Bayesian Incentive-Compatible Bandit Exploration

Yishay Mansour*
Tel Aviv University, Tel Aviv, Israel, mansour.yishay@gmail.com

Aleksandrs Slivkins
Microsoft Research, New York, NY 10011, USA, slivkins@microsoft.com

Vasilis Syrgkanis
Microsoft Research, Cambridge, MA 02142, USA, vasy@microsoft.com

As self-interested individuals ("agents") make decisions over time, they utilize information revealed by other agents in the past, and produce information that may help agents in the future. This phenomenon is common in a wide range of scenarios in the Internet economy, as well as in medical decisions. Each agent would like to *exploit*: select the best action given the current information, but would prefer the previous agents to *explore*: try out various alternatives to collect information. A social planner, by means of a carefully designed recommendation policy, can incentivize the agents to balance the exploration and exploitation so as to maximize social welfare.

We model the planner's recommendation policy as a multi-arm bandit algorithm under incentive-compatibility constraints induced by agents' Bayesian priors. We design a bandit algorithm which is incentive-compatible and has asymptotically optimal performance, as expressed by regret. Further, we provide a black-box reduction from an arbitrary multi-arm bandit algorithm to an incentive-compatible one, with only a constant multiplicative increase in regret. This reduction works for very general bandit setting that incorporate contexts and arbitrary partial feedback.

Key words: mechanism design, multi-armed bandits, regret, Bayesian incentive-compatibility
History: First version: February 2015. This version: May 2019.[†]

Contents

1. Introduction

Decisions made by an individual often reveal information about the world that can be useful to others. For example, the decision to dine in a particular restaurant may reveal some observations about this restaurant. This revelation could be achieved, for example, by posting a photo, tweeting, or writing a review. Others can consume this information either directly (via photo, review, tweet, etc.) or indirectly through aggregations, summarizations or recommendations. Thus, individuals have a dual role: they both consume information from previous individuals and produce information for future consumption. This phenomenon applies very broadly: the choice of a product or experience, be it a movie, hotel, book, home appliance, or virtually any other consumer's choice, leads to an individual's subjective observations pertaining to this choice. These subjective observations can be recorded and collected, e.g., when the individual ranks a product or leaves a review, and can help others make similar choices in similar circumstances in a more informed way. Collecting, aggregating and presenting such observations is a crucial value proposition of numerous businesses in the modern Internet economy, such as TripAdvisor, Yelp, Netflix, Amazon, Waze and many others (see Table 1). Similar issues, albeit possibly with much higher stakes, arise in medical decisions: selecting a doctor or a hospital, choosing a drug or a treatment, or deciding whether to participate in a medical trial. First the individual can consult information from similar individuals in the past, to the extent that such information is available, and later he can contribute her experience as a review or as an outcome in a medical trial.

Watch this movie	Netflix
Dine in this restaurant	Yelp
Vacation in this resort	TripAdvisor
Buy this product	Amazon
Drive this route	Waze
Do this exercise	FitBit
See this doctor	SuggestADoctor
Take this medicine	medical trial

Table 1   Systems for recommendations and collecting feedback.

If a social planner were to direct the individuals in the information-revealing decisions discussed above, she would have two conflicting goals: *exploitation*, choose the best alternative given the information available so far, and *exploration*, trying out less known alternatives for the sake of gathering more information, at the risk of worsening the individual experience. A social planner would like to combine exploration and exploitation so as to maximize the social welfare, which results in the *exploration-exploitation tradeoff*, a well-known subject in Machine Learning, Operation Research and Economics.

However, when the decisions are made by individuals rather than enforced by the planner, we have another problem dimension based on the individuals' incentives. While the social planner benefits from both exploration and exploitation, each individuals' incentives are typically skewed in favor of the latter. (In particular, many people prefer to benefit from exploration done by others.) Therefore, the society as a whole may suffer from insufficient amount of exploration. In particular, if a given alternative appears suboptimal given the information available so far, however sparse and incomplete, then this alternative may remain unexplored – even though it may be the best.

The focus of this work is how to incentivize self-interested decision-makers to explore. We consider a social planner who cannot control the decision-makers, but can communicate with them, e.g., recommend an action and observe the outcome later on. Such a planner would typically be implemented via a website, either one dedicated to recommendations and feedback collection (such as Yelp or Waze), or one that actually provides the product or experience being recommended (such as Netflix or Amazon). In medical applications, the planner would be either a website that

rates/recommends doctors and collects reviews on their services (such as *rateMDs.com* or *SuggestADoctor.com*), or an organization conducting a medical trial. We are primarily interested in exploration that is efficient from the social planner's perspective, *i.e.,* exploration that optimizes the social welfare.[1]

Following Kremer et al. [2014], we consider a basic scenario when the only incentive offered by the social planner is the recommended experience itself (or rather, the individual's belief about the expected utility of this experience). In particular, the planner does not offer payments for following the recommendation. On a technical level, we study a mechanism design problem with an explore-exploit tradeoff and auxiliary incentive-compatibility constraints. Absent these constraints, our problem reduces to *multi-armed bandits* (*MAB*) with stochastic rewards, the paradigmatic and well-studied setting for exploration-exploitation tradeoffs, and various generalizations thereof. The interaction between the planner and a single agent can be viewed as a version of the *Bayesian Persuasion* game [Kamenica and Gentzkow 2011] in which the planner has more information due to the feedback from the previous agents; in fact, this *information asymmetry* is crucial for ensuring the desired incentives.

1.1. Our model and scope

We consider the following abstract framework, called *incentive-compatible exploration.* The social planner is an algorithm that interacts with the self-interested decision-makers (henceforth, *agents*) over time. In each round, an agent arrives, chooses one action among several alternatives, receives a reward for the chosen action, and leaves forever. Before an agent makes her choice, the planner sends a message to the agent which includes a recommended action. Everything that happens in a given round is observed by the planner, but not by other agents. The agent has a Bayesian prior on the reward distribution, and chooses an action that maximizes its Bayesian expected reward given the algorithm's message (breaking ties in favor of the recommended action). The agent's prior is known to the planner, either fully or partially. We require the planner's algorithm to be *Bayesian incentive-compatible* (henceforth, *BIC*), in the sense that each agent's Bayesian expected reward is maximized by the recommended action. The basic goal is to design a BIC algorithm so as to maximize social welfare, *i.e.,* the cumulative reward of all agents.

The algorithm's message to each agent is restricted to the recommended action (call such algorithms *message-restricted*). Any BIC algorithm can be turned into a message-restricted BIC algorithm which chooses the same actions, as long as the agents' priors are exactly known to the planner.[2] Note that a message-restricted algorithm (BIC or not) is simply an MAB-like learning algorithm for the same setting.

A paradigmatic example is the setting where the reward is an independent draw from a distribution determined only by the chosen action. All agents share a common Bayesian prior on the reward distribution; the prior is also known to the planner. No other information is received by the algorithm or an agent (apart from the prior, the recommended action, and the reward for the chosen action). We call this setting *BIC bandit exploration.* Absent the BIC constraint, it reduces to the MAB problem with IID rewards and Bayesian priors. We also generalize this setting in several directions, both in terms of the machine learning problem being solved by the planner's algorithm, and in terms of the mechanism design assumptions on the information structure.

We assume that each agent knows which round he is arriving in. This assumption is without loss of generality, because an algorithm that is BIC under this assumption is also BIC for a more general version in which each agent has a Bayesian belief about his round. (Also, the round can simply be included in the planner's message to this agent.)

---

[1] In the context of Internet economy, the "planner" would be a for-profit company. Yet, the planner's goal, for the purposes of incentivinzing exploration, would typically be closely aligned with the social welfare.

[2] This is by "revelation principle", as observed in Kremer et al. [2014] (see Remark 3.1 in the Preliminaries).

Discussion. BIC exploration can be interpreted as a protection from the *selection bias*, when the population that participates in the experiment differs from the target population, potentially affecting the findings. More specifically, both action and the observed outcome may depend on agents' properties, *e.g.,* the people who rate a particular niche movie are the people who chose to see this movie, and therefore (this being a niche movie) are much more likely to enjoy it.

BIC exploration does not rely on "external" incentives such as monetary payments or discounts, social status distinctions (*e.g.,* leaderboards or merit badges for prolific feedback contributors), or people's affinity towards experimentation. Such approaches can themseleves cause some selection bias. Indeed, paying patients for participation in a medical trial may be more appealing to poorer patients; offering discounts for new services may attract customers who are more sensitive to such discounts; and relying on people who like to explore for themselves would lead to a dataset that represents this category of people rather than the general population. While all these approaches are reasonable and in fact widely used (with well-developed statistical tools to mitigate the selection bias), an alternative intrinsically less prone to selection bias is, in our opinion, worth investigating.

The "intrinsic" incentives offered by BIC exploration can be viewed as a guarantee of *fairness* for the agents. Indeed, even though the planner imposes experimentation on the agents, the said experimentation does not degrade expected utility of any one agent. This is simply because an agent can always choose to ignore the planner's recommendation and select an action with the highest prior mean reward. This is particularly important for settings in which "external" incentives described above do not fully substitute for low intrinsic utility of the chosen actions. For example, a monetary payment does not fully substitute for an adverse outcome of a medical trial, and a discounted meal at a restaurant does not fully substitute for a bad experience.

However, the BIC property is stronger than the fairness property discussed above. Indeed, following the recommendation should be compared not only to ignoring it (and going with the best action according to the prior), but also to choosing some other action *given the recommendation*. In other words, an agent may see a particular action being recommended, do a Bayesian update on this fact, and decide to choose a different action; BIC property does not allow this to happen.

We focus on message-restricted algorithms, and rely on the BIC property to convince agents to follow our recommendations. We do not attempt to make our recommendations more convincing by revealing additional information, because doing so does not help in our model, and because the desirable kinds of additional information to be revealed are likely to be application-specific (whereas with message-restricted algorithms we capture many potential applications at once). Further, message-restricted algorithms are allowed, and even recommended, in the domain of medical trials (see Section 1.4 for discussion).

We make the standard assumptions of trust and rationality. We assume the "power of commitment": the planner can commit to a particular policy, and the agents trust that the planner actually implements it. We posit that the agents are Bayesian and risk-neutral.[3] Finally, we assume that the agents either trust that our mechanisms are Bayesian incentive-compatible, or are able to verify this property themselves or via third party. While non-trivial, such assumptions are standard and tremendously useful in many scenarios throughout economics.

Objectives. We seek BIC algorithms whose performance is near-optimal for the corresponding setting without the BIC constraint. This is a common viewpoint for welfare-optimizing mechanism design problems, which often leads to strong positive results, both prior-independent and Bayesian, even if Bayesian-optimal BIC algorithms are beyond one's reach. Prior-independent guarantees are particularly desirable because priors are almost never completely correct in practice.

We express prior-independent performance of an algorithm via a standard notion of *regret*: the difference, in terms of the cumulative expected reward, between the algorithm and the the best fixed action. Intuitively, it is the extent to which the algorithm "regrets" not knowing the best action in advance. For Bayesian performance, we consider *Bayesian regret*: ex-post regret in expectation

---

[3] A version of risk-aversion may be modeled in a standard way, by redefining rewards for the undesirable outcomes.

over the prior, and also the average Bayesian-expected reward per agent. (For clarity, we will refer to the prior-independent version as *ex-post regret*.) Moreover, we consider a version in which the algorithm outputs a prediction after each round (visible only to the planner), e.g., the predicted best action; then we are interested in the rate at which this prediction improves over time.

1.2. Our contributions

On a high level, we make two contributions:

Regret minimization We provide an algorithm for BIC bandit exploration whose ex-post regret is asymptotically optimal among all MAB algorithms (assuming a constant number of actions). Our algorithm is *detail-free*, in that it requires only a limited knowledge of the prior.

Black-box reduction We provide a reduction from an arbitrary learning algorithm to a BIC one, with only a minor loss in performance; this reduction "works" for a very general setting.

In what follows we discuss our results in more detail.

Regret minimization. Following the literature on regret minimization, we focus on the asymptotic ex-post regret rate as a function of the time horizon (which in our setting corresponds to the number of agents). We establish that the BIC restriction does not affect the asymptotically optimal ex-post regret rate. The optimality is two-fold: in the worst case over all realizations of the common prior (*i.e.,* for every possible vector of expected rewards), and for every particular realization of the prior (which may allow much smaller ex-post regret than in the worst-case). More formally, if T is the time horizon and Δ is the "gap" in expected reward between the best action and the second-best action, then our algorithm achieves ex-post regret

$$c_{\mathcal{P}} + c_0 \cdot \min(\tfrac{m}{\Delta} \log T, \sqrt{mT \log T}), \tag{1.1}$$

where c_0 is an absolute constant, and $c_{\mathcal{P}} < 0$ depends only on the common prior $\mathcal{P}$. Without the BIC restriction, bandit algorithms achieve a similar regret bound, but without the $c_{\mathcal{P}}$ term [Auer et al. 2002a]. No bandit algorithm, BIC or not, can achieve ex-post regret better than $O(\min(\tfrac{m}{\Delta} \log T, \sqrt{mT}))$ [Lai and Robbins 1985, Auer et al. 2002b]. [4]

Conceptually, our algorithm implements *adaptive exploration*: the exploration schedule is adapted to the observations, so that low-performing actions are phased out early. This is known to be a vastly superior approach compared to exploration schedules that are fixed in advance.[5]

Further, our algorithm is *detail-free*, requiring only a limited knowledge of the common prior. This is desirable because in practice it may be complicated or impossible to elicit the prior exactly. Moreover, this feature allows the agents to have different priors (as long as they are "compatible" with the planner's prior, in a precise sense specified later). In fact, an agent does not even need to know her prior exactly: instead, she would trust the planner's recommendation as long as she believes that their priors are compatible.

The prior-dependent additive term $c_{\mathcal{P}}$ in (1.1) can get arbitrarily large depending on the prior, even for two arms. Informally, the magnitude of $c_{\mathcal{P}}$ is proportional to how difficult it is to convince an agent to try a seemingly suboptimal action. This term can also grow exponentially with the number of arms m in some examples. While (very) suboptimal for multi-armed bandits, we conjecture that such dependence is an inevitable "price of incentive-compatibility".

Black-box reduction. Given an arbitrary MAB algorithm $\mathcal{A}$, we provide a BIC algorithm $\mathcal{A}^{\text{IC}}$ which internally uses $\mathcal{A}$ as a "black-box". That is, $\mathcal{A}^{\text{IC}}$ simulates a run of $\mathcal{A}$, providing inputs and

---

[4] More precisely, this is the best regret bound that can be achieved for all problem instances with gap Δ.

[5] For multi-armed bandit algorithms with fixed exploration schedules, regret can be no better than $\Omega(m^{1/3} T^{2/3})$ in the worst case over all problem instances, and (essentially) no better than $O(\Delta T^{\Omega(1)})$ in the worst case over all problem instances with gap Δ [Babaioff et al. 2014].

recording the respective outputs, but does not depend on the internal workings of $\mathcal{A}$. In addition to recommending an action, the original algorithm $\mathcal{A}$ can also output a *prediction* after each round (visible only to the planner), e.g. the predicted best action; then $\mathcal{A}^{\text{IC}}$ outputs a prediction, too. A reduction such as ours allows a modular design: one can design a non-BIC algorithm (or take an existing one), and then use the reduction to inject incentive-compatibility. Modular designs are very desirable in complex economic systems, especially for settings such as MAB with a rich body of existing work.

Our reduction incurs only a small loss in performance, which can be quantified in several ways. In terms of Bayesian regret, the performance of $\mathcal{A}^{\text{IC}}$ worsens by at most a constant multiplicative factor that only depends on the prior. In terms of the average rewards, we guarantee the following: for any duration T, the average Bayesian-expected reward of $\mathcal{A}^{\text{IC}}$ between rounds $c_{\mathcal{P}}$ and $c_{\mathcal{P}} + T$ is at least that of the first $T/L_{\mathcal{P}}$ rounds in the original algorithm $\mathcal{A}$; here $c_{\mathcal{P}}$ and $L_{\mathcal{P}}$ are prior-dependent constants.[6] Finally, if $\mathcal{A}$ outputs a prediction ϕ_t after each round t, then $\mathcal{A}^{\text{IC}}$ learns as fast as $\mathcal{A}$, up to a prior-dependent constant factor $c_{\mathcal{P}}$: for every realization of the prior, its prediction in round t has the same distribution as $\phi_{\lfloor t/c_{\mathcal{P}} \rfloor}$.[7]

The black-box reduction has several benefits other than "modular design". Most immediately, one can plug in an MAB algorithm that takes into account the Bayesian prior or any other auxiliary information that a planner might have. Moreover, one may wish to implement a particular approach to exploration, e.g., incorporate some constraints on the losses, or preferences about which arms to favor or to throttle. Further, the planner may wish to predict things other than the best action. To take a very stark example, the planner may wish to learn what are the *worst* actions (in order to eliminate these actions later by other means such as legislation). While the agents would not normally dwell on low-performing actions, our reduction would then incentivize them to explore these actions in detail.

Beyond BIC bandit exploration. Our black-box reduction supports much richer scenarios than BIC *bandit* exploration. Most importantly, it allows for agent heterogeneity, as expressed by observable signals. We adopt the framework of *contextual bandits*, well-established in the machine learning literature (see Section 2 for citations). In particular, each agent is characterized by a signal, called *context*, observable by both the agent and the planner before the planner issues the recommendation. The context can include demographics, tastes, preferences and other agent-specific information. It impacts the expected rewards received by this agent, as well as the agent's beliefs about these rewards. Rather than choose the best action, the planner now wishes to optimize a *policy* that maps contexts to actions. This type of agent heterogeneity is practically important: for example, websites that issue recommendations may possess a huge amount of information about their customers, and routinely use this "context" to adjust their recommendations (e.g., Amazon and Netflix). Our reduction turns an arbitrary contextual bandit algorithm into a BIC one, with performance guarantees similar to those for the non-contextual version. From the selection bias perspective, this extension distinguishes the agents based on their observed attributes (*i.e.,* those included in the context), but protects against the selection bias that arises due to the *unobserved* attributes.

Moreover, the reduction allows learning algorithms to incorporate arbitrary *auxiliary feedback* that agents' actions may reveal. For example, a restaurant review may contain not only the overall evaluation of the agent's experience (*i.e.,* her reward), but also reveal her culinary preferences, which in turn may shed light on the popularity of other restaurants (*i.e.,* on the expected rewards of other actions). Further, an action can consist of multiple "sub-actions", perhaps under common constraints, and the auxiliary feedback may reveal the reward for each sub-action. For instance,

---

[6] As in (1.1), the prior-dependent constants can get arbitrarily large depending on the prior, even for $m = 2$ arms, and can scale exponentially with m in some examples.

[7] So if the original algorithm $\mathcal{A}$ gives an asymptotically optimal error rate as a function of t, compared to the "correct" prediction, then so does the transformed algorithm $\mathcal{A}^{\text{IC}}$, up to a prior-dependent multiplicative factor.

a detailed restaurant recommendation may include suggestions for each course, and a review may contain evaluations thereof. Such problems (without incentive constraints) have been actively studied in machine learning, under the names "MAB with partial monitoring" and "combinatorial semi-bandits"; see Section 2 for relevant citations.

We allow for scenarios when the planner wishes to optimize his own utility which is misaligned with the agents'. Then rewards in our model still correspond to the agents' utilities, and planner's utility is observed by the algorithm as auxiliary feedback. For example, a vendor who recommends products to customers may favor more expensive products or products that are tied in with his other offerings. In a different setting, a planner may prefer *less* expensive options: in a medical trial with substantial treatment costs, patients (who are getting these treatments for free) are only interested in their respective health outcomes, whereas a socially responsible planner may also factor in the treatment costs. Another example is a medical trial of several available immunizations for the same contagious disease, potentially offering different tradeoffs between the strength and duration of immunity and the severity of side effects. Hoping to free-ride on the immunity of others, a patient may assign a lower utility to a successful outcome than the government, and therefore prefer safer but less efficient options.

