration set size must at all times be fixed at K, in a K-typed setting. Any more is redundant while a reduction in size shall cause the algorithm to spend a positive fraction of its sampling effort (in expectation) on inferior consideration sets, thereby incurring linear regret.

Performance of existing algorithms for infinite-armed bandits (Reviewers 2,3). A suitable modification of the algorithm in [20] (references) provably achieves poly-log regret on our problem, a significant performance degradation.

UCB vs. Thompson Sampling on the zero gap problem. In a nutshell, UCB's faster convergence on the zero gap problem has to do with the presence of the $\sqrt{\log n}$ additive bias in the UCB score. Thompson Sampling lacks in this regard, thereby causing the fraction of samples from a given arm to converge to a non-degenerate limit, evident empirically. This is also suggestive of UCB being better suited to adversarial settings than Thompson Sampling. However, this is only an empirical observation and remains a conjecture at the moment. More work on the matter is presently underway. While we have not been able to address here each and every remark by the reviewers, we have taken due note of all the points and hope that the most substantive ones have been satisfactorily answered in this rebuttal.