A black-box reduction such as ours is particularly desirable for the extended setting described above, essentially because it is not tied up to a particular variant of the problem. Indeed, contextual bandit algorithms in the literature heavily depend on the class of policies to optimize over, whereas our reduction does not. Likewise, algorithms for bandits with auxiliary feedback heavily depend on the particular kind of feedback.

1.3. Our techniques and assumptions

An essential challenge in BIC exploration is to incentivize agents to explore actions that appear suboptimal according to the agent's prior and/or the information currently available to the planner. The desirable incentives are created due to *information asymmetry*: the planner knows more than the agents do, and the recommendation reveals a carefully calibrated amount of additional information. The agent's beliefs are then updated so that the recommended action now seems preferable to others, even though the algorithm may in fact be exploring in this round, and/or the prior mean reward of this action may be small.

Our problem is hopeless for some priors. For a simple example, consider a prior on two actions (whose expected rewards are denoted μ_1 and μ_2) such that $\mathbb{E}[\mu_1] > \mathbb{E}[\mu_2]$ and μ_1 is statistically independent from $\mu_1 - \mu_2$. Then, since no amount of samples from action 1 has any bearing on $\mu_1 - \mu_2$, a BIC algorithm cannot possibly incentivize agents to try action 2. To rule out such pathological examples, we make some assumptions. Our detail-free result assumes that the prior is independent across actions, and additionally posits minor restrictions in terms of bounded rewards and full support. The black-box reduction posits a more abstract condition which allows for correlated priors, and includes independent priors as a special case (with similar minor restrictions). This condition is necessary for the case of two arms, if the algorithm is "strongly BIC", *i.e.,* if each agent's utility is strictly maximized by the recommended action.

Our algorithms are based on (versions of) a common building block: an algorithm that incentivizes agents to explore at least once during a relatively short time interval (a "phase"). The idea is to hide one round of exploration among many rounds of exploitation. An agent receiving a recommendation does not know whether this recommendation corresponds to exploration or to exploitation. However, the agents' Bayesian posterior favors the recommended action because the exploitation is much more likely. Information asymmetry arises because the agent cannot observe the previous rounds and the algorithm's randomness.

To obtain BIC algorithms with good performance, we overcome a number of technical challenges, some of which are listed below. First, an algorithm needs to convince an agent not to switch to several other actions: essentially, all actions with larger prior mean reward than the recommended action. In particular, the algorithm should accumulate sufficiently many samples of these actions

beforehand. Second, we ensure that phase length — *i.e.,* the sufficient size of exploitation pool — does not need to grow over time. In particular, it helps not to reveal any information to future agents (e.g., after each phase or at other "checkpoints" throughout the algorithm). Third, for the black-box reduction we ensure that the choices of the bandit algorithm that we reduce from do not reveal any information about rewards. In particular, this consideration was essential in formulating the main assumption in the analysis (Property (P2) on page 20). Fourth, the detail-free algorithm cannot use Bayesian inference, and relies on sample average rewards to make conclusions about Bayesian posterior rewards, even though the latter is only an approximation for the former.

1.4. Further discussion on medical trials

We view patients' incentives as one of the major obstacles that inhibit medical trials in practice, or prevent some of them from happening altogether. This obstacle may be particularly damaging for large-scale trials that concern wide-spread medical conditions with relatively inexpensive treatments. Then finding suitable patients and providing them with appropriate treatments would be fairly realistic, but incentivizing patients to participate in sufficient numbers may be challenging. BIC exploration is thus a theoretical (and so far, highly idealized) attempt to mitigate this obstacle.

Medical trials have been one of the original motivations for studying MAB and exploration-exploitation tradeoff [Thompson 1933, Gittins 1979]. Bandit-like designs for medical trials belong to the realm of *adaptive* medical trials [Chow and Chang 2008], which can also include other "adaptive" features such as early stopping, sample size re-estimation, and changing the dosage.

"Message-restricted" algorithms (which recommend particular treatments to patients and reveal no other information) are appropriate for this domain. Revealing some (but not all) information about the medical trial is required to meet the standards of "informed consent", as prescribed by various guidelines and regulations [see Arango et al. 2012, for background]. However, revealing information about clinical outcomes in an ongoing trial is currently not required, to the best of our understanding. In fact, revealing such information is seen as a significant threat to the statistical validity of the trial (because both patients and doctors may become biased in favor of better-performing treatments), and care is advised to prevent information leaks as much as possible [see Detry et al. 2012, for background and discussion, particularly pp. 26-30].

Medical trials provide additional motivation for BIC bandit exploration with multiple actions. While traditional medical trials compare a new treatment against the placebo or a standard treatment, designs of medical trials with multiple treatments have been studied in the biostatistics literature [e.g., see Hellmich 2001, Freidlin et al. 2008], and are becoming increasingly important in practice [Parmar et al. 2014, Redig and Jänne 2015]. Note that even for the placebo or the standard treatment the expected reward is often not known in advance, as it may depend on the particular patient population.

BIC *contextual* bandit exploration is particularly relevant to medical trials, as patients come with a lot of "context" which can be used to adjust and personalize the recommended treatment. The context can include age, fitness levels, race or ethnicity, various aspects of the patient's medical history, as well as genetic markers (increasingly so as genetic sequencing is becoming more available). Context-dependent treatments (known as *personalized medicine*) has been an important trend in the pharmaceutical industry in the past decade, especially genetic-marker-dependent treatments in oncology [e.g., see Maitland and Schilsky 2011, Garimella 2015, Vaidyanathan 2012]. Medical trials for context-dependent treatments are more complex, as they must take the context into account. To reduce costs and address patient scarcity, a number of novel designs for such trials have been deployed [e.g., see Maitland and Schilsky 2011, Garimella 2015]. Some of the deployed designs are explicitly "contextual", in that they seek the best policy — mapping from patient's context to treatment [Redig and Jänne 2015]. More advanced "contextual" designs have been studied in biostatistics [e.g., see Freidlin and Simon 2005, Freidlin et al. 2007, 2010].

2. Related work

Exploration, exploitation, and incentives. There is a growing literature about a three-way interplay of exploration, exploitation, and incentives, comprising a variety of scenarios.

The study of mechanisms to incentivize exploration has been initiated by Kremer et al. [2014]. They mainly focus on deriving the Bayesian-optimal policy for the case of only two actions and deterministic rewards, and only obtain a preliminary result for stochastic rewards; a detailed comparison is provided below. Bimpikis et al. [2018][8] consider a similar model with time-discounted rewards, focusing on the case of two arms. If expected rewards are known for one arm, they provide a BIC algorithm that achieves the "first-best" utility. For the general case, they design an optimal BIC algorithm that is computationally inefficient, and propose a tractable heuristic based on the same techniques. Motivated by similar applications in the Internet economy, Che and Hörner [2018] propose a model with a continuous information flow and a continuum of consumers arriving to a recommendation system and derive a Bayesian-optimal incentive-compatible policy. Their model is technically different from ours, and is restricted to two arms and binary rewards. Frazier et al. [2014] consider a similar setting with monetary transfers, where the planner not only recommends an action to each agent, but also offers a payment for taking this action. In their setting, incentives are created via the offered payments rather than via information asymmetry.

Exploration-exploitation problems with self-interested agents have been studied in several other scenarios: multiple agents engaging in exploration and benefiting from exploration performed by others, without a planner to coordinate them [*e.g.,* Bolton and Harris 1999, Keller et al. 2005] dynamic pricing with model uncertainty [*e.g.,* Kleinberg and Leighton 2003, Besbes and Zeevi 2009, Badanidiyuru et al. 2018], dynamic auctions [*e.g.,* Athey and Segal 2013, Bergemann and Välimäki 2010, Kakade et al. 2013], pay-per-click ad auctions with unknown click probabilities [*e.g.,* Babaioff et al. 2014, Devanur and Kakade 2009, Babaioff et al. 2015b], as well as human computation [*e.g.,* Ho et al. 2016, Ghosh and Hummel 2013, Singla and Krause 2013]. In particular, a black-box reduction from an arbitrary MAB algorithm to an incentive-compatible algorithm is the main result in Babaioff et al. [2015b], in the setting of pay-per-click ad auctions with unknown click probabilities. The technical details in all this work (who are the agents, what are the actions, etc.) are very different from ours, so a direct comparison of results is uninformative.

Subsequent work on incentivized exploration. Several papers appeared subsequent to the conference publication of this paper. Kleinberg et al. [2016] consider costly information acquisition by self-interested agents (*e.g.,* real estate buyers or venture capitalists investigating a potential purchase/invenstment), and design a BIC mechanism to coordinate this process so as to improve social welfare. Bahar et al. [2016] extend the setting in Kremer et al. [2014] — BIC exploration with two actions and deterministic rewards — to a known social network on the agents, where each agent can observe friend's recommendations in the previous rounds (but not their rewards). Mansour et al. [2016] extend BIC exploration to scenarios when multiple agents arrive in each round and their chosen actions may affect others' rewards. The paradigmatic motivating example is driving directions on *Waze*, where drivers' choice of routes can create congestion for other drivers. The mechanism's recommendation is correlated across agents. Formally, it is a randomized mapping from history to joint actions which must be incentive-compatible in the sense of Bayesian correlated equilibrium. Chen et al. [2018], Immorlica et al. [2019] investigated scenarios with heterogenous agents. Finally, Kannan et al. [2017] focuses on incentivizing *fair* exploration, in the sense that worse arms are never (probabilistically) preferred to better ones.

Mansour et al. [2016], Immorlica et al. [2019] take a different conceptual approach which is complementary to ours: rather than making assumptions to ensure that all agents are "explorable" (*i.e.,* can be eventually chosen by an incentive-compatible algorithm), they start with no assumptions, and strive to explore all actions that can possibly be explored.

---

[8] Bimpikis et al. [2018] is concurrent and independent work with respect to the conference publication of this paper.

A closely related line of work [Bastani et al. 2018, Kannan et al. 2018, Schmit and Riquelme 2018, Raghavan et al. 2018] considers the "full revelation" scenario, when each agent sees the full history and chooses an action based on her own incentives. This corresponds to the "greedy algorithm" in multi-armed bandits. While the greedy algorithm is known to suffer linear regret in the worst case, it performs well under some (fairly strong) assumptions on the diversity of agent population.

Information design. As mentioned in the Introduction, a single round of our model is as a version of the *Bayesian Persuasion* game [Kamenica and Gentzkow 2011] in which the planner's signal is the "history" of the previous rounds. Rayo and Segal [2010] examine a related setting in which a planner incentivizes agents to make better choices via information disclosure. Ely et al. [2015] and Hörner and Skrzypacz [2015] consider information disclosure over time, in very different models: resp., releasing news over time to optimize suspense and surprise, and selling information over time. More background on the "design of information structures" in theoretical economics can be found in [Bergemann and Morris 2016, Taneva 2016].

Detailed comparison to Kremer, Mansour, and Perry [2014]. While the expected reward is determined by the chosen action, we allow the realized reward to be stochastic. Kremer et al. [2014] mainly focus on deterministic rewards, and only obtain a preliminary result for stochastic rewards. We improve over the latter result in several ways.

Kremer et al. [2014] only consider the case of two actions, whereas we handle an arbitrary constant number of actions. Handling more than two actions is important in practice, because recommendation systems for products or experiences are rarely faced with only binary decisions. Further, medical trials with multiple treatments are important, too [Hellmich 2001, Freidlin et al. 2008, Parmar et al. 2014]. For multiple actions, convergence to the best action is a new and interesting result on its own, even regardless of the rate of convergence. Our extension to more than two actions is technically challenging, requiring several new ideas compared to the two-action case; especially so for the detail-free version.

We implement *adaptive exploration*, when the exploration schedule is adapted to the previous observations, whereas the algorithm in Kremer et al. [2014] is a BIC implementation of *fixed exploration* — a "naive" MAB algorithm in which the exploration schedule is fixed in advance. This leads to a stark difference in ex-post regret. To describe these improvements, let us define the *MAB instance* as a mapping from actions to their expected rewards, and let us be more explicit about the asymptotic ex-post regret rates as a function of the time horizon T (i.e., the number of agents). Kremer et al. [2014] only provides regret bound of $\tilde{O}(T^{2/3})$ for all MAB instances,[9] whereas our algorithm achieves ex-post regret $\tilde{O}(\sqrt{T})$ for all MAB instances, and polylog(T) for MAB instances with constant "gap" in the expected reward between the best and the second-best action. The literature on MAB considers this a significant improvement (more on this below). In particular, the polylog(T) result is important in two ways: it quantifies the advantage of "nice" MAB instances over the worst case, and of IID rewards over adversarially chosen rewards.[10] The sub-par regret rate in Kremer et al. [2014] is indeed a consequence of fixed exploration: such algorithms cannot have worst-case regret better than $\Omega(T^{2/3})$, and cannot achieve polylog(T) regret for MAB instances with constant gap [Babaioff et al. 2014].

In terms of information structure, the algorithm in Kremer et al. [2014] requires all agents to have the same prior, and requires a very detailed knowledge of that prior; both are significant impediments in practice. Whereas our detail-free result allows the planner to have only a very limited knowledge of the prior, and allows the agents to have different priors.

Finally, Kremer et al. [2014] does not provide an analog of our black-box reduction, and does not handle the various generalizations of BIC bandit exploration that the reduction supports.

---

[9] Here and elsewhere, the $\tilde{O}(\cdot)$ notation hides polylog(T) factors.

[10] The MAB problem with adversarially chosen rewards only admits $\tilde{O}(\sqrt{T})$ ex-post regret in the worst case.

Multi-armed bandits. Multi-armed bandits (MAB) have been studied extensively in Economics, Operations Research and Computer Science since [Thompson 1933]. Motivations and applications range from medical trials to pricing and inventory optimization to driving directions to online advertising to human computation. A reader may refer to [Bubeck and Cesa-Bianchi 2012] and [Gittins et al. 2011] for background on regret-minimizing and Bayesian formulations, respectively. Further background on related machine learning problems can be found in Cesa-Bianchi and Lugosi [2006]. Our results are primarily related to regret-minimizing MAB formulations with IID rewards [Lai and Robbins 1985, Auer et al. 2002a].

Our detail-free algorithm builds on *Hoeffding races* [Maron and Moore 1993, 1997], a well-known technique in reinforcement learning. Its incarnation in the context of MAB is also known as *active arms elimination* [Even-Dar et al. 2006].

The general setting for the black-box reduction is closely related to three prominent directions in the work on MAB: contextual bandits, MAB with budgeted exploration, and MAB with partial monitoring. Contextual bandits have been introduced, under various names and models, in [Woodroofe 1979, Auer et al. 2002b, Wang et al. 2005, Langford and Zhang 2007], and actively studied since then. We follow the formulation from [Langford and Zhang 2007] and a long line of subsequent work [*e.g.,* Dudík et al. 2011, Agarwal et al. 2014]. In MAB with budgeted exploration, algorithm's goal is to predict the best arm in a given number of rounds, and performance is measured via prediction quality rather than cumulative reward [Even-Dar et al. 2002, Mannor and Tsitsiklis 2004, Guha and Munagala 2007, Goel et al. 2009, Bubeck et al. 2011]. In MAB with partial monitoring, auxiliary feedback is revealed in each round along with the reward. A historical account for this direction can be found in Cesa-Bianchi and Lugosi [2006]; see Audibert and Bubeck [2010], Bartók et al. [2014] for examples of more recent progress. Important special cases are MAB with graph-structured feedback [Alon et al. 2013, 2015], where choosing an arm would also bring feedback for adjacent arms, and "combinatorial semi-bandits" [György et al. 2007, Kale et al. 2010, Audibert et al. 2011, Wen et al. 2015], where in each round the algorithm chooses a subset from some fixed ground set, and the reward for each chosen element of this set is revealed.

Improving the asymptotic regret rate from $\tilde{O}(T^{\gamma})$, $\gamma > \frac{1}{2}$ to $\tilde{O}(\sqrt{T})$ and from $\tilde{O}(\sqrt{T})$ to polylog(T) has been a dominant theme in the literature on regret-minimizing MAB (a through survey is beyond our scope, see Bubeck and Cesa-Bianchi [2012] for background and citations). In particular, the improvement from $\tilde{O}(T^{2/3})$ to $\tilde{O}(\sqrt{T})$ regret due to the distinction between fixed and adaptive exploration has been a major theme in [Babaioff et al. 2014, Devanur and Kakade 2009, Babaioff et al. 2015a].

As much of our motivation comes from human computation, we note that MAB-like problems have been studied in several other setting motivated by human computation. Most of this work has focused on crowdsourcing markets, addressing topics such as task-worker matching and pricing decisions; see Slivkins and Vaughan [2013] for a survey and discussion.

Fairness in ML. The connection between BIC exploration and fairness, observed in Section 1.1, relates this paper to a growing body of work on fairness in machine learning [*e.g.,* Dwork et al. 2012, Hardt et al. 2016, Kleinberg et al. 2017, Chouldechova 2017]. Some of this work studies fairness in the context of multi-armed bandits [*e.g.,* Joseph et al. 2016, Liu et al. 2017, Celis and Vishnoi 2017], albeit with very different motivations and technical formulations. Most notably, [Joseph et al. 2016, Kearns et al. 2017, Kannan et al. 2017] treat arms as applicants for limited resources such as bank loans, and require that worse applicants are not preferred to better ones.

3. Model and preliminaries

We define the basic model, called *BIC bandit exploration*; the generalizations are discussed in Section 7. A sequence of T agents arrive sequentially to the planner. In each round t, the interaction

protocol is as follows: a new agent arrives, the planner sends this agent a signal σ_t, the agent chooses an action i_t in a set A of m possible actions, receives a reward for this action, and leaves forever. The signal σ_t includes a recommended action $I_t \in A$. This entire interaction is not observed by other agents. The planner knows the value of T. However, a coarse upper bound would suffice in most cases, with only a constant degradation of our results.

The planner chooses signals σ_t using an algorithm, called *recommendation algorithm*. If $\sigma_t = I_t = i_t$ (i.e., the signal is restricted to the recommended action, which is followed by the corresponding agent) then the setting reduces to multi-armed bandits (MAB), and the recommendation algorithm is a bandit algorithm. To follow the MAB terminology, we will use *arms* synonymously with *actions*; we sometimes write "play/pull an arm" rather than "choose an action".

Rewards. For each arm i there is a parametric family $\mathcal{D}_i(\cdot)$ of reward distributions, parameterized by the expected reward μ_i. The *reward vector* $\mu = (\mu_1, \ldots, \mu_m)$ is drawn from some prior distribution $\mathcal{P}^0$. Conditional on the mean μ_i, the realized reward when a given agent chooses action i is drawn independently from distribution $\mathcal{D}_i(\mu_i)$. In this paper we restrict attention to the case of single parameter families of distributions, however we do not believe this is a real restriction for our results to hold.

The prior $\mathcal{P}^0$ and the tuple $\mathcal{D} = (\mathcal{D}_i(\cdot) : i \in A)$ constitute the (full) Bayesian prior on rewards, denoted $\mathcal{P}$. It is known to all agents and to the planner.[11] The expected rewards μ_i are not known to either.

For each arm i, let $\mathcal{P}_i^0$ be the marginal of $\mathcal{P}^0$ on this arm, and let $\mu_i^0 = \mathbb{E}[\mu_i]$ be the prior mean reward. W.l.o.g., re-order the arms so that $\mu_1^0 \geq \mu_2^0 \geq \ldots \geq \mu_m^0$. The prior $\mathcal{P}$ is independent across arms if the distribution $\mathcal{P}^0$ is a product distribution, i.e., $\mathcal{P}^0 = \mathcal{P}_1^0 \times \ldots \times \mathcal{P}_m^0$.

Incentive-compatibility. Each agent t maximizes her own Bayesian expected reward, conditional on any information that is available to him. Recall that the agent observes the planner's message σ_t, and does not observe anything about the previous rounds. Therefore, the agent simply chooses an action i that maximizes the posterior mean reward $\mathbb{E}[\mu_i | \sigma_t]$. In particular, if the signal does not contain any new information about the reward vector μ, then the agent simply chooses an action i that maximizes the prior mean reward $\mathbb{E}[\mu_i]$.

We are interested in algorithms that respect incentives, in the sense that the recommended action $i = I_t$ maximizes the posterior mean reward. To state this property formally, let $\mathcal{E}_{t-1}$ be the event that the agents have followed the algorithm's recommendations up to (and not including) round t.

DEFINITION 3.1. A recommendation algorithm is *Bayesian incentive-compatible* (BIC) if

$$\mathbb{E}[\mu_i \mid \sigma_t,\, I_t = i, \mathcal{E}_{t-1}] \geq \max_{j \in A} \mathbb{E}[\mu_j \mid \sigma_t,\, I_t = i, \mathcal{E}_{t-1}] \qquad \forall t \in [T],\, \forall i \in A. \tag{3.1}$$

The algorithm is *strongly BIC* if Inequality (3.1) is always strict.

REMARK 3.1. Throughout this paper, we focus on BIC recommendation algorithms with $\sigma_t = I_t$. As observed in Kremer et al. [2014], this restriction is w.l.o.g. in the following sense. First, any recommendation algorithm can be made BIC by re-defining I_t to lie in $\arg\max_i \mathbb{E}[\mu_i | \sigma_t]$. Second, any BIC recommendation algorithm can be restricted to $\sigma_t = I_t$, preserving the BIC property. Note that the first step may require full knowledge of the prior and may be computationally expensive.

Thus, for each agent the recommended action is at least as good as any other action. For simplicity, we assume the agents break ties in favor of the recommended action. Then the agents *always* follow the recommended action.

Regret. The goal of the recommendation algorithm is to maximize the expected social welfare, i.e., the expected total reward of all agents. For BIC algorithms, this is just the total expected reward of the algorithm in the corresponding instance of MAB.

We measure algorithm's performance via the standard definitions of *regret*.

---

[11] As mentioned in Introduction, we also consider an extension to a partially known prior.

DEFINITION 3.2 (EX-POST REGRET). The *ex-post regret* of the algorithm is:

$$R_\mu(T) = T(\max_i \mu_i) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t} \mid \mu\right] \tag{3.2}$$

The *Bayesian regret* of the algorithm is:

$$R_\mathcal{P}(T) = \mathbb{E}\left[T(\max_i \mu_i) - \sum_{t=1}^{T} \mu_{I_t}\right] = \mathop{\mathbb{E}}_{\mu \sim \mathcal{P}^0}\left[R_\mu(T)\right]. \tag{3.3}$$

The ex-post regret is specific to a particular reward vector μ; in the MAB terminology, it is sometimes called *MAB instance*. In (3.2) the expectation is taken over the randomness in the realized rewards and the algorithm, whereas in (3.3) it is also taken over the prior. The last summand in (3.2) is the expected reward of the algorithm.

Ex-post regret allows to capture "nice" MAB instances. Formally, we consider the *gap* of a problem instance μ, defined as

$$\Delta = \mu^* - \max_{i:\mu_i < \mu^*} \mu_i, \quad \text{where} \quad \mu^* = \max_i \mu_i. \tag{3.4}$$

In words: the difference between the largest and the second largest expected reward. One way to characterize "nice" MAB instances is via large gap. There are MAB algorithms with regret $O(\min(\frac{1}{\Delta}\log T, \sqrt{T \log T}))$ for a constant number of arms Auer et al. [2002a].

The basic performance guarantee is expressed via Bayesian regret. Bounding the ex-post regret is particularly useful if the prior is not exactly known to the planner, or if the prior that everyone believes in may not quite the same as the true prior. In general, a ex-post regret also guards against "unlikely realizations". Besides, a bound on the ex-post regret is valid for every realization of the prior, which is reassuring for the social planner and also could take advantage of "nice" realizations such as the ones with large "gap".

Concentration inequalities. For our detail-free algorithm, we will use a well-known concentration inequality known as *Chernoff-Hoeffding Bound*. This concentration inequality, in slightly different formulations, can be found in many textbooks and surveys [e.g., Mitzenmacher and Upfal 2005]. We use a formulation from the original paper [Hoeffding 1963].

THEOREM 3.1 (**Chernoff-Hoeffding Bound**). *Consider n I.I.D. random variables $X_1 \ldots X_n$ with values in $[0,1]$. Let $X = \frac{1}{n}\sum_{i=1}^{n} X_i$ be their average, and let $\mu = \mathbb{E}[X]$. Then:*

$$\Pr\left(|X - \mu| < \delta\right) \geq 1 - 2\,e^{-2n\delta^2} \qquad \forall \delta \in (0,1]. \tag{3.5}$$

In a typical usage, we consider the high-probability event in Eq. 3.5 for a suitably chosen random variables X_i (which we call the *Chernoff event*), use the above theorem to argue that the failure probability is negligible, and proceed with the analysis conditional on the Chernoff event.

Conditional expectations. Throughout, we often use conditional expectations of the form $\mathbb{E}[A|B = b]$ and $\mathbb{E}[A|B]$ where A and B are random variables. To avoid confusion, let us emphasize that $\mathbb{E}[A|B = b]$ evaluates to a scalar, whereas $\mathbb{E}[A|B]$ is a random variable that maps values of B to the corresponding conditional expectations of A. At a high level, we typically use $\mathbb{E}[A|B = b]$ in the algorithms' specifications, and we often consider $\mathbb{E}[A|B]$ in the analysis.

4. Basic technique: sampling the inferior arm

A fundamental sub-problem in BIC bandit exploration is to incentivize agents to choose any arm $i \geq 2$ even once, as initially they would only choose arm 1. (Recall that arms are ordered according to their prior mean rewards.) We provide a simple stand-alone BIC algorithm that samples each arm at least k times, for a given k, and completes in time k times a prior-dependent constant. This algorithm is the initial stage of the black-box reduction, and (in a detail-free extension) of the detail-free algorithm.

In this section we focus on the special case of two arms, so as to provide a lucid introduction to the techniques and approaches in this paper. (Extension to many arms, which requires some additional ideas, is postponed to Section 5). We allow the common prior to be correlated across arms, under a mild assumption which we prove is necessary. For intuition, we explicitly work out the special cases of Gaussian priors and beta-binomial priors.

4.1. Restricting the prior

While consider the general case of correlated per-action priors, we need to restrict the common prior $\mathcal{P}$ so as to give our algorithms a *fighting chance*, because our problem is hopeless for some priors. As discussed in the Introduction, an easy example is a prior such that μ_1 and $\mu_1 - \mu_2$ are independent. Then samples from arm 1 have no bearing on the conditional expectation of $\mu_1 - \mu_2$, and therefore cannot possibly incentivize an agent to try arm 2.

For a "fighting chance", we assume that after seeing sufficiently many samples of arm 1 there is a positive probability that arm 2 is better. To state this property formally, we denote with S_1^k the random variable that captures the first k outcomes of arm 1, and we let

$$X^k := \mathbb{E}\left[\mu_2 - \mu_1 \mid S_1^k\right] \tag{4.1}$$

be the conditional expectation of $\mu_2 - \mu_1$ as a function S_1^k. We make the following assumption:

(P1) $\Pr\left(X^k > 0\right) > 0$ for some prior-dependent constant $k = k_{\mathcal{P}} < \infty$.

In fact, Property (P1) is "almost necessary": it is necessary for a strongly BIC algorithm.

LEMMA 4.1. *Consider an instance of BIC bandit exploration with two actions such that Property (P1) does not hold. Then a strongly BIC algorithm never plays arm 2.*

Proof. Since Property (P1) does not hold, by Borel-Cantelli Lemma we have

$$\Pr\left(X^t \leq 0\right) = 1 \quad \forall t \in \mathbb{N}. \tag{4.2}$$

We prove by induction on t that the t-th agent cannot be recommended arm 2. This is trivially true for $t = 1$. Suppose the induction hypothesis is true for some t. Consider an execution of a strongly BIC algorithm. Such algorithm cannot recommend arm 2 before round $t + 1$. Then the decision whether to recommend arm 2 in round $t + 1$ is determined by the t outcomes of action 1. In other words, the event $U = \{I_{t+1} = 2\}$ belongs to $\sigma(S^t)$, the sigma-algebra generated by X^t. Therefore,

$$\mathbb{E}[\mu_2 - \mu_1 | U] = \frac{1}{\Pr\left(U\right)} \int_U \mu_2 - \mu_1 = \mathbb{E}[X^t | U] \leq 0.$$

The last inequality holds by (4.2). So a strongly BIC algorithm cannot recommend arm 2 in round $t + 1$. This completes the induction proof. $\square$

4.2. Algorithm and analysis

We provide a simple BIC algorithm that samples both arms at least k times. The time is divided into $k + 1$ phases of L rounds each, except the first one, where L is a parameter that depends on the prior. In the first phase the algorithm recommends arm 1 to $K = \max\{L, k\}$ agents, and picks the "exploit arm" a^* as the arm with a highest posterior mean conditional on these K samples. In each subsequent phase, it picks one agent independently and uniformly at random from the L agents in this phase, and use this agent to explore arm 2. All other agents exploit using arm a^*. A more formal description is given in Algorithm 1.

ALGORITHM 1: A BIC algorithm to collect k samples of both arms.

Parameters: $k, L \in \mathbb{N}$

1 In the first $K = \max\{L, k\}$ rounds recommend arm 1 ;
2 Let $s_1^K = (r_1^1, \ldots, r_1^K)$ be the corresp. tuple of rewards;
3 Let $a^* = \arg\max_{a \in \{1,2\}} \mathbb{E}[\mu_a | s_1^K]$, breaking ties in favor of arm 1;
   for *each phase* $n = 2, \ldots, k + 1$ **do**
4   From the set P of the next L agents, pick one agent p_0 uniformly at random;
5   Every agent $p \in P - \{p_0\}$ is recommended arm a^*;
6   Player p_0 is recommended arm 2
   end

We prove that the algorithm is BIC as long as L is larger than some prior-dependent constant. The idea is that, due to information asymmetry, an agent recommended arm 2 does not know whether this is because of exploration or exploitation, but knows that the latter is much more likely. For large enough L, the expected gain from exploitation exceeds the expected loss from exploration, making the recommendation BIC.

LEMMA 4.2. *Consider BIC bandit exploration with two arms. Assume Property (P1) holds for some $k_{\mathcal{P}}$, and let $Y = X^{k_{\mathcal{P}}}$. Algorithm 1 with parameters k, L is BIC as long as*

$$L \geq \max\left(k_{\mathcal{P}}, \, 1 + \frac{\mu_1^0 - \mu_2^0}{\mathbb{E}[Y | Y > 0] \Pr(Y > 0)}\right). \tag{4.3}$$

The algorithm collects at least k samples from each arm, and completes in $kL + \max(k, L)$ rounds.

Note that the denominator in (4.3) is strictly positive by Property (P1). Indeed, the property implies that $\Pr(Y > \tau) = \delta > 0$ for some $\tau > 0$, so $\mathbb{E}[Y | Y > 0] \geq \tau\delta > 0$.

Proof. The BIC constraint is trivially satisfied in the initial phase. By a simple observation from Kremer et al. [2014], which also applies to our setting, it suffices to show that an agent is incentivized to follow a recommendation for arm 2. Focus on one phase $n \geq 2$ of the algorithm, and fix some agent p in this phase. Let $Z = \mu_2 - \mu_1$. It suffices to prove that $\mathbb{E}[Z | I_p = 2] \Pr(I_p = 2) \geq 0$, i.e., that an agent recommended arm 2 is not incentivized to switch to arm 1.

Let S_1^K be the random variable which represents the tuple of the initial samples s_1^K. Observe that the event $I_p = 2$ is uniquely determined by the initial samples S_1^K and by the random bits of the algorithm, which are independent of Z. Thus by the law of iterated expectations:

$$\mathbb{E}[Z | I_p = 2] = \mathbb{E}\left[\mathbb{E}\left[Z | S_1^K\right] | I_p = 2\right] = \mathbb{E}[X | I_p = 2],$$

where $X = X^K = \mathbb{E}[Z | S_1^K]$. There are two possible disjoint events under which agent p is recommended arm 2: either $\mathcal{E}_1 = \{X > 0\}$ or $\mathcal{E}_2 = \{X \leq 0 \text{ and } p = p_0\}$. Thus,

$$\mathbb{E}[Z | I_p = 2] \Pr(I_p = 2) = \mathbb{E}[X | I_p = 2] \Pr(I_p = 2)$$

$$= \mathbb{E}[X | \mathcal{E}_1] \Pr(\mathcal{E}_1) + \mathbb{E}[X | \mathcal{E}_2] \Pr(\mathcal{E}_2),$$

Observe that:

$$\Pr\left(\mathcal{E}_2\right) = \Pr\left(X \leq 0\right) \cdot \Pr\left(p = p_0 \mid X \leq 0\right) = \tfrac{1}{L}\Pr\left(X \leq 0\right).$$

Moreover, X is independent of the event $\{p = p_0\}$. Therefore, we get:

$$\mathbb{E}[Z|I_p = 2]\Pr\left(I_p = 2\right) = \ \mathbb{E}[X|X > 0]\Pr\left(X > 0\right) + \tfrac{1}{L}\mathbb{E}[X|X \leq 0]\Pr\left(X \leq 0\right).$$

Thus, for $\mathbb{E}[Z|I_p = 2] \geq 0$ it suffices to pick L such that:

$$L \geq \frac{-\mathbb{E}[X|X \leq 0]\Pr\left(X \leq 0\right)}{\mathbb{E}[X|X > 0]\Pr\left(X > 0\right)} = 1 - \frac{\mathbb{E}[X]}{\mathbb{E}[X|X > 0]\Pr\left(X > 0\right)}.$$

The Lemma follows because $\mathbb{E}[X] = \mathbb{E}[Z] = \mu_2^0 - \mu_1^0$ and

$$\mathbb{E}[X|X > 0]\Pr\left(X > 0\right) \geq \mathbb{E}[Y|Y > 0]\Pr\left(Y > 0\right).$$

The latter inequality follows from Lemma 4.3, as stated and proved below. Essentially, it holds because $X = X^K$ conditions on more information than $Y = X^{k_{\mathcal{P}}}$: respectively, the first K samples and the first $k_{\mathcal{P}}$ samples of arm 1, where $K \geq k_{\mathcal{P}}$. $\qquad\square$

LEMMA 4.3. *Let S and T be two random variables such that T contains more information than S, i.e. if $\sigma(S)$ and $\sigma(T)$ are the sigma algebras generated by these random variables then $\sigma(S) \subset \sigma(T)$. Let $X^S = \mathbb{E}[f(\mu)|S]$ and $X^T = \mathbb{E}[f(\mu)|T]$ for some function $f(\cdot)$ on the vector of means μ. Then:*

$$\mathbb{E}[X^T|X^T > 0]\Pr\left(X^T > 0\right) \geq \mathbb{E}[X^S|X^S > 0]\Pr\left(X^S > 0\right). \tag{4.4}$$

Proof. Since the event $\{X^S > 0\}$ is measurable in $\sigma(S)$ we have:

$$\mathbb{E}[X^S|X^S > 0] = \mathbb{E}\left[\mathbb{E}[f(\mu)|S]|X^S > 0\right] = \mathbb{E}[f(\mu)|X^S > 0].$$

Since $\sigma(S) \subset \sigma(T)$:

$$\mathbb{E}[f(\mu)|X^S > 0] = \mathbb{E}[\mathbb{E}[f(\mu)|T]|X^S > 0] = \mathbb{E}[X^T|X^S > 0].$$

Thus we can conclude:

$$\begin{aligned}
\mathbb{E}[X^S|X^S > 0]\Pr\left(X^S > 0\right) &= \mathbb{E}[X^T|X^S > 0]\Pr\left(X^S > 0\right) \\
&= \int_{\{X^S > 0\}} X^T \\
&\leq \int_{\{X^T > 0\}} X^T = \mathbb{E}[X^T|X^T > 0]\Pr\left(X^T > 0\right).
\end{aligned}$$

$\qquad\square$

4.3. Example: Gaussian priors

To give more intuition on constants $L, k_{\mathcal{P}}$ and random variable $Y = X^{k_{\mathcal{P}}}$ from Lemma 4.2, we present a concrete example where the prior on the expected reward of each arm is a Normal distribution, and the rewards are normally distributed conditional on their mean. We are interested in X^k, as defined in (4.1); we derive an explicit formula for its distribution. We will use $\mathcal{N}(\mu, \sigma^2)$ to denote a normal distribution with mean μ and variance σ^2.

EXAMPLE 4.1. Assume the prior $\mathcal{P}$ is independent across arms, $\mu_i \sim \mathcal{N}(\mu_i^0, \sigma_i^2)$ for each arm i, and the respective reward r_i^t is conditionally distributed as $\mathbb{E}[r_i^t \mid \mu_i] \sim \mathcal{N}(\mu_i, \rho_i^2)$. Then for each $k \in \mathbb{N}$

$$X^k \sim \mathcal{N}\left(\mu_2^0 - \mu_1^0, \ \frac{\sigma_1^2}{1 + \frac{\rho_1^2/\sigma_1^2}{k}}\right). \tag{4.5}$$

It follows that Property (P1) is satisfied for any $k = k_{\mathcal{P}} \geq 1$.

   Proof. Observe that $r_1^t = \mu_1 + \epsilon_1^t$, where $\epsilon_1^t \sim \mathcal{N}(0, \rho_1^2)$ and $\mu_1 = \mu_1^0 + \zeta_1$, with $\zeta_1 \sim \mathcal{N}(0, \sigma_1^2)$. Therefore,

$$
\begin{aligned}
\mathbb{E}[\mu_1 \mid S_1^k] &= \frac{\frac{1}{\sigma_1^2}\mu_1^0 + \frac{1}{\rho_1^2}\sum_{t=1}^k r_1^t}{\frac{1}{\sigma_1^2} + k \cdot \frac{1}{\rho_1^2}} \\
&= \mu_1^0 + \frac{k \cdot \frac{1}{\rho_1^2}}{\frac{1}{\sigma_1^2} + k \cdot \frac{1}{\rho_1^2}}\left(\zeta_1 + \frac{1}{k}\sum_{t=1}^k \epsilon_1^t\right) \\
&= \mu_1^0 + \frac{k \cdot \sigma_1^2}{\rho_1^2 + k \cdot \sigma_1^2}\left(\zeta_1 + \frac{1}{k}\sum_{t=1}^k \epsilon_1^t\right).
\end{aligned}
$$

Since $\zeta_1 + \frac{1}{k}\sum_{t=1}^k \epsilon_1^t \sim \mathcal{N}(0, \sigma_1^2 + \rho_1^2/k)$, it follows that $\mathbb{E}[\mu_1|S_1^k]$ is a linear transformation of a Normal distribution and therefore

$$\mathbb{E}[\mu_1 \mid S_1^k] \sim \mathcal{N}\left(\mu_1^0, \left(\frac{k \cdot \sigma_1^2}{\rho_1^2 + k \cdot \sigma_1^2}\right)^2 \cdot \left(\sigma_1^2 + \frac{\rho_1^2}{k}\right)\right) = \mathcal{N}\left(\mu_1^0, \sigma_1^2 \frac{k \cdot \sigma_1^2}{\rho_1^2 + k \cdot \sigma_1^2}\right).$$

Eq. 4.5 follows because $X^k = \mathbb{E}[\mu_2 - \mu_1 | S_1^k] = \mu_2^0 - \mathbb{E}[\mu_1|S_1^k]$. $\square$

   Observe that as $k \to \infty$, the variance of X^k increases and converges to the variance of μ_1. Intuitively, more samples can move the posterior mean further away from the prior mean μ_i^0; in the limit of many samples X^k will be distributed exactly the same as $\mu_2^0 - \mu_1$.

   While any value of the parameter $k_{\mathcal{P}}$ is sufficient for BIC, one could choose $k_{\mathcal{P}}$ to minimize the phase length L. Recall from Lemma 4.2 that the (smallest possible) value of L is inversely proportional to the quantity $\mathbb{E}[Y|Y>0]\Pr(Y>0)$. This is simply an integral over the positive region of a normally distributed random variable with negative mean, as given in Eq. 4.5. Thus, increasing $k_{\mathcal{P}}$ increases the variance of this random variable, and therefore increases the integral. For example, $k_{\mathcal{P}} = \rho_1^2/\sigma_1^2$ is a reasonable choice if $\rho_1^2 \gg \sigma_1^2$.

4.4. Example: beta-binomial priors

Our second example concerns beta-binomial priors, when the prior on the expected reward of each arm is a beta distribution, and the rewards are 0 or 1. We are interested in X^k, as defined in (4.1). We use $\texttt{Beta}(\alpha, \beta)$ to denote the beta distribution with parameters α, β; its expectation is $\frac{\alpha}{\alpha+\beta}$.

EXAMPLE 4.2. Assume there are only two arms, the prior $\mathcal{P}$ is independent across arms, $\mu_i \sim \texttt{Beta}(\alpha_i, \beta_i)$ for each arm i, and the rewards are 0 or 1. Let $\mu_i^0 = \frac{\alpha_i}{\alpha_i+\beta_i}$ be the prior mean reward of arm i. Let Z_k be the total reward of arm 1 in the first k samples. Then:
   (a) For each $\phi \in (0, \mu_2^0]$,

$$\mathbb{E}[\mu_1 \mid S_1^k] < \phi \quad \Leftrightarrow \quad \frac{Z_k}{k} < \phi - \frac{G(\phi)}{k},$$

where $G(\phi) := \alpha_1 - \phi(\alpha_1 + \beta_1) = \alpha_1(1 - \phi/\mu_1^0) > 0$.
   (b) Take $\phi = \mu_2^0$ and $k > G(\phi)/\phi$. Then $\Pr(X^k > 0) > 0$.

(c) Take $\phi = \mu_2^0/2$ and $k > 2 \cdot G(\phi)/\phi$. Then $\Pr\left(X^k \geq \phi\right) \geq \frac{1}{2} \Pr\left(\mu_1 \leq \phi/4\right)$. Therefore,

$$\mathbb{E}[X^k \mid X^k > 0] \Pr\left(X_k > 0\right) \geq \phi \cdot \Pr\left(X^k \geq \phi\right)$$

$$\geq \frac{\phi}{2} \cdot \Pr\left(\mu_1 \leq \phi/4\right).$$

Thus, Property (P1) is satisfied whenever $k > G(\mu_2^0)/\mu_2^0$ according to part (b). Part (c) provides a more quantitative perspective on Lemma 4.2 (and also on Property (P2) in the next section).

Proof. We start with a well-known property of beta-binomial priors: after observing k samples from arm 1, the posterior distribution of μ_1 is $\texttt{Beta}(\alpha_1 + Z_k, \beta_1 + k - Z_k)$. Accordingly, the posterior mean reward of this arm is

$$\mathbb{E}[\mu_1 \mid S_1^k] = \frac{\alpha_1 + Z_k}{\alpha_1 + \beta_1 + k}.$$

Part (a) is obtained by a simple rearrangement of the terms.

Part (b) easily follows from Part (a):

$$\Pr\left(X^k > 0\right) = \Pr\left(\mathbb{E}[\mu_1 \mid S_1^k] < \mu_2^0\right) = \Pr\left(Z_k/k < \epsilon\right) > 0,$$

where $\epsilon = \mu_2^0 - G(\mu_2^0)/k > 0$ (and the last inequality holds for any $\epsilon > 0$).

Part (c) follows from part (a) as follows: $X^k \geq \phi$ holds whenever $Z_k/k \leq \phi/2$, so

$$\Pr\left(X^k \geq \phi\right) \geq \Pr\left(Z_k/k \leq \phi/2\right)$$

$$\geq \Pr\left(Z_k/k \leq \phi/2 \mid \mu_1 \leq \phi/4\right) \cdot \Pr\left(\mu_1 \leq \phi/4\right)$$

$$\geq \tfrac{1}{2} \cdot \Pr\left(\mu_1 \leq \phi/4\right).$$

The last inequality follows from Markov's inequality. $\qquad\square$

5. Black-box reduction from any algorithm

We present a "black-box reduction" that uses a given bandit algorithm $\mathcal{A}$ as a black box to create a BIC algorithm $\mathcal{A}^{\texttt{IC}}$ for BIC bandit exploration, with only a minor increase in performance.

The reduction inputs two parameters k, L and works as follows. We start with a "sampling stage" which collects k samples of each arm in only a constant number of rounds. The sampling stage for two arms is already implemented in Section 4. The sampling stage for $m > 2$ arms is a non-trivial extension of the two-arms case, described below. Then we proceed with a "simulation stage", divided in phases of L rounds each. In each phase, we pick one round uniformly at random and "dedicate" it to running the original algorithm $\mathcal{A}$. We run $\mathcal{A}$ in the "dedicated" rounds only, ignoring the feedback from all other rounds. In the non-dedicated rounds, we recommend an "exploit arm": an arm with the largest posterior mean reward conditional on the samples collected in the previous phases. A formal description is given in Algorithm 2.

The sampling stage for $m > 2$ arms works as follows. In order to convince an agent to pull an arm $i > 1$, we need to convince him not to switch to any other arm $j < i$. We implement a "sequential contest" among the arms. We start with collecting samples of arm 1. The rest of the process is divided into $m - 1$ phases $2, 3, \ldots, m$, where the goal of each phase i is to collect the samples of arm i. We maintain the "exploit arm" a^*, the arm with the best posterior mean given the previously collected samples of arms $j < i$. The i-th phase collects the samples of arm i using a procedure similar to that in Section 4: k agents chosen u.a.r. are recommended arm i, and the rest are recommended the exploit arm. Then the current exploit arm is compared against arm i, and the "winner" of this contest becomes the new exploit arm. The pseudocode is in Algorithm 3.

ALGORITHM 2: Black-box reduction: the simulation stage

Input : A bandit algorithm $\mathcal{A}$; parameters $k, L \in \mathbb{N}$.
Input : A dataset $\mathcal{S}^1$ which contains k samples from each arm.

1 Split all rounds into consecutive phases of L rounds each;
  for *each phase* $n = 1, \ldots$ **do**
2   Let $a_n^* = \arg\max_{i \in A} \mathbb{E}[\mu_i | \mathcal{S}^n]$ be the "exploit arm";
3   Query algorithm $\mathcal{A}$ for its next arm selection i_n;
4   Pick one agent p from the L agents in the phase uniformly at random;
5   Every agent in the phase is recommended a_n^*, except agent p who is recommended i_n;
6   Return to algorithm $\mathcal{A}$ the reward of agent p;
7   Set $\mathcal{S}^{n+1} = \mathcal{S}^n \cup \{$all samples collected in this phase$\}$
  end

ALGORITHM 3: Black-box reduction: the sampling stage.

Parameters: $k, L \in \mathbb{N}$; the number of arms m.

1 For the first k agents recommend arm 1, let $s_1^k = \{r_1^1, \ldots, r_1^k\}$ be the tuple of rewards;
  for *each arm* $i > 1$ *in increasing lexicographic order* **do**
2   Let $a^* = \arg\max_{a \in A} \mathbb{E}[\mu_a | s_1^k, \ldots, s_{i-1}^k]$, breaking ties lexicographically;
3   From the set P of the next $L \cdot k$ agents, pick a set Q of k agents uniformly at random;
4   Every agent $p \in P - Q$ is recommended arm a^*;
5   Every agent $p \in Q$ is recommended arm i;
6   Let s_i^k be the tuple of rewards of arm i returned by agents in Q
  end

Incentive-compatibility guarantees. We make an assumption on the prior $\mathcal{P}$:

(P2) Let S_i^k be the the random variable representing k rewards of action i, and let

$$X_i^k = \min_{j \neq i} \mathbb{E}[\mu_i - \mu_j | S_1^k, \ldots, S_{i-1}^k]. \tag{5.1}$$

There exist prior-dependent constants $k_\mathcal{P}, \tau_\mathcal{P}, \rho_\mathcal{P} > 0$ such that

$$\Pr\left(X_i^k > \tau_\mathcal{P}\right) \geq \rho_\mathcal{P} \qquad \forall k > k_\mathcal{P}, i \in A.$$

Informally: any given arm i can "a posteriori" be the best arm by margin $\tau_\mathcal{P}$ with probability at least $\rho_\mathcal{P}$ after seeing sufficiently many samples of each arm $j < i$. We deliberately state this property in such an abstract manner in order to make our BIC guarantees as inclusive as possible, e.g., allow for correlated priors, and avoid diluting our main message by technicalities related to convergence properties of Bayesian estimators.

REMARK 5.1. For the special case of two arms, Property (P2) follows from Property (P1), see Section 8.1 for a proof. Recall that the latter property is necessary for any strongly BIC algorithm.

REMARK 5.2. For quantitative perspective, consider the special case of two arms and beta-binomial priors. See Example 4.2(c) for an instantiation of Property (P2) with specific parameters.

THEOREM 5.1 (BIC). *Assume Property (P2) holds with constants $k_\mathcal{P}, \tau_\mathcal{P}, \rho_\mathcal{P}$. Then the black-box reduction is BIC, for any bandit algorithm $\mathcal{A}$, if the parameters k, L are larger than some prior-dependent constant. More precisely, it suffices to take*

$$k \geq k_\mathcal{P} \text{ and } L \geq 2 + \frac{\mu_{\max}^0 - \mu_m^0}{\tau_\mathcal{P} \cdot \rho_\mathcal{P}}, \text{ where } \mu_{\max}^0 = \mathbb{E}\left[\max_{i \in A} \mu_i\right].$$

We prove Theorem 5.1 in Sections 5.3 and 5.2, separately for the two stages of the algorithm.

If the common prior $\mathcal{P}$ puts positive probability on the event that $\{\mu_i - \max_{j \neq i} \mu_j \geq \tau\}$, then Property (P2) is plausible because the conditional expectations tend to approach the true means as the number of samples grows. We elucidate this intuition via a concrete, and yet fairly general, example, where we focus on priors that are independent across arms. (See Section 8 for the proof).

LEMMA 5.1. *Property (P2) holds for a constant number of arms under the following assumptions:*
 (i)   The prior $\mathcal{P}$ is independent across arms.
 (ii) The prior on each μ_i has full support on some bounded interval $[a, b]$. [12]
 (iii) The prior mean rewards are all distinct: $\mu_1^0 > \mu_2^0 > \ldots > \mu_m^0$.
 (iv) The realized rewards are bounded (perhaps on a larger interval).

Performance guarantees. The algorithm's performance can be quantified in several ways: in terms of the total expected reward, the average expected reward, and in terms of Bayesian regret; we state guarantees for all three. Unlike many results on regret minimization, we do *not* assume that expected rewards are bounded.

THEOREM 5.2 (**Performance**). *Consider the black-box reduction with parameters k, L, applied to bandit algorithm $\mathcal{A}$. Let $c = k + Lk$ be the number of rounds in the sampling stage. Then:*
 (a) Let $\widehat{U}_{\mathcal{A}}(\tau_1, \tau_2)$ be the average Bayesian-expected reward of algorithm $\mathcal{A}$ in rounds $[\tau_1, \tau_2]$. Then

$$\widehat{U}_{\mathcal{A}^{\mathrm{IC}}}(c+1, c+\tau) \geq \widehat{U}_{\mathcal{A}}(1, \lfloor \tau/L \rfloor) \qquad \text{for any duration } \tau.$$

 (b) Let $U_{\mathcal{A}}(t)$ be the total Bayesian-expected reward of algorithm $\mathcal{A}$ in the first $\lfloor t \rfloor$ rounds. Then

$$U_{\mathcal{A}^{\mathrm{IC}}}(t) \geq L \cdot U_{\mathcal{A}}\left(\frac{t-c}{L}\right) + c \cdot \mathbb{E}\left[\min_{i \in A} \mu_i\right].$$

 (c) Let $R_{\mathcal{A}}(t)$ be the Bayesian regret of algorithm $\mathcal{A}$ in the first $\lfloor t \rfloor$ rounds.

$$R_{\mathcal{A}^{\mathrm{IC}}}(t) \leq L \cdot R_{\mathcal{A}}(t/L) + c \cdot \mathbb{E}\left[\max_{i \in A} \mu_i - \min_{i \in A} \mu_i\right].$$

In particular, if $\mu_i \in [0, 1]$ for all arms i, and the original algorithm $\mathcal{A}$ achieves asymptotically optimal Bayesian regret $O(\sqrt{t})$, then so does $\mathcal{A}^{\mathrm{IC}}$.

Proof.   Recall that a_n^* and i_n are, resp., the "exploit" arm and the original algorithm's selection in phase n. Part (a) follows because $\mathbb{E}\left[\sum_{n=1}^N \mu_{i_n}\right] = \widehat{U}_{\mathcal{A}}(1, N)$, where N is the number of phases, and by definition of the "exploit" arm we have $\mathbb{E}[\mu_{a_n^*} | \mathcal{S}^n] \geq \mathbb{E}[\mu_{i_n} | \mathcal{S}^n]$ for each phase n.

Parts (b) and (c) follow easily from part (a) by going, resp., from average rewards to total rewards, and from total rewards to regret. We have no guarantee whatsoever on the actions chosen in the first c rounds, so we assume the worst: in part (b) we use the worst possible Bayesian-expected reward $\mathbb{E}[\max_{i \in A} \mu_i]$, and in part (c) we use worst possible per-round Bayesian regret $\mathbb{E}[\max_{i \in A} \mu_i - \min_{i \in A} \mu_i]$. □

5.1. The black-box reduction with predictions

We can also allow the original algorithm $\mathcal{A}$ to output a *prediction* in each round. For our purposes, the prediction is completely abstract: it can be any element in the given set of possible predictions. It can be a predicted best arm, the estimated gap between the best and the second-best arms (in terms of their expected rewards), or any other information that the planner wishes to learn. For example, as pointed out in the Introduction, the planner may wish to learn the worst action (or

---

[12] That is, the prior on μ_i assigns a strictly positive probability to every open subinterval of $[a, b]$, for each arm i.

several worst actions, and possibly along with their respective rewards), in order to eliminate these actions via regulation and legislation.

The black-box reduction can be easily modified to output predictions, visible to the planner but not to agents. In each phase n, the reduction records the prediction made by $\mathcal{A}$, denote it by ϕ_n. In each round of the initial sampling stage, and in each round of the first phase ($n = 1$) the reduction outputs the "null" prediction. In each round of every subsequent phase $n \geq 2$, the reduction outputs ϕ_{n-1}.

Predictions are not visible to agents and therefore have no bearing on their incentives; therefore, the resulting algorithm is BIC by Theorem 5.1.

We are interested in the *sample complexity* of $\mathcal{A}^{\text{IC}}$: essentially, how the prediction quality grows over time. Such guarantee is meaningful even if the performance guarantees from Theorem 5.2 are not strong. We provide a very general guarantee: essentially, $\mathcal{A}^{\text{IC}}$ learns as fast as $\mathcal{A}$, up to a constant factor.

THEOREM 5.3. *Consider the black-box reduction with parameters k, L, and let $c = k + Lk$ be the number of rounds in the sampling stage. Then the prediction returned by algorithm $\mathcal{A}^{\text{IC}}$ in each round $t > c + L$ has the same distribution as that returned by algorithm $\mathcal{A}$ in round $\lfloor (t-c)/L \rfloor$.*

5.2. Proof of Theorem 5.1: the sampling stage

To guarantee the BIC property for the sampling stage of the algorithm, it suffices to have a somewhat weaker bound on parameter L than stated in Theorem 5.1. Namely, we will only assume

$$L \geq 1 + \frac{\mu_1^0 - \mu_m^0}{\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}}}.$$

The algorithm can be split into m phases, where each phase except the first one lasts $L \cdot k$ rounds, and agents are aware which phase they belong to. We will argue that each agent p in each phase i has no incentive not to follow the recommended arm. If an agent in phase i is recommended any arm $j \neq i$, then she knows that this is because this is the exploit arm a_i^*; so by definition of the exploit arm it is incentive-compatible for the agent to follow it. Thus it remains to argue that an agent p in phase i who is recommended arm i does not want to deviate to some other arm $j \neq i$, i.e., that $\mathbb{E}[\mu_i - \mu_j | I_p = i] \Pr(I_p = i) \geq 0$.

Denote $X_{ij} := \mathbb{E}[\mu_i - \mu_j | S_1^k, \ldots, S_{i-1}^k]$ and $X_i := X_i^k = \min_{j \neq i} X_{ij}$.

By the law of iterated expectations:

$$\mathbb{E}[\mu_i - \mu_j | I_p = i] \Pr(I_p = i) = \mathbb{E}[\mathbb{E}[\mu_i - \mu_j | S_1^k, \ldots, S_{i-1}^k] | I_p = i] \ \Pr(I_p = i)$$

$$= \mathbb{E}[X_{ij} | I_p = i] \Pr(I_p = i).$$

There are two possible disjoint events under which agent p is recommended arm i, either $\mathcal{E}_1 = \{X_i > 0\}$ or $\mathcal{E}_2 = \{X_i \leq 0 \text{ and } p \in Q\}$. Thus:

$$\mathbb{E}[\mu_i - \mu_j | I_p = i] \Pr(I_p = i) = \mathbb{E}[X_{ij} | I_p = i] \Pr(I_p = i)$$

$$= \mathbb{E}[X_{ij} | \mathcal{E}_1] \Pr(\mathcal{E}_1) + \mathbb{E}[X_{ij} | \mathcal{E}_2] \Pr(\mathcal{E}_2).$$

Observe that:

$$\Pr(\mathcal{E}_2) = \Pr(X_i \leq 0) \cdot \Pr(p \in Q \mid X_i \leq 0) = \frac{1}{L} \Pr(X_i \leq 0).$$

Moreover, X_{ij} is independent of the event $\{p \in Q\}$. Therefore, we get:

$$\mathbb{E}[\mu_i - \mu_j | I_p = i] \Pr(I_p = i) = \ \mathbb{E}[X_{ij} | X_i > 0] \Pr(X_i > 0) + \tfrac{1}{L} \mathbb{E}[X_{ij} | X_i \leq 0] \Pr(X_i \leq 0)$$

$$= \ \mathbb{E}[X_{ij} | X_i > 0] \Pr(X_i > 0) \left(1 - \tfrac{1}{L}\right) + \tfrac{1}{L} \mathbb{E}[X_{ij}]$$

$$= \ \mathbb{E}[X_{ij} | X_i > 0] \Pr(X_i > 0) \left(1 - \tfrac{1}{L}\right) + \tfrac{1}{L} \left(\mu_i^0 - \mu_j^0\right)$$

$$\geq \ \mathbb{E}[X_i | X_i > 0] \Pr(X_i > 0) \left(1 - \tfrac{1}{L}\right) + \tfrac{1}{L} \left(\mu_i^0 - \mu_j^0\right).$$

In the above, we used the facts that $\mathbb{E}[X_{ij}] = \mathbb{E}[\mu_i - \mu_j] = \mu_i^0 - \mu_j^0$ and that $X_{ij} \geq X_i$. Thus it suffices to pick L such that:

$$L \geq 1 + \frac{\mu_j^0 - \mu_i^0}{\mathbb{E}[X_i | X_i > 0] \Pr(X_i > 0)}.$$

The theorem follows by our choice of L and since $\mu_j^0 - \mu_i^0 \leq \mu_1^0 - \mu_i^0$ for all $j \in A$, and by Property (P2) we have that $\mathbb{E}[X_i | X_i > 0] \Pr(X_i > 0) \geq \tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}}$.

5.3. Proof of Theorem 5.1: the simulation stage

Instead using Property (P2) directly, we will work with a somewhat weaker corollary thereof:

(P3) Let Λ_k be a random variable representing $k_i \geq k$ samples from each action i. There exist prior-dependent constants $k_{\mathcal{P}}^* < \infty$ and $\tau_{\mathcal{P}}, \rho_{\mathcal{P}} > 0$ such that

$$\forall k \geq k_{\mathcal{P}}^* \quad \forall i \in A \qquad \Pr\left(\min_{j \in A - \{i\}} \mathbb{E}[\mu_i - \mu_j | \Lambda_k] \geq \tau_{\mathcal{P}}\right) \geq \rho_{\mathcal{P}}.$$

Property (P3) is derived in Section 8.1.

Let us consider phase n of the algorithm. We will argue that any agent p who is recommended some arm j does not want to switch to some other arm i. More formally, we assert that

$$\mathbb{E}[\mu_j - \mu_i | I_p = j] \Pr(I_p = j) \geq 0.$$

Let S^n be the random variable which represents the dataset $\mathcal{S}^n$ collected by the algorithm by the beginning of phase n. Let $X_{ji}^n = \mathbb{E}[\mu_j - \mu_i | S^n]$ and $X_j^n = \min_{i \in A/\{j\}} X_{ji}^n$. It suffices to show:

$$\mathbb{E}\left[X_j^n | I_p = j\right] \Pr(I_p = j) \geq 0.$$

Call an agent *unlucky* if it is chosen as agent p in Algorithm 2. There are two possible ways that agent p is recommended arm j: either arm j is the best posterior arm, i.e., $X_j^n > 0$, and agent i is not the unlucky agent or $X_j^n < 0$, in which case arm j cannot possibly be the exploit arm a_n^* and therefore, $I_p = j$ happens only if agent p was the unlucky one agent among the L agents of the phase to be recommended the arm of algorithm $\mathcal{A}$ and algorithm $\mathcal{A}$ recommended arm j. There are also the followings two cases: (i) $X_j^n = 0$ or (ii) $X_j^n > 0$, agent p is the unlucky agent and algorithm $\mathcal{A}$ recommends arm j. However, since we only want to lower bound the expectation and since under both these events $X_j^n \geq 0$, we will disregard these events, which can only contribute positively.

Denote with U the index of the "unlucky" agent selected in phase n to be given recommendation of the original algorithm $\mathcal{A}$. Denote with $\mathcal{A}_j^n$ the event that algorithm $\mathcal{A}$ pulls arm j at iteration n. Thus by our reasoning in the previous paragraph:

$$\mathbb{E}\left[X_j^n | I_p = j\right] \Pr(I_p = j) \geq \mathbb{E}[X_j^n | X_j^n \geq 0, U \neq p] \Pr(X_j^n \geq 0, U \neq p)$$
$$+ \mathbb{E}[X_j^n | X_j^n < 0, U = p, \mathcal{A}_j^n] \Pr(X_j^n < 0, U = p, \mathcal{A}_j^n) \qquad (5.2)$$

Since X_j^n is independent of U, we can drop conditioning on the event $\{U = p\}$ from the conditional expectations in Eq. 5.2. Further, observe that:

$$\Pr(X_j^n < 0, U = p, \mathcal{A}_j^n) = \Pr(U = p \mid X_j^n < 0, \mathcal{A}_j^n) \Pr(X_j^n < 0, \mathcal{A}_j^n)$$
$$= \tfrac{1}{L} \Pr(X_j^n < 0, \mathcal{A}_j^n)$$
$$\Pr(X_j^n \geq 0, U \neq p) = \Pr(U \neq p \mid X_j^n \geq 0) \Pr(X_j^n \geq 0)$$
$$= \left(1 - \tfrac{1}{L}\right) \Pr(X_j^n \geq 0)$$

Plugging this into Eq. 5.2, we can re-write it as follows:

$$\mathbb{E}\left[X_j^n | I_p = j\right] \Pr\left(I_p = j\right) \geq \left(1 - \tfrac{1}{L}\right) \mathbb{E}[X_j^n | X_j^n \geq 0] \Pr\left(X_j^n \geq 0\right)$$
$$+ \tfrac{1}{L} \mathbb{E}[X_j^n | X_j^n < 0, \mathcal{A}_j^n] \Pr\left(X_j^n < 0, \mathcal{A}_j^n\right). \tag{5.3}$$

Observe that conditional on $\{X_j^n < 0\}$, the integrand of the second part in the latter summation is always non-positive. Thus integrating over a bigger set of events can only decrease it, i.e.:

$$\mathbb{E}[X_j^n | X_j^n < 0, \mathcal{A}_j^n] \Pr\left(X_j^n < 0, \mathcal{A}_j^n\right) \geq \mathbb{E}[X_j^n | X_j^n < 0] \Pr\left(X_j^n < 0\right)$$

Plugging this into Eq. 5.3, we can lower bound the expected gain as:

$$\mathbb{E}\left[X_j^n | I_p = j\right] \Pr\left(I_p = j\right) \geq \left(1 - \tfrac{1}{L}\right) \mathbb{E}[X_j^n | X_j^n \geq 0] \Pr\left(X_j^n \geq 0\right) + \tfrac{1}{L} \mathbb{E}[X_j^n | X_j^n < 0] \Pr\left(X_j^n < 0\right)$$
$$= \mathbb{E}[X_j^n | X_j^n \geq 0] \Pr\left(X_j^n \geq 0\right) \tfrac{L-2}{L} + \tfrac{1}{L} \mathbb{E}[X_j^n].$$

Now observe that:

$$\mathbb{E}[X_j^n] = \mathbb{E}\left[\min_{i \in A/\{j\}} \mathbb{E}[\mu_j - \mu_i | S^n]\right] \geq \mathbb{E}\left[\mathbb{E}\left[\min_{i \in A/\{j\}} (\mu_j - \mu_i) | S^n\right]\right] = \mathbb{E}\left[\mu_j - \max_{i \in A/\{j\}} \mu_i\right]$$
$$= \mu_j^0 - \mu_{\max}^0.$$

By property (P3), we have that $\Pr\left(X_j^n \geq \tau_{\mathcal{P}}\right) \geq \rho_{\mathcal{P}}$. Therefore:

$$\mathbb{E}[X_j^n | X_j^n \geq 0] \Pr\left(X_j^n \geq 0\right) \geq \tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}} > 0.$$

Thus the expected gain is lower bounded by:

$$\mathbb{E}\left[X_j^n | I_p = j\right] \Pr\left(I_p = j\right) \geq \tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}} \frac{L-2}{L} + \frac{1}{L}(\mu_j^0 - \mu_{\max}^0).$$

The latter is non-negative if:

$$L \geq \frac{\mu_{\max}^0 - \mu_j^0}{\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}}} + 2$$

Since this must hold for every any j, we get that for incentive compatibility it suffices to pick:

$$L \geq \frac{\max_{j \in A}(\mu_{\max}^0 - \mu_j^0)}{\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}}} + 2 = \frac{\mu_{\max}^0 - \mu_m^0}{\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}}} + 2.$$

6. Ex-post regret bounds with a detail-free algorithm

We present a BIC algorithm with near-optimal bounds on ex-post regret, both for "nice" MAB instances and in the worst case over all MAB instances. Our algorithm, called DETAILFREERACE, is "detail-free", as it requires only a limited knowledge of the prior. This is an improvement over the results in the preceding section, whcih relied on the precise knowledge of the prior, and only provided performance guarantees in expectation over the prior.

6.1. Results and discussion

We use an assumption to restrict the common prior, like in the previous sections. We need this assumption not only to give the algorithm a "fighting chance", but also to ensure that the sample average of each arm is a good proxy for the respective posterior mean. Hence, we focus on priors that are independent across arms, assume bounded rewards, and full support for each μ_i.

(P4) The prior is independent across arms; rewards lie in $[0,1]$; each μ_i has full support on $[0,1]$.

However, we make no assumptions on the distribution of each μ_i and on the parameterized reward distributions $\mathcal{D}(\mu_i)$, other than the bounded support.

Our algorithm needs to know the ordering of the arms, an approximation to the smallest prior mean reward, and an upper bound on a certain prior-independent threshold.

ASSUMPTION 6.1. *Algorithm* DETAILFREERACE *knows the following:*
- *the ordering of the arms according to their prior mean rewards;*
- *the smallest prior mean reward μ_m^0, or a parameter $\hat{\mu}$ such that $\mu_m^0 \le \hat{\mu} \le 2\,\mu_m^0$;*
- *a certain threshold $N_{\mathcal{P}}$ that depends only on the prior $\mathcal{P}$, or a parameter $N \ge N_{\mathcal{P}}$.*

A loose upper bound on the threshold $N_{\mathcal{P}}$ suffices to guarantee BIC property, but hurts algorithm's regret. Since the algorithm does not know the exact prior, it avoids posterior updates, and instead estimates the expected reward of each arm with its sample average.

The high-level guarantees for DETAILFREERACE can be stated as follows:

THEOREM 6.1. *Algorithm* DETAILFREERACE *is parameterized by two positive numbers* $(\hat{\mu}, N)$.
(a) *The algorithm is BIC if Property (P4) holds, $\mu_m^0 > 0$, and Assumption 6.1 is satisfied.*
(b) *The algorithm achieves ex-post regret*

$$R(t) \le \min\left(C_N \cdot \Delta + \frac{O(m \log TN)}{\Delta},\ t\Delta \right) \le C_N + O\left(\sqrt{mt \cdot \log(TN)} \right).$$

Here Δ is the gap, defined as per (3.4), T is the time horizon, m is the number of arms, and $C_N = N + N^2(m-1)$. This regret bound holds for each round $t \le T$ and any value of the parameters.

The dependence on the prior $\mathcal{P}$ creeps in via threshsold $\mathbb{N}_{\mathcal{P}}$ and parameter $N \ge \mathbb{N}_{\mathcal{P}}$.

Due to the detail-free property we can allow agents to have different priors (subject to some assumptions). For example, all agents can start with the same prior, and receive idiosyncratic signals before the algorithm begins. Further, agents do not even need to know their own priors exactly. We state this formally as follows:

COROLLARY 6.1. *The BIC property holds even if agents have different priors, and incomplete knowledge of the said priors, as long as (i) all agents agree on the ordering of prior mean rewards (which is known to the algorithm), (ii) each agent's prior satisfies the conditions in Assumption 6.1 for the chosen parameters $(\hat{\mu}, N)$, and (iii) each agent knows that her prior satisfies (i) and (ii).*

Our algorithm consists of two stages. The *sampling stage* samples each arm a given number of times; it is a detail-free version of Algorithm 3. The racing stage is a BIC version of *Active Arms Elimination*, a bandit algorithm from Even-Dar et al. [2006]. Essentially, arms "race" against one another and drop out over time, until only one arm remains.[13]

We present a somewhat more flexible parametrization of the algorithm than in Theorem 6.1, which allows for better additive constants. In particular, instead of one parameter N we use several parameters, so that each parameter needs to be larger than a separate prior-dependent threshold in order to guarantee the BIC property. To obtain Theorem 6.1, we simply make all these parameters equal to N, and let $N_{\mathcal{P}}$ be the largest threshold.

The rest of this section is organized as follows: we present the sampling stage in Section 6.2, then the racing stage in Section 6.3, and wrap up the proof of the main theorem in Section 6.4. The special case of two arms is substantially simpler, we provide a standalone exposition in Appendix A.

---

[13] This approach is known as *Hoeffding race* in the literature, as Azuma-Hoeffding inequality is used to guarantee the high-probability selection of the best arm; see Related Work for relevant citations.

6.2. The detail-free sampling stage

To build up intuition, consider the special case of two arms. Our algorithm is similar to Algorithm 1, but chooses the "exploit" arm a^* in a different way. Essentially, we use the sample average reward $\hat{\mu}_1$ of arm 1 instead of its posterior mean reward, and compare it against the prior mean reward μ_2^0 of arm 2 (which is the same as the posterior mean of arm 2, because the common prior is independent across arms). To account for the fact that $\hat{\mu}_1$ is only an imprecise estimate of the posterior mean reward, we pick arm 2 as the exploit arm only if μ_2^0 exceeds $\hat{\mu}_1$ by a sufficient margin. The detailed pseudocode and a standalone analysis of this algorithm can be found in Appendix A.

Now we turn to the general case of $m > 2$ arms. On a very high level, the sampling stage proceeds as in Algorithm 3: after collecting some samples from arm 1, the process again is split into $m - 1$ phases such that arm i is explored in phase i. More precisely, rounds with arm i are inserted uniformly at random among many rounds in which the "exploit arm" a_i^* is chosen.

However, the definition of the exploit arm is rather tricky, essentially because it must be made "with high confidence" and cannot based on brittle comparisons. Consider a natural first attempt: define the "winner" w_i of the previous phases as an arm $j < i$ with the highest average reward, and pick the exploit arm to be either w_i or i, depending on whether the sample average reward of arm w_i is greater than the prior mean reward of arm i. Then it is not clear that a given arm, conditional on being chosen as the exploit arm, will have a higher expected reward than any other arm: it appears that no Chernoff-Hoeffding Bound argument can lead to such a conclusion. We resolve this issue by always using arm 1 as the winner of previous phases ($w_i = 1$), so that the exploit arm at phase i is either arm 1 or arm i. Further, we alter the condition to determine which of the two arms is the exploit arm: for each phase $i > 2$, this condition now involves comparisons with the previously considered arms. Namely, arm i is the exploit arm if only if the sample average reward of each arm $j \in (1, i)$ is larger than the sample average reward of arm 1 and smaller than the prior mean reward of arm i, and both inequalities hold by some margin.

The pseudocode is in Algorithm 4. The algorithm has three parameters k, L, C, whose meaning is as follows: the algorithm collects k samples from each arm, each phase lasts for kL rounds, and C is the safety margin in a rule that defines the exploit arm. Note that the algorithm completes in $k + Lk(m - 1)$ rounds.

ALGORITHM 4: The detail-free sampling stage: samples each arm k times.

Parameters: $k, L \in \mathbb{N}$ and $C \in (0, 1)$

1 For the first k agents recommend arm 1, and let $\hat{\mu}_1^k$ be their average reward;
   for *each arm $i > 1$ in increasing lexicographic order* **do**
      if $\hat{\mu}_1^k < \mu_i^0 - C$ *and* $\hat{\mu}_1^k + C < \hat{\mu}_j^k < \mu_i^0 - C$ *for all arms* $j \in (1, i)$ **then**
        | $a_i^* = i$
      else
        | $a_i^* = 1$
      end
2   From the set P of the next $L \cdot k$ agents, pick a set Q of k agents uniformly at random;
3   Every agent $p \in P - Q$ is recommended arm a_i^*;
4   Every agent $p \in Q$ is recommended arm i;
5   Let $\hat{\mu}_i^k$ be the average reward among the agents in Q;
   end

The provable guarantee for Algorithm 4 is stated as follows.

LEMMA 6.1. *Assume Property (P4). Fix $C \in (0, \mu_m^0/3)$ and consider event*

$$\mathcal{E} = \left\{ \forall j \in A - \{1, m\} : \mu_1 + \tfrac{3C}{2} \leq \mu_j \leq \mu_m^0 - \tfrac{3C}{2} \right\}.$$

Algorithm 4 is BIC if parameters k, L are large enough:

$$k \geq 8C^{-2} \log\left(\frac{16m}{C \cdot \Pr(\mathcal{E})}\right) \qquad and \qquad L \geq 1 + 2\frac{\mu_1^0 - \mu_m^0}{C \cdot \Pr(\mathcal{E})}. \tag{6.1}$$

Observe that $\Pr(\mathcal{E}) > 0$ by Property (P4), using the fact that $C < \mu_m^0/3$.

Proof. Consider agent p in phase i of the algorithm. Recall that she is recommended either arm 1 or arm i. We will prove incentive-compatibility for the two cases separately.

Let $\hat{\mathcal{E}}_i$ be the event that the exploit arm a_i^* in phase i is equal to i. By algorithm's specification,

$$\hat{\mathcal{E}}_i = \{\hat{\mu}_1^k \leq \min_{1 < j < i} \hat{\mu}_j - C \text{ and } \max_{1 \leq j < i} \hat{\mu}_j \leq \mu_i^0 - C\}.$$

Part I: Recommendation for arm i. We will first argue that an agent p who is recommended arm i does not want to switch to any other arm j. The case $j > i$ is trivial because no information about arms j or i has been collected by the algorithm (because the prior is independent across arms), and $\mu_i^0 \geq \mu_j^0$. Thus it suffices to consider $j < i$. We need to show that

$$\mathbb{E}[\mu_i^0 - \mu_j | I_p = i] \Pr(I_p = i) \geq 0.$$

Agent p is in the "explore group" Q with probability $\frac{1}{L}$ and in the "exploit group" $P - Q$ with the remaining probability $1 - \frac{1}{L}$. Conditional on being in the explore group, the expected gain from switching to arm j is $\mu_i^0 - \mu_j^0$, since that event does not imply any information about the rewards. If she is in the exploit group, then all that recommendation for arm i implies is that event $\hat{\mathcal{E}}_i$ has happened. Thus:

$$\mathbb{E}[\mu_i^0 - \mu_j | I_p = i] \Pr(I_p = i) = \mathbb{E}[\mu_i^0 - \mu_j | \hat{\mathcal{E}}_i] \Pr\left(\hat{\mathcal{E}}_i\right) \cdot \left(1 - \tfrac{1}{L}\right) + \frac{\mu_i^0 - \mu_j^0}{L}.$$

Thus for the algorithm to be BIC we need to pick a sufficiently large parameter L:

$$L \geq 1 + \frac{\mu_j^0 - \mu_i^0}{\mathbb{E}[\mu_i^0 - \mu_j | \hat{\mathcal{E}}_i] \Pr\left(\hat{\mathcal{E}}_i\right)}.$$

It remains to lower-bound the denominator in the above equation. Consider the Chernoff event:

$$\mathcal{C} = \{\forall i \in A : |\hat{\mu}_i^k - \mu_i| \leq \epsilon\}.$$

Observe that by Chernoff-Hoeffding Bound and the union bound:

$$\forall \mu_1, \ldots, \mu_m : \Pr(\mathcal{C} \mid \mu_1, \ldots, \mu_m) \geq 1 - m \cdot 2\,e^{-2\epsilon^2 k}.$$

Now observe that since $\mu_i \in [0, 1]$ and by the Chernoff-Hoeffding Bound:

$$
\begin{aligned}
\mathbb{E}[\mu_i^0 - \mu_j | \hat{\mathcal{E}}_i] \Pr\left(\hat{\mathcal{E}}_i\right) = & \ \mathbb{E}[\mu_i^0 - \mu_j | \hat{\mathcal{E}}_i, \mathcal{C}] \Pr\left(\hat{\mathcal{E}}_i, \mathcal{C}\right) + \mathbb{E}[\mu_i^0 - \mu_j | \hat{\mathcal{E}}_i, \neg\mathcal{C}] \Pr\left(\hat{\mathcal{E}}_i, \neg\mathcal{C}\right) \\
\geq & \ \mathbb{E}[\mu_i^0 - \mu_j | \hat{\mathcal{E}}_i, \mathcal{C}] \Pr\left(\hat{\mathcal{E}}_i, \mathcal{C}\right) - 2\,e^{-2\epsilon^2 k}.
\end{aligned}
$$

Observe that conditional on the events $\hat{\mathcal{E}}_i$ and $\mathcal{C}$, we know that $\mu_i^0 - \mu_j \geq C - \epsilon$, since event $\hat{\mathcal{E}}_i$ implies that $\mu_i^0 \geq \hat{\mu}_j^k + C$ and event $\mathcal{C}$ implies that $\hat{\mu}_j^k \geq \mu_j - \epsilon$. Thus we get:

$$\mathbb{E}[\mu_i^0 - \mu_j | \hat{\mathcal{E}}_i] \Pr\left(\hat{\mathcal{E}}_i\right) \geq (C - \epsilon) \Pr\left(\hat{\mathcal{E}}_i, \mathcal{C}\right) - m \cdot 2\,e^{-2\epsilon^2 k}.$$

Now let:

$$\mathcal{E}_i = \left\{ \mu_1 \leq \min_{1 < j < i} \mu_j - C - 2\epsilon \text{ and } \max_{1 \leq j < i} \mu_j \leq \mu_i^0 - C - 2\epsilon \right\},$$

i.e., $\mathcal{E}_i$ is a version of $\hat{\mathcal{E}}_i$ but on the true mean rewards, rather than the sample averages. Observe that under event $\mathcal{C}$, event $\mathcal{E}_i$ implies event $\hat{\mathcal{E}}_i$. Thus:

$$\Pr(\mathcal{E}_i, \mathcal{C}) \leq \Pr\left(\hat{\mathcal{E}}_i, \mathcal{C}\right). \tag{6.2}$$

Moreover, by Chernoff-Hoeffding Bound:

$$\Pr(\mathcal{E}_i, \mathcal{C}) = \Pr(\mathcal{C} \mid \mathcal{E}_i) \Pr(\mathcal{E}_i) \geq \left(1 - m \cdot e^{-2\epsilon^2 k}\right) \Pr(\mathcal{E}_i). \tag{6.3}$$

Hence, we conclude that:

$$\mathbb{E}[\mu_i^0 - \mu_j | \hat{\mathcal{E}}_i] \Pr\left(\hat{\mathcal{E}}_i\right) \geq (C - \epsilon)\left(1 - m \cdot e^{-2\epsilon^2 k}\right) \Pr(\mathcal{E}_i) - m \cdot 2 e^{-2\epsilon^2 k}.$$

By taking $\epsilon = \frac{C}{4}$ and $k \geq 2\epsilon^{-2} \log\left(\frac{16m}{C \cdot \Pr(\mathcal{E}_i)}\right)$ we get that:

$$\mathbb{E}[\mu_i^0 - \mu_j | \hat{\mathcal{E}}_i] \Pr\left(\hat{\mathcal{E}}_i\right) \geq \tfrac{3}{4}C\left(1 - \tfrac{1}{8}\right) \Pr(\mathcal{E}_i) - \tfrac{1}{8}C \Pr(\mathcal{E}_i) \geq \tfrac{1}{2}C \Pr(\mathcal{E}_i).$$

Thus it suffices to pick:

$$L \geq 1 + 2 \frac{\mu_j^0 - \mu_i^0}{C \Pr(\mathcal{E}_i)}.$$

In turn, since $\mu_j^0 \leq \mu_1^0$ it suffices to pick the above for $j = 1$.

Part II: Recommendation for arm 1. When agent p is recommended arm 1 she knows that she is not in the explore group, and therefore all that is implied by the recommendation is the event $\neg \mathcal{E}_i$. Thus we need to show that for any arm $j \geq 1$:

$$\mathbb{E}[\mu_1 - \mu_j | \neg \mathcal{E}_i] \Pr(\neg \mathcal{E}_i) \geq 0.$$

It suffices to consider arms $j \leq i$. Indeed, if an agent does not want to deviate to arm $j = i$ then she also does not want to deviate to any arm $j > i$, because at phase i the algorithm has not collected any information about arms $j \geq i$.

Now observe the following:

$$\begin{aligned} \mathbb{E}[\mu_1 - \mu_j | \neq \mathcal{E}_i] \Pr(\neg \mathcal{E}_i) &= \mathbb{E}[\mu_1 - \mu_j] - \mathbb{E}[\mu_1 - \mu_j | \mathcal{E}_i] \Pr(\mathcal{E}_i) \\ &= \mu_1^0 - \mu_j^0 + \mathbb{E}[\mu_j - \mu_1 | \mathcal{E}_i] \Pr(\mathcal{E}_i) \end{aligned}$$

Since, by definition $\mu_1^0 \geq \mu_j^0$, it suffices to show that for any $j \in [2, i]$:

$$\mathbb{E}[\mu_j - \mu_1 | \mathcal{E}_i] \Pr(\mathcal{E}_i) \geq 0.$$

If $j = i$, the above inequality is exactly the incentive-compatibility constraint that we showed in the first part of the proof. Thus we only need to consider arms $j \in [2, i-1]$.

Similar to the first part of the proof, we can write:

$$\mathbb{E}[\mu_j - \mu_1 | \mathcal{E}_i] \Pr(\mathcal{E}_i) \geq \mathbb{E}[\mu_j - \mu_1 | \mathcal{E}_i, \mathcal{C}] \Pr(\mathcal{E}_i, \mathcal{C}) - 2m e^{-2\epsilon^2 k}.$$

By similar arguments, since $\mathcal{E}_i$ implies that $\hat{\mu}_j^k \geq \hat{\mu}_1^k + C$ and since $\mathcal{C}$ implies that $\mu_j \geq \hat{\mu}_j^k - \epsilon$ and $\hat{\mu}_1^k \geq \mu_1 - \epsilon$, we have that the intersection of the two events implies $\mu_j \geq \mu_1 + C - 2\epsilon$. Hence:

$$\mathbb{E}[\mu_j - \mu_1 | \mathcal{E}_i] \Pr(\mathcal{E}_i) \geq (C - 2\epsilon) \Pr\left(\hat{\mathcal{E}}_i, \mathcal{C}\right) - 2me^{-2\epsilon^2 k}.$$

By Equations (6.2) and (6.3)

$$\mathbb{E}[\mu_j - \mu_1 | \mathcal{E}_i] \Pr(\mathcal{E}_i) \geq (C - 2\epsilon)\left(1 - 2me^{-2\epsilon^2 k}\right) \Pr(\mathcal{E}_i) - me^{-2\epsilon^2 k}.$$

Setting ϵ and k same way as in the first part, we have

$$\mathbb{E}[\mu_j - \mu_1 | \mathcal{E}_i] \Pr(\mathcal{E}_i) \geq \tfrac{1}{2}C\left(1 - \tfrac{1}{8}\right) \Pr(\mathcal{E}_i) - \tfrac{1}{8}C \Pr(\mathcal{E}_i) = \tfrac{5}{16}C \Pr(\mathcal{E}_i) \geq 0.$$

Wrapping up. We proved that both potential recommendations in phase i are BIC if we take $\epsilon = C/4$ and

$$k \geq 8C^{-2} \log\left(\frac{8m}{C \cdot \Pr(\mathcal{E}_i)}\right) \qquad \text{and} \qquad L \geq 1 + 2\frac{\mu_1^0 - \mu_m^0}{C \cdot \Pr(\mathcal{E}_i)}.$$

Eq. 6.1 suffices to ensure the above for all i because $\mathcal{E} = \mathcal{E}_m \subset \mathcal{E}_i$ for all phases $i < m$. $\qquad\square$

6.3. The detail-free racing stage

The detail-free racing stage is a BIC version of *Active Arms Elimination*, a bandit algorithm from Even-Dar et al. [2006]. Essentially, arms "race" against one another and drop out over time, until only one arm remains. More formally, the algorithm proceeds as follows. Each arm is labeled "active" initially, but may be permanently "deactivated" once it is found to be sub-optimal with high confidence. Only active arms are chosen. The time is partitioned into phases: in each phase the set of active arms is recomputed, and then all active arms are chosen sequentially. To guarantee incentive-compatibility, the algorithm is initialized with some samples of each arm, collected in the sampling stage. The pseudocode is in Algorithm 5.

ALGORITHM 5: The detail-free racing stage.

Input : parameters $k \in \mathbb{N}$ and $\theta \geq 1$; time horizon T.

Input : k samples of each arm i, denoted $r_i^1, \ldots, r_i^k$.

1 Initialize the set of active arms: $B = \{\text{all arms}\}$;

2 Split the remainder into consecutive phases, starting from phase $n = k$;

repeat

    // Phase n

3   Let $\hat{\mu}_i^n = \frac{1}{n}\sum_{t=1}^{n} r_i^t$ be the sample average for each arm $i \in B$;

4   Let $\hat{\mu}_*^n = \max_{i \in B} \hat{\mu}_i^n$ be the largest average reward so far;

5   Recompute $B = \left\{\text{arms } i : \hat{\mu}_*^n - \hat{\mu}_i^n \leq \sqrt{\frac{\log(T\theta)}{n}}\right\}$;

6   To the next $|B|$ agents, recommend each arm $i \in B$ sequentially in lexicographic order;

7   Let r_i^{n+1} be the reward of each arm $i \in B$ in this phase;

8     $n = n + 1$;

until $|B| = 1$;

9 For all remaining agents recommend the single arm a^* that remains in B

LEMMA 6.2 (**BIC**). *Fix an absolute constant $\tau \in (0,1)$ and let*

$$\theta_\tau = \frac{4m^2}{\tau} \Big/ \min_{i \in A} \Pr\left(\mu_i - \max_{j \neq i} \mu_j \geq \tau\right).$$

Algorithm 5 is BIC if Property (P4) holds and the parameters satisfy $\theta \geq \theta_\tau$ and $k \geq \theta_\tau^2 \log(T)$.

Proof. We will show that for any agent p in the racing stage and for any two arms $i, j \in A$:

$$\mathbb{E}[\mu_i - \mu_j | I_p = i]\Pr\left(I_p = i\right) \geq 0.$$

In fact, we will prove a stronger statement:

$$\mathbb{E}[\mu_i - \max_{j \neq i}\mu_j | I_p = i]\Pr\left(I_p = i\right) \geq 0. \tag{6.4}$$

For each pair of arms $i, j \in A$, denote $Z_{ij} = \mu_i - \mu_j$ and

$$Z_i = \mu_i - \max_{j \neq i}\mu_j = \min_{j \neq i} Z_{ij}.$$

For each phase n, let $\hat{z}_{ij}^n = \hat{\mu}_i^n - \hat{\mu}_j^n$. For simplicity of notation, we assume that samples of eliminated arms are also drawn at each phase but simply are not revealed or used by the algorithm.

Let $c_n = \sqrt{\frac{\log(T\theta)}{n}}$ be the threshold in phase n of the algorithm. Consider the "Chernoff event":

$$\mathcal{C} = \left\{\forall n \geq k, \forall i, j \in A : |Z_{ij} - \hat{z}_{ij}^n| < c_n\right\}.$$

By Chernoff-Hoeffding Bound and the union bound, for any Z, we have:

$$\Pr\left(\neg\mathcal{C}|Z\right) \leq \sum_{n \geq n^*}\sum_{i,j \in A}\Pr\left(|Z_{ij} - \hat{z}_{ij}^n| \geq c_n n\right) \leq \sum_{n \geq n^*}\sum_{i,j \in A} e^{-2n \cdot c_n^2}$$

$$\leq T \cdot m^2 \cdot e^{-2\log(T\theta)} \leq \frac{m^2}{\theta}.$$

Given the above and since $Z_i \geq -1$ we can write:

$$\mathbb{E}[Z_i | I_p = i]\Pr\left(I_p = i\right) = \ \mathbb{E}[Z_i | I_p = i, \mathcal{C}]\Pr\left(I_p = i, \mathcal{C}\right) + \mathbb{E}[Z_i | I_p = i, \neg\mathcal{C}]\Pr\left(I_p = i, \neg\mathcal{C}\right)$$

$$\geq \ \mathbb{E}[Z_i | I_p = i, \mathcal{C}]\Pr\left(I_p = i, \mathcal{C}\right) - \frac{m^2}{\theta}. \tag{6.5}$$

From here on, we will focus on the first term in Eq. 6.5. Essentially, we will assume that Chernoff-Hoeffding Bound hold deterministically.

Consider phase n. Denote $n^* = \theta_\tau^2 \log(T)$, and recall that $n \geq k \geq n^*$.

We will split the integral $\mathbb{E}[Z_i | I_p = i, \mathcal{C}]$ into cases for the value of Z_i. Observe that by the definition of c_n we have $c_n \leq c_{n^*} \leq \frac{\tau \Pr(Z_i \geq \tau)}{4}$. Hence, conditional on $\mathcal{C}$, if $Z_i \geq \tau \geq 2c_{n^*}$, then we have definitely eliminated all other arms except arm i at phase $n = k \geq n^*$, since

$$\hat{z}_{ij}^n > Z_{ij} - c_n \geq Z_i - c_n \geq c_n.$$

Moreover, if $Z_i \leq -2c_n$, then by similar reasoning, we must have already eliminated arm i, since

$$\hat{\mu}_i^n - \hat{\mu}_*^n = \min_{j \neq i}\hat{z}_{ij}^n < \min_{j \neq i}Z_{ij} + c_n = Z_i + c_n \leq -c_n.$$

Thus in that case $I_p = i$, cannot occur. Moreover, if we ignore the case when $Z_i \in [0, \tau)$, then the integral can only decrease, since the integrand is non-negative. Since we want to lower bound it, we will ignore this region. Hence:

$$\mathbb{E}[Z_i | I_p = i, \mathcal{C}] \Pr(I_p = i, \mathcal{C}) \geq \tau \Pr(\mathcal{C}, Z_i \geq \tau) - 2 \cdot c_n \Pr(\mathcal{C}, -2c_n \leq Z_i \leq 0)$$

$$\geq \tau \Pr(\mathcal{C}, Z_i \geq \tau) - 2 \cdot c_n \tag{6.6}$$

$$\geq \tau \Pr(\mathcal{C}, Z_i \geq \tau) - \frac{\tau \cdot \Pr(Z_i \geq \tau)}{2}. \tag{6.7}$$

Moreover, we can lower-bound the term $\Pr(\mathcal{C}, Z_i \geq \tau)$ using Chernoff-Hoeffding Bound:

$$\Pr(\mathcal{C}, Z_i \geq \tau) = \Pr(\mathcal{C} | Z_i \geq \tau) \Pr(Z_i \geq \tau) \geq \left(1 - \sum_{t \geq n^*} \sum_{i,j \in A} e^{-2c_n^2 t}\right) \cdot \Pr(Z_i \geq \tau)$$

$$\geq (1 - m^2/\theta) \cdot \Pr(Z_i \geq \tau) \geq \tfrac{3}{4} \Pr(Z_i \geq \tau).$$

Plugging this back into Eq. 6.7, we obtain:

$$\mathbb{E}[Z_i | I_p = i, \mathcal{C}] \Pr(I_p = i, \mathcal{C}) \geq \tfrac{\tau}{4} \cdot \Pr(\mathcal{C}, Z_i \geq \tau).$$

Plugging this back into Eq. 6.5 and using the fact that $\theta \geq \theta_\tau$, we obtain Eq. 6.4. $\qquad\square$

LEMMA 6.3 (**Regret**). *Algorithm 5 with any parameters $k \in \mathbb{N}$ and $\theta \geq 1$ achieves ex-post regret*

$$R(T) \leq \sum_{arms \ i} \frac{18 \log(T\theta)}{\mu^* - \mu_i}, \quad where \ \mu^* = \max_{i \in A} \mu_i. \tag{6.8}$$

Denoting $\Delta = \min_{i \in A} \mu^* - \mu_i$, and letting n^* be the duration of the initial sampling stage, the ex-post regret for each round t of the entire algorithm (both stages) is

$$R(t) \leq \min\left(n^*\Delta + \frac{18 \, m \, \log(T\theta)}{\Delta}, \ t\Delta\right) \leq n^* + O\left(\sqrt{m \, t \log(T\theta)}\right). \tag{6.9}$$

The "logarithmic" regret bound (6.8) is proved via standard techniques, e.g., see Even-Dar et al. [2006]. We omit the details from this version. To derive the corollary (6.9), observe that the ex-post regret of the entire algorithm (both stages) is at most that of the second stage plus Δ per each round of the first stage; alternatively, we can also upper-bound it by Δ per round.

6.4. Wrapping up: proof of Theorem 6.1

To complete the proof of Theorem 6.1, it remains to transition to the simpler parametrization, using only two parameters $\hat{\mu} \in [\mu_m^0, 2\mu_m^0]$ and $N \geq N_{\mathcal{P}}$, where $N_{\mathcal{P}}$ is a prior-dependent constant.

Let us define the parameters for both stages of the algorithm as follows. Define parameter C for the sampling stage as $C = \hat{\mu}/6$, and set the remaining parameters k^*, L for the sampling stage and k, θ for the racing stage be equal to N. To define $N_{\mathcal{P}}$, consider the thresholds for k^*, L in Lemma 6.1 with $C = \mu_m^0/6$, and the thresholds for k, θ in Lemma 6.2 with any fixed $\tau \in (0, 1)$, and let $N_{\mathcal{P}}$ be the largest of these four thresholds. This completes the algorithm's specification.

It is easy to see that both stages are BIC. (For the sampling stage, this is because $\Pr(\mathcal{E})$ is monotonically non-decreasing in parameter C.) Further, the sampling stage lasts for $f(N) = N + N^2(m - 1)$ rounds. So the regret bound in the theorem follows from that in Lemma 6.3.

7. Extensions: contextual bandits and auxiliary feedback

We now extend the black-box reduction result in two directions: contexts and auxiliary feedback. Each agent comes with a signal (*context*), observable by both the agent and the algorithm, which impacts the rewards received by this agent as well as the agent's beliefs about these rewards. Further, after the agent chooses an arm, the algorithm observes not only the agent's reward but also some auxiliary feedback. As discussed in Introduction and Related Work, each direction is well-motivated in the context of BIC exploration, and is closely related to a prominent line of work in the literature on multi-armed bandits.

7.1. Setting: BIC contextual bandit exploration with auxiliary feedback

In each round t, a new agent arrives, and the following happens:
- context $x_t \in \mathcal{X}$ is observed by both the algorithm and the agent,
- algorithm recommends action $a_t \in A$ and computes prediction ϕ_t (invisible to agents),
- reward $r_t(a_t) \in \mathbb{R}$ and auxiliary feedback $f_t(a_t) \in \mathcal{F}$ are observed by the algorithm.

The sets $\mathcal{X}, A, \mathcal{F}$ are fixed throughout, and called, resp., the *context space*, the *action space*, and the *feedback space*. Here $r_t(\cdot)$ and $f_t(\cdot)$ are functions from the action space to, resp., $\mathbb{R}$ and $\mathcal{F}$.

We assume an IID environment. Formally, the tuple $(x_t \in \mathcal{X}; \ r_t : A \to \mathbb{R}; \ f_t : A \to \mathcal{F})$ is chosen independently from some fixed distribution Ψ that is not known to the agents or the algorithm.

There is a common prior on Ψ, denoted $\mathcal{P}$, which is known to the algorithm and the agents. Specifically, we will assume the following parametric structure. The context x_t is an independent draw from some fixed distribution $\mathcal{D}_{\mathbf{x}}$ over the context space. Let $\mu_{a,x}$ denote the expected reward corresponding to context x and arm a. There is a single-parameter family $\mathcal{D}(\mu)$ of reward distributions, parameterized by the mean μ. The vector of expected rewards $\mu = (\mu_{a,x} : a \in A, x \in \mathcal{X})$ is drawn according to a prior $\mathcal{P}^0$. Conditional on μ and the context $x = x_t$, the reward $r_t(a)$ of every given arm $a = a_t$ is distributed according to $\mathcal{D}(\mu_{a,x})$. Similarly, there is a single-parameter family of distributions $\mathcal{D}_{\mathtt{fb}}(\cdot)$ over the feedback space, so that conditional on μ and the context $x = x_t$, the feedback $f_t(a)$ for every given arm $a = a_t$ is distributed according to $\mathcal{D}_{\mathtt{fb}}(\mu_{a,x})$. In what follows, we will sometimes write $\mu(a,x)$ instead of $\mu_{a,x}$.

The incentive constraint is now conditioned on the context. To state it formally, the planner's recommendation in round t is denoted with I_t^x, where $x = x_t$ is the context (recall that the recommendation can depend on the observed context). As in Definition 3.1, let $\mathcal{E}_{t-1}$ be the event that the agents have followed the algorithm's recommendations up to (and not including) round t.

DEFINITION 7.1 (CONTEXTUAL BIC). A recommendation algorithm is *Bayesian incentive-compatible* (*BIC*) if

$$\mathbb{E}[\mu_{a,x} \mid I_t^x = a, \mathcal{E}_{t-1}] \geq \max_{a' \in A} \mathbb{E}[\mu_{a',x} \mid I_t^x = a, \mathcal{E}_{t-1}] \qquad \forall t \in [T], \forall x \in \mathcal{X}, \forall a \in A.$$

Performance measures. Algorithm's performance is primarily measured in terms of *contextual regret*, defined as follows. A *policy* $\pi : \mathcal{X} \to A$ is a mapping from contexts to actions. The number of possible policies increases exponentially in the cardinality of the context space. Since in realistic applications the context space tends to be huge, learning over the space of all possible policies is usually intractable. A standard way to resolve this (following Langford and Zhang [2007]) is to explicitly restrict the class of policies. Then the algorithm is given a set Π of policies, henceforth, the *policy set*, so that the algorithm's performance is compared to the best policy in Π. More specifically, Bayesian contextual regret of algorithm $\mathcal{A}$ relative to the policy class Π in the first t rounds is defined as

$$R_{\mathcal{A},\Pi}(t) = \mathbb{E}_{\mu \sim \mathcal{P}^0} \left[t \cdot \sup_{\pi \in \Pi} \mathbb{E}_{x \in \mathcal{D}_{\mathbf{x}}} \left[\mu(\pi(x), x) \right] - \mathbb{E}_{\mathcal{A}} \left[\sum_{s=1}^{t} \mu(I_s^x, x) \right] \right]. \tag{7.1}$$

As in Section 5, we will also provide performance guarantees in terms of average rewards and in terms of the quality of predictions.

Discussion. In the line of work on contextual bandits with policy sets, a typical algorithm is in fact a family of algorithms, parameterized by an oracle that solves a particular optimization problem for a given policy class Π. The algorithm is then oblivious to how this oracle is implemented. This is a powerful approach, essentially reducing contextual bandits to machine learning on a given data set, so as to take advantage of the rich body of work on the latter. In particular, while the relevant optimization problem is NP-hard for most policy classes studied in the literature, it is often solved very efficiently in practice.

Thus, it is highly desirable to have a BIC contextual bandit algorithm which can work with different oracles for different policy classes. However, an explicit design of such algorithm would necessitate BIC implementations for particular oracles, which appears very tedious. A black-box reduction such as ours circumvents this difficulty.

A note on notation. We would like to re-order the actions according to their prior mean rewards, as we did before. However, we need to do it carefully, as this ordering is now context-specific. Let $\mu^0_{a,x} = \mathbb{E}[\mu_{a,x}]$ be the prior mean reward of arm a given context x. Denote with $\sigma(x,i) \in A$ the i-th ranked arm for x according to the prior mean rewards:

$$\mu^0_{\sigma(x,1),\,x} \geq \mu^0_{\sigma(x,2),\,x} \geq \cdots \geq \mu^0_{\sigma(x,m),\,x},$$

where m is the number of arms and ties are broken arbitrarily. The *arm-rank* of arm a given context x is the i such that $a = \sigma(x,i)$. We will use arm-ranks as an equivalent representation of the action space: choosing arm-rank i corresponds to choosing action $\sigma(x,i)$ for a given context x.

A *rank-sample* of arm-rank i is a tuple (x,a,r,f), where x is a context, $a = \sigma(x,i)$ is an arm, and r, f are, resp., the reward and auxiliary feedback received for choosing this arm for this context. Unless the x is specified explicitly, it is drawn from distribution $\mathcal{D}_{\mathtt{x}}$.

7.2. The black-box reduction and provable guarantees

Given an arbitrary algorithm $\mathcal{A}$, the black-box reduction produces an algorithm $\mathcal{A}^{\mathtt{IC}}$. It proceeds in two stages: the sampling stage and the simulation stage, which extend the respective stages for BIC bandit exploration. In particular, the sampling stage does not depend on the original algorithm $\mathcal{A}$.

The main new idea is that the sampling stage considers arm-ranks instead of arms: it proceeds in phases $i = 1, 2, \ldots$, so that in each phase i it collects samples of arm-rank i, and the exploit arm is (implicitly) the result of a contest between arm-ranks $j < i$. Essentially, this design side-steps the dependence on contexts.

Each stage of the black-box reduction has two parameters, $k, L \in \mathbb{N}$. While one can set these parameters separately for each stage, for the sake of clarity we consider a slightly suboptimal version of the reduction in which the parameters are the same for both stages. The pseudo-code is given in Algorithm 6. The algorithm maintains the "current" dataset $\mathcal{S}$ and keeps adding various rank-samples to this dataset. The "exploit arm" now depends on the current context, maximizing the posterior mean reward given the current dataset $\mathcal{S}$.

The black-box accesses the common prior via the following two oracles. First, given a context x, rank all actions according to their prior mean rewards $\mu^0_{a,x}$. Second, given a context x and a set S of rank-samples, compute the action that maximizes the posterior mean, i.e., $\arg\max_{a \in A} \mathbb{E}[\mu_{a,x}|S]$.

As before, we need to restrict the common prior $\mathcal{P}$ to guarantee incentive-compatibility. Our assumption talks about posterior mean rewards conditional on several rank-samples. For a particular context x and arm-rank i, we are interested in the smallest difference, in terms of posterior mean rewards, between the arm-rank i and any other arm. We state our assumption as follows:

ALGORITHM 6: Black-box reduction with contexts, auxiliary feedback and predictions.

Parameters: $k, L \in \mathbb{N}$; contextual bandit algorithm $\mathcal{A}$

    `// Sampling stage: obtains k samples from each arm-rank`

1   Initialize the dataset: $\mathcal{S} = \emptyset$;

2   For each context $x \in \mathcal{X}$: denote $a_x^*(\mathcal{S}) = \arg\max_{a \in A} \mathbb{E}[\mu_{a,x}|\mathcal{S}]$;

3   Recommend arm-rank 1 to the first k agents, add the rank-samples to $\mathcal{S}$;

    for *each phase $i = 2$ to m* **do**

        `// each phase lasts KL rounds`

4   From the set P of the next kL agents, pick a set Q of k agents uniformly at random;

5   For each agent $p \in P - Q$, recommend the "exploit arm" $a_{x_p}^*(\mathcal{S})$;

6   For each agent in Q, recommend arm-rank i;

7   Add to $\mathcal{S}$ the rank-samples from all agents in Q

    end

    `// Simulation stage (each phase lasts L rounds)`

    for *each phase $n = 1, 2, \ldots$* **do**

8   Pick one agent p from the set P of the next L agents uniformly at random;

9   For agent p: send context x_p to algorithm $\mathcal{A}$, get back arm a_n and prediction ϕ_n;

10   recommend arm a_n to agent p;

11   For every agent $t \in P \setminus \{p\}$: recommend the "exploit arm" $a_{x_t}^*(\mathcal{S})$;

12   For every agent $t \in P$ in phase $n \geq 2$: record prediction ϕ_{n-1};

13   Return the rank sample from agent p to algorithm $\mathcal{A}$;

14   Add to $\mathcal{S}$ the rank-samples from all agents in this phase;

    end

(P5) For a given arm-rank $i \in [m]$ and parameter $k \in \mathbb{N}$, let Λ_i^k be a random variable representing $k_i \geq k$ rank-samples of arm-rank i. Given context x, denote

$$X_{(i,j,x)}^k = \min_{\text{arms } a \neq \sigma(x,i)} \mathbb{E}\left[\mu_{\sigma(x,i),x} - \mu_{a,x} | \Lambda_1^k, \ldots, \Lambda_j^k\right].$$

There exist prior-dependent constants $k_{\mathcal{P}} < \infty$ and $\tau_{\mathcal{P}}, \rho_{\mathcal{P}} > 0$ such that

$$\Pr_{x \sim \mathcal{D}_{\mathbb{X}}}\left(X_{(i,j,x)}^k > \tau_{\mathcal{P}}\right) \geq \rho_{\mathcal{P}}$$

for any $k \geq k_{\mathcal{P}}$, any arm-rank $i \in [m]$, and $j \in \{i-1, m\}$.

The analysis is essentially the same as before. (For the sampling stage, the only difference is that the whole analysis for any single agent is done conditional on her context x and uses arm-ranks instead of arms.)

Property (P5) is used with $j = i - 1$ for the sampling stage, and with $j = m$ for the simulation stage. For each stage, the respective property is discovered naturally as the condition needed to complete the proof.

THEOREM 7.1 **(BIC)**. *Assume the common prior $\mathcal{P}$ satisfies Property (P5) with constants $(k_{\mathcal{P}}, \tau_{\mathcal{P}}, \rho_{\mathcal{P}})$. The black-box reduction is BIC (applied to any algorithm) if its parameters satisfy $k \geq k_{\mathcal{P}}$ and $L \geq L_{\mathcal{P}}$, where $L_{\mathcal{P}}$ is a finite prior-dependent constant. In particular, one can take*

$$L_{\mathcal{P}} = 1 + \sup_{x \in \mathcal{X}} \max_{a,a' \in A} \frac{\mu_{a,x}^0 - \mu_{a',x}^0}{\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}}}.$$

The performance guarantees are similar to those in Theorem 5.2.

THEOREM 7.2 **(Performance)**. *Consider the black-box reduction with parameters k, L, applied to some algorithm $\mathcal{A}$. The sampling stage then completes in $c = mLk + k$ rounds, where m is the number of arms.*

(a) Let $\widehat{U}_{\mathcal{A}}(\tau_1, \tau_2)$ be the average Bayesian-expected reward of algorithm $\mathcal{A}$ in rounds $[\tau_1, \tau_2]$. Then

$$\widehat{U}_{\mathcal{A}^{\text{IC}}}(c+1, c+\tau) \geq \widehat{U}_{\mathcal{A}}(1, \lfloor \tau/L \rfloor) \qquad \text{for any duration } \tau.$$

(c) Let $R_{\mathcal{A}, \Pi}(t)$ be the Bayesian contextual regret of algorithm $\mathcal{A}$ in the first t rounds, relative to some policy class Π. Then:

$$R_{\mathcal{A}^{\text{IC}}, \Pi}(t) \leq L \cdot R_{\mathcal{A}, \Pi}(\lfloor t/L \rfloor) + c \cdot \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathbf{X}}} \left[\max_{a \in A} \mu_{a,x} - \min_{a \in A} \mu_{a,x} \right].$$

In particular, if $\mu_{a,x} \in [0, 1]$ and the original algorithm $\mathcal{A}$ achieves asymptotically optimal Bayesian contextual regret $O(\sqrt{t \log |\Pi|})$, then so does $\mathcal{A}^{\text{IC}}$ (assuming L is a constant).

(c) The prediction returned by algorithm $\mathcal{A}^{\text{IC}}$ in each round $t > c + L$ has the same distribution as that returned by algorithm $\mathcal{A}$ in round $\lfloor (t-c)/L \rfloor$.

8. Properties of the common prior

Property (P2), our main assumption on the common prior for the black-box reduction, is stated in a rather abstract manner for the sake of generality (e.g., so as to allow correlated priors), and in order to avoid some excessive technicalities. In this section we discuss some necessary or sufficient conditions. First, we argue that *conditioning on more samples can only help*, i.e. that if a similar property holds conditioning on less samples, it will still hold conditioning on more samples. This is needed in the incentives analysis of the simulation stage, and to show the "almost necessity" for two arms. Second, we exhibit some natural sufficient conditions if the prior is independent across arms, and specifically prove Lemma 5.1.

8.1. Conditioning on more samples can only help

Fix arm i and vector $\mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{N}^m$, where m is the number of arms. Let $\Lambda_{\mathbf{a}}$ be a random variable representing the first a_j samples from each arm j. Denote

$$X_{i,\mathbf{a}} = \min_{\text{arms } j \neq i} \mathbb{E}\left[\mu_i - \mu_j | \Lambda_{\mathbf{a}} \right].$$

We are interested in the following property:

(P6) There exist prior-dependent constants $\tau_{\mathcal{P}}, \rho_{\mathcal{P}} > 0$ such that $\Pr\left(X_{i,\mathbf{a}} \geq \tau_{\mathcal{P}} \right) \geq \rho_{\mathcal{P}}$.

We argue that Property (P6) stays true if we condition on more samples.

LEMMA 8.1. *Fix arm i and consider vectors $\mathbf{a}, \mathbf{b} \in \mathbb{N}^m$ such that $a_j \leq b_j$ for all arms j. If Property (P6) holds for arm i and vector $\mathbf{a}$, then it also holds for arm i and vector $\mathbf{b}$.*

Proof. For the sake of contradiction, assume that the property for this arm holds for $\mathbf{a}$, but not for $\mathbf{b}$. Denote $X = X_{i,\mathbf{a}}$ and $Y = X_{i,\mathbf{b}}$.

Then, in particular, we have $\Pr\left(Y > 1/\ell^2 \right) < 1/\ell^2$ for each $\ell \in \mathbb{N}$. It follows that $\Pr\left(Y \leq 0 \right) = 1$. (To prove this, apply Borel-Cantelli Lemma with sets $S_\ell = \{ Y > 1/\ell^2 \}$, noting that $\sum_\ell \Pr\left(S_\ell \right)$ is finite and $\limsup_\ell S_\ell = \{ Y > 0 \}$.)

The samples in $\Lambda_{\mathbf{a}}$ are a subset of those for $\Lambda_{\mathbf{b}}$, so $\sigma(\Lambda_{\mathbf{a}}) \subset \sigma(\Lambda_{\mathbf{b}})$, and consequently $X = \mathbb{E}[Y | \Lambda_{\mathbf{a}}]$. Since $Y \leq 0$ a.s., it follows that $X \leq 0$ a.s. as well, contradicting our assumption. $\square$

Therefore, Property (P2) implies Property (P3) with the same constant $k_{\mathcal{P}}$; this is needed in the incentives analysis of the simulation stage. Also, Property (P1) implies Property (P2) with the same $k_{\mathcal{P}}$, so the latter property is (also) "almost necessary" for two arms, in the sense that it is necessary for a strongly BIC algorithm (see Lemma 4.1).

8.2. Sufficient conditions: Proof of Lemma 5.1

For each arm i, let S_i^k be a random variable that represents n samples of arm i, and let $Y_i^k = \mathbb{E}[\mu_i | S_i^k]$ be the posterior mean reward of this arm conditional on this random variable.

Since the prior is independent across arms, Eq. 5.1 in Property (P2) can be simplified as follows:

$$\Pr\left(\mu_i^0 - \max\left\{Y_1^k, \ldots, Y_{i-1}^k; \mu_{i+1}^0\right\} \geq \tau_{\mathcal{P}}\right) \geq \rho_{\mathcal{P}} \qquad \forall k > k_{\mathcal{P}}, i \in A. \tag{8.1}$$

Fix some $\tau_{\mathcal{P}}$ such that $0 < \tau_{\mathcal{P}} < \min_{j \in A}(\mu_j^0 - \mu_{j+1}^0)$. Note that the right-hand side is strictly positive because the prior mean rewards are all distinct (condition (iii) in the lemma).

Since $(\mu_i : i \in A)$ are mutually independent, and for fixed k each S_i^k, $i \in A$ is independent of everything else conditional on μ_i, we have that $(Y_i^k : i \in A)$ are mutually independent. Thus, one can rewrite Eq. 8.1 as follows:

$$\prod_{j < i} \Pr\left(Y_j^k \leq \mu_i^0 - \tau_{\mathcal{P}}\right) \geq \rho_{\mathcal{P}} \qquad \forall k > k_{\mathcal{P}}, i \in A.$$

Therefore it suffices to prove that $\tau_{\mathcal{P}}$ as fixed above, and some prior-dependent constant $k_{\mathcal{P}}$, for any two arms $j < i$ we have:

$$\Pr\left(Y_j^k \leq \mu_i^0 - \tau_{\mathcal{P}}\right) \geq q > 0 \qquad \forall k > k_{\mathcal{P}}. \tag{8.2}$$

where q is some prior-dependent constant. Observe that event $\{Y_j^k \leq \mu_i^0 - \tau_{\mathcal{P}}\}$ is implied by the event $\{Y_j^k - \mu_j \leq \epsilon\} \cap \{\mu_j \leq \mu_i^0 - \tau_{\mathcal{P}} - \epsilon\}$, where $\epsilon > 0$ is a prior-dependent constant to be fixed later. Invoking also the union bound we get:

$$\Pr\left(Y_j^k \leq \mu_i^0 - \tau_{\mathcal{P}}\right) \geq \Pr\left((Y_j^k - \mu_j \leq \epsilon) \cap (\mu_j \leq \mu_i^0 - \tau_{\mathcal{P}} - \epsilon)\right)$$
$$\geq \Pr\left(\mu_j \leq \mu_i^0 - \tau_{\mathcal{P}} - \epsilon\right) - \Pr\left(Y_j^k - \mu_j > \epsilon\right). \tag{8.3}$$

By the full support assumption (condition (ii) in the lemma), the first summand in (8.3) is bounded from below by some constant $\rho > 0$ when $\mu_i^0 - \tau_{\mathcal{P}} - \epsilon > b$. The latter is achieved when $\tau_{\mathcal{P}}, \epsilon$ are sufficiently small (because $\mu_i^0 > a$, also by the full support assumption). Fix such $\tau_{\mathcal{P}}, \epsilon$ from here on.

It suffices to bound the second summand in (8.3) from above by $\rho/2$. This follows by a convergence property (P7) which we articulate below. By this property, there exists a prior-dependent constant k^* such that $\Pr\left(Y_j^k - \mu_j > \epsilon\right) \leq \rho/2$ for any $k > k^*$. Thus, we have proved Eq. 8.2 for $k_{\mathcal{P}} = k^*$ and $q = \rho/2$. This completes the proof of the lemma.

It remains to state and prove the convergence property used above.

(P7) For each arm i, Y_i^k converges to μ_i in probability as $k \to \infty$. That is,

$$\forall i \in A \quad \forall \epsilon, \delta > 0 \quad \exists k^*(\epsilon, \delta) < \infty \quad \forall k \geq k^*(\epsilon, \delta) \quad \Pr\left(|Y_i^k - \mu_i| > \epsilon\right) \leq \delta.$$

This property follows from the convergence theorem in Doob [1949], because the realized rewards are bounded (condition (iv) in the lemma). For a more clear exposition of the aforementioned theorem, see e.g., Theorem 1 in Ghosal [1996]. There the sufficient condition is that the parameterized reward distribution $\mathcal{D}_i(\mu_i)$ of each arm i is pointwise dominated by an integrable random variable; the latter is satisfied if the realized rewards are bounded.

9. Conclusions and open questions

We have resolved the asymptotic regret rates achievable with incentive compatible exploration for a constant number of actions, as outlined in the Introduction. We provided an algorithm for BIC bandit exploration with asymptotically optimal regret (for a constant number of actions), and a general "black-box" reduction from arbitrary bandit algorithms to BIC ones that works in a very general explore-exploit setting and increases the regret by at most a constant multiplicative factor.

This paper sets the stage for future work in several directions. First, the most immediate technical questions left open by our work is whether one can achieve Bayesian regret with constants that do not depend on the prior (for a fixed m, the number of actions), and with only a polynomial dependence on m. Second, one would like to handle bandit problems with a large number of actions and a known structure on the action set.[14] This would require handling agents' priors with complex correlations across actions, which do not necessarily satisfy the properties assumed in this paper. On the other hand, algorithms for such problems typically explore only a sparse subset of the arms, so perhaps the assumptions on the prior could be relaxed. Third, the mechanism design setup could be generalized to incorporate constraints on how much information about the previous rounds must be revealed to the agents. Such constraints, arising for practical reasons in the Internet economy and for legal reasons in medical decisions, typically work against the information asymmetry, and hence make the planner's problem more difficult. Fourth, while our detail-free result does not require the algorithm to have full knowledge of the priors, ideally the planner would like to start with little or no knowledge of the priors, and elicit the necessary knowledge directly from the agents. Then the submitted information becomes subject to agent's incentives, along with the agent's choice of action. Finally, it would be tantalizing to bring the theory of BIC exploration closer to the design of large-scale medical trials (see Section 1.4 for discussion). In particular, one needs to account for heterogenous, risk-averse patients whose beliefs and risk preferences are largely unknown (and probably need to be elicited), and whose decisions may deviate from the standard economic assumptions.

Acknowledgments

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *31st Intl. Conf. on Machine Learning (ICML)*, 2014.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In *27th Advances in Neural Information Processing Systems (NIPS)*, pages 1610–1618, 2013.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *28th Conf. on Learning Theory (COLT)*, pages 23–35, 2015.

Jaime Arango, Karen Arts, Paul Braunschweiger, Gary L. Chadwick an David G. Forster, and Karen Hansen, editors. *Good Clinical Practice Guide.* CITI Program at The University of Miami, 2012. Available at www.citiprogrampublications.org.

Susan Athey and Ilya Segal. An efficient dynamic mechanism. *Econometrica*, 81(6):2463–2485, November 2013. A preliminary version has been available as a working paper since 2007.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. In *24th Conf. on Learning Theory (COLT)*, pages 107–132, 2011.

---

[14] Structured action sets have been one of the major directions in the literature on MAB in the past decade. Several structures have been studied, e.g., linearity, convexity, Lipschitzness, and combinatorial structure; see Bubeck and Cesa-Bianchi [2012] for background and references.

J.Y. Audibert and S. Bubeck. Regret Bounds and Minimax Policies under Partial Monitoring. *J. of Machine Learning Research (JMLR)*, 11:2785–2836, 2010. Preliminary version in *COLT 2009*.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b. Preliminary version in *36th IEEE FOCS*, 1995.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. *SIAM J. on Computing (SICOMP)*, 43(1):194–230, 2014. Preliminary version in *10th ACM EC*, 2009.

Moshe Babaioff, Shaddin Dughmi, Robert D. Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. *ACM Trans. on Economics and Computation*, 3(1):4, 2015a. Special issue for *13th ACM EC*, 2012.

Moshe Babaioff, Robert Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. *J. of the ACM*, 62(2):10, 2015b. Subsumes the conference papers in *ACM EC 2010* and *ACM EC 2013*.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. of the ACM*, 65(3), 2018. Preliminary version in *FOCS 2013*.

Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. Economic recommendation systems. In *16th ACM Conf. on Electronic Commerce (EC)*, 2016.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–997, 2014.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *CoRR arXiv:1704.09011*, 2018. Working paper.

Dirk Bergemann and Stephen Morris. Information design, bayesian persuasion and bayes correlated equilibrium, 2016. Working paper.

Dirk Bergemann and Juuso Välimäki. The dynamic pivot mechanism. *Econometrica*, 78(2):771–789, 2010. Preliminary versions have been available since 2006.

Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57:1407–1420, 2009.

Kostas Bimpikis, Yiangos Papanastasiou, and Nicos Savva. Crowdsourcing exploration. *Management Science*, 64:1477–1973, 2018.

Patrick Bolton and Christopher Harris. Strategic Experimentation. *Econometrica*, 67(2):349–374, 1999.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1), 2012.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure Exploration in Multi-Armed Bandit Problems. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

L. Elisa Celis and Nisheeth K Vishnoi. Fair personalization. *CoRR arXiv:1707.02260, also appeared at the Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge Univ. Press, 2006.

Yeon-Koo Che and Johannes Hörner. Optimal design for social learning. *Quarterly Journal of Economics*, 2018. Forthcoming. First published draft: 2013.

Bangrui Chen, Peter I. Frazier, and David Kempe. Incentivizing exploration by heterogeneous users. In *Conf. on Learning Theory (COLT)*, pages 798–818, 2018.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 2017.

Shein-Chung Chow and Mark Chang. Adaptive design methods in clinical trials a review. *Orphanet Journal of Rare Diseases*, 3(11):1750–1172, 2008.

Michelle A. Detry, Roger J. Lewis, Kristine R. Broglio, Jason T. Connor, Scott M. Berry, and Donald A. Berry. Standards for the design, conduct, and evaluation of adaptive randomized clinical trials. *Patient-Centered Outcomes Research Institute (PCORI) Guidance Report*, March 2012.

Nikhil Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *10th ACM Conf. on Electronic Commerce (EC)*, pages 99–106, 2009.

Joseph L. Doob. Application of the theory of martingales. *Le Calcul des Probabilités et ses Applications, Colloques Internationaux du Centre National de la Recherche Scientifique*, 13:23–27, 1949.

Miroslav Dudík, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal leanring for contextual bandits. In *27th Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2011.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conf. (ITCS)*, 2012.

Jeff Ely, Alex Frankel, and Emir Kamenica. Suspense and Surprise. *J. of Political Economy*, 2015. forthcoming.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *15th Conf. on Learning Theory (COLT)*, pages 255–270, 2002.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. of Machine Learning Research (JMLR)*, 7:1079–1105, 2006.

Peter Frazier, David Kempe, Jon M. Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *ACM Conf. on Economics and Computation (ACM EC)*, pages 5–22, 2014.

Boris Freidlin and Richard Simon. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*, 11:7872–7878, 2005.

Boris Freidlin, Wenyu Jiang, and Richard Simon. Biomarker adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *J. of the National Cancer Institute*, 99:1036–1043, 2007.

Boris Freidlin, Edward L. Korn, Robert Gray, and Alison Martin. Multi-arm clinical trials of new agents: Some design considerations. *Clinical Cancer Research*, 14:43–68, 2008.

Boris Freidlin, Wenyu Jiang, and Richard Simon. Cross-validated adaptive signature design. *Clinical Cancer Research*, 16(2), 2010.

Surabhidanhi Garimella. New oncology clinical trial designs: What works and what doesnt? *Evidence-Based Oncology*, 21(10):349–350, 2015.

Subhashis Ghosal. A review of consistency and convergence rates of posterior distribution, 1996. In "Proceedings of Varanashi Symposium in Bayesian Inference", Banaras Hindu University. Available from the author's webpage at `www4.stat.ncsu.edu/ sghosal`.

Arpita Ghosh and Patrick Hummel. Learning and incentives in user-generated content: multi-armed bandits with endogenous arms. In *Innovations in Theoretical Computer Science Conf. (ITCS)*, pages 233–246, 2013.

J. C. Gittins. Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Statist. Soc. Ser. B*, 41:148–177, 1979.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

Ashish Goel, Sanjeev Khanna, and Brad Null. The Ratio Index for Budgeted Learning, with Applications. In *20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 18–27, 2009.

Sudipta Guha and Kamesh Munagala. Approximation algorithms for budgeted learning problems. In *39th ACM Symp. on Theory of Computing (STOC)*, pages 104–113, 2007.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *J. of Machine Learning Research (JMLR)*, 8:2369–2403, 2007.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Martin Hellmich. Monitoring clinical trials with multiple arms. *Biometrics*, 57:892–898, 2001.

Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowd-sourcing markets: Bandit algorithms for repeated principal-agent problems. *J. of Artificial Intelligence Research*, 55:317–359, 2016. Preliminary version appeared in *ACM EC 2014*.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):pp. 13–30, 1963.

Johannes Hörner and Andrzej Skrzypacz. Selling Information. *J. of Political Economy*, 2015. forthcoming.

Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Steven Wu. Bayesian exploration with heterogenous agents. In *The Web Conference (formerly known as* WWW*)*, 2019.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Sham M. Kakade, Ilan Lobel, and Hamid Nazerzadeh. Optimal dynamic mechanism design and the virtual-pivot mechanism. *Operations Research*, 61(4):837–854, 2013.

Satyen Kale, Lev Reyzin, and Robert E. Schapire. Non-stochastic bandit slate problems. In *24th Advances in Neural Information Processing Systems (NIPS)*, pages 1054–1062, 2010.

Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra, and Z Steven Wu. Fairness incentives for myopic agents. In *ACM Conf. on Economics and Computation (ACM EC)*, 2017.

Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *Intl. Conf. on Machine Learning (ICML)*, 2017.

Godfrey Keller, Sven Rady, and Martin Cripps. Strategic Experimentation with Exponential Bandits. *Econometrica*, 73(1):39–68, 2005.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science Conf. (ITCS)*, 2017.

Robert D. Kleinberg and Frank T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, 2003.

Robert D. Kleinberg, Bo Waggoner, and E. Glen Weyl. Descending price optimally coordinates search. Working paper, 2016. Preliminary version in *ACM EC 2016*.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the wisdom of the crowd. *J. of Political Economy*, 122:988–1012, 2014. Preliminary version in *ACM EC 2014*.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

John Langford and Tong Zhang. The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. In *21st Advances in Neural Information Processing Systems (NIPS)*, 2007.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *CoRR arXiv:1707.01875*, 2017.

Michael L. Maitland and Richard L. Schilsky. Clinical trials in the era of personalized oncology. *CA: A Cancer Journal for Clinicians*, 61(6):365–381, 2011.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *J. of Machine Learning Research (JMLR)*, 5:623–648, 2004.

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Steven Wu. Bayesian exploration: Incentivizing exploration in Bayesian games. In *16th ACM Conf. on Economics and Computation (ACM EC)*, 2016.

Oded Maron and Andrew W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In *7th Advances in Neural Information Processing Systems (NIPS)*, pages 59–66, 1993.

Oded Maron and Andrew W. Moore. The racing algorithm: Model selection for lazy learners. *Artif. Intell. Rev.*, 11(1-5):193–225, 1997.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press, 2005.

Mahesh K.B. Parmar, James Carpenter, and Matthew R. Sydes. More multiarm randomised trials of superiority are needed. *Lancet*, 384(9940):283284, 2014.

Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. In *Conf. on Learning Theory (COLT)*, pages 1724–1738, 2018.

Luis Rayo and Ilya Segal. Optimal Information Disclosure. *J. of Political Economy*, 118(5):949–987, 2010.

Amanda Redig and Pasi Jänne. Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J. of Clinical Oncology*, 33, 2015.

Sven Schmit and Carlos Riquelme. Human interaction with recommendation systems. In *Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, pages 862–870, 2018.

Adish Singla and Andreas Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *22nd Intl. World Wide Web Conf. (WWW)*, pages 1167–1178, 2013.

Aleksandrs Slivkins and Jennifer Wortman Vaughan. Online decision making in crowdsourcing markets: Theoretical challenges. *SIGecom Exchanges*, 12(2), December 2013.

Ina Taneva. Information design, 2016. Working paper.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Gayathri Vaidyanathan. Redefining clinical trials: The age of personalized medicine. *Cell*, 148, 2012.

Chih-Chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor. Bandit problems with side observations. *IEEE Trans. on Automatic Control*, 50(3):338–355, 2005.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *32nd Intl. Conf. on Machine Learning (ICML)*, pages 1113–1122, 2015.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.*, 74 (368), 1979.

Appendix A: A simpler detail-free algorithm for two arms

The detail-free algorithm becomes substantially simpler for the special case of two arms. For the sake of clarity, we provide a standalone exposition. The sampling stage is similar to Algorithm 1, except that the exploit arm is chosen using the sample average reward of arm 1 instead of its posterior mean reward, and picking arm 2 as the exploit arm only if it appears better by a sufficient margin. The racing stage is a simple "race" between the two arms: it alternates the two arms until one of them can be eliminated with high confidence, regardless of the prior, and uses the remaining arm from then on.

Below we provide the algorithm and analysis for each stage separately. The corresponding lemmas can be used to derive Theorem 6.1 for two arms, essentially same way as in Section 6.4.

Compared to the general case of $m > 2$ arms, an additional provable guarantee is that even computing the threshold $N_{\mathcal{P}}$ exactly does not require much information: it only requires knowing the prior mean rewards of both arms, and evaluating the CDFs for μ_1 and $\mu_1 - \mu_2$ at a single point each.

A.1. The sampling stage for two arms

The initial sampling algorithm is similar in structure to that in Section 4, but chooses the "exploit" arm a^* in a different way (after several samples of arm 1 are drawn). Previously, a^* was the arm with the highest posterior mean given the collected samples, whereas now the selection of a^* is based only on the sample average $\hat{\mu}_1$ of the previously drawn samples of arm 1.

We use $\hat{\mu}_1$ instead of the posterior mean of arm 1, and compare it against the prior mean μ_2^0. (Note that μ_2^0 is also the posterior mean of arm 2, because the prior is independent across arms.) We still need $\mathbb{E}[\mu_i - \mu_j | a^* = i]$ to be positive for all arms i, j, even though $\hat{\mu}_1$ is only an imprecise estimate of the posterior mean of μ_1. For this reason, we pick $a^* = 2$ only if μ_2^0 exceeds $\hat{\mu}_1$ by a sufficient margin. Once a^* is selected, it is recommended in all of the remaining rounds, except for a few rounds chosen u.a.r. in which arm 2 is recommended. See Algorithm 7 for the details.

ALGORITHM 7: The sampling stage: samples both arms k times.

Parameters: $k, L \in \mathbb{N}$ and $C \in (0, 1)$.

1 Let $\hat{\mu}_1$ be the sample average of the resp. rewards;
2 if $\hat{\mu}_1 \leq \mu_2^0 - C$ then $a^* = 2$ else $a^* = 1$;
3 From the set P of the next $L \cdot k$ agents, pick a set Q of k agents uniformly at random;
4 Every agent $p \in P - Q$ is recommended arm a^*;
5 Every agent $p \in Q$ is recommended arm 2

We prove that the algorithm is BIC as long as $C \approx \frac{2}{3}\mu_2^0$, and parameters k^*, L are larger than some prior-dependent thresholds. We write out these thresholds explicitly, parameterizing them by the ratio $\lambda = C/\mu_2^0$.

LEMMA A.1. *Consider   BIC   bandit   exploration   with   two   arms.   Algorithm* DETAILFREESAMPLEBIC(k) *with parameters* (C, k^*, L) *completes in* $Lk + \max(k, k^*)$ *rounds. The algorithm is BIC if Property (P4) holds and* $\mu_2^0 > 0$, *and the parameters satisfy*

$$
\begin{aligned}
\lambda &\triangleq C/\mu_2^0 \in (0, \tfrac{2}{3}), \\
k^* &\geq 2 \left(\lambda \cdot \mu_2^0\right)^{-2} \log \tfrac{4}{\beta(\lambda)}, \\
L &\geq 1 + \tfrac{8}{\beta(\lambda)} \left(\mu_1^0 - \mu_2^0\right),
\end{aligned}
\tag{A.1}
$$

where $\beta(\lambda) = \lambda \cdot \mu_2^0 \cdot \Pr\left(\mu_1/\mu_2^0 \leq 1 - \tfrac{3\lambda}{2}\right)$.

Proof. Let $Z = \mu_2 - \mu_1$. Consider an agent $p > k$. As in the proof of Lemma 4.2 it suffices to show that $\mathbb{E}[Z|I_p = 2] \Pr(I_p = 2) \geq 0$.

Denote $\mathcal{E}_1 = \{\mu_1 \leq \mu_2^0 - C\}$. As in the proof of Lemma 4.2,

$$\mathbb{E}[Z|I_p = 2] \Pr(I_p = 2) = \mathbb{E}[Z|\mathcal{E}_1] \Pr(\mathcal{E}_1) \cdot \left(1 - \tfrac{1}{L}\right) + \mathbb{E}[Z] \cdot \tfrac{1}{L}.$$

Thus for the algorithm to be incentive compatible we need to pick L:

$$L \geq 1 + \frac{\mu_1^0 - \mu_2^0}{\mathbb{E}[Z|\mathcal{E}_1] \Pr(\mathcal{E}_1)}. \tag{A.2}$$

It remains to lower-bound the quantity $\mathbb{E}[Z|\mathcal{E}_1] \Pr(\mathcal{E}_1)$. Since the means μ_i are independent,

$$\mathbb{E}[Z|\mathcal{E}_1] \Pr(\mathcal{E}_1) = \mathbb{E}[\mu_2^0 - \mu_1|\mu_2^0 - \hat{\mu}_1 \geq C] \Pr\left(\mu_2^0 - \hat{\mu}_1 \geq C\right).$$

Denote $X = \mu_2^0 - \mu_1$ and $\hat{x}^k = \mu_2^0 - \mu_1^k$. Observe that $\hat{x}_k$ is the sample average of k i.i.d. samples from some distribution with mean X. Then the quantity that we want to lower bound is:

$$\mathbb{E}[Z|\mathcal{E}_1] \Pr(\mathcal{E}_1) = \mathbb{E}[X|\hat{x}_k \geq C] \Pr(\hat{x}_k \geq C). \tag{A.3}$$

We will use Chernoff-Hoeffding Bound to relate the right-hand side with quantities that are directly related to the prior distribution $\mathcal{P}$. More precisely, we use a corollary: Lemma A.4, which we state and prove in Appendix A.3.

If we apply Lemma A.4 with $C = \lambda \cdot \mu_2^0$ and $\zeta = \kappa = \tfrac{1}{2}$, then for $k \geq 2(\lambda \cdot \mu_2^0)^{-2} \log \frac{4}{\beta(\lambda)}$ the right-hand side of Eq. A.3 is at least $\tfrac{1}{8}\beta(\lambda)$. Therefore, $\mathbb{E}[Z|\mathcal{E}_1] \Pr(\mathcal{E}_1) \geq \tfrac{1}{8}\beta(\lambda)$. Plugging this back into Eq. A.2, it follows that the conditions (A.1) suffice to guarantee BIC. $\qquad\square$

A.2. The racing stage for two arms

The racing phase alternates both arms until one of them can be eliminated with high confidence. Specifically, we divide time in phases of two rounds each, in each phase select each arm once, and after each phase check whether $|\hat{\mu}_1^n - \hat{\mu}_2^n|$, is larger than some threshold, where $\hat{\mu}_i^n$ is the sample average of arm i at the beginning of phase n. If that happens then the arm with the higher sample average "wins the race", and we only pull this arm from then on. The threshold needs to be sufficiently large to ensure that the "winner" is indeed the best arm with very high probability, regardless of the prior. The pseudocode is in Algorithm 8.

ALGORITHM 8: BIC race for two arms.

Input : parameters $k \in \mathbb{N}$ and $\theta \geq 1$; time horizon T.

Input : k samples of each arm i, denoted $r_i^1, \ldots, r_i^k$.

1 Let $\hat{\mu}_i^k = \frac{1}{k}\sum_{t=1}^{k} r_i^t$ be the sample average for each arm i;

2 Split remainder into consecutive phases of two rounds each, starting from phase $n = k$;

   while $|\hat{\mu}_1^n - \hat{\mu}_2^n| \leq \sqrt{\frac{\log(T\theta)}{n}}$ **do**

3 | The next two agents are recommended both arms sequentially;

4 | Let r_i^n be the reward of each arm $i = 1, 2$ in this phase, and $\hat{\mu}_i^{n+1} = \frac{1}{n+1}\sum_{t=1}^{n} r_i^t$;

5 | $n = n + 1$;

   end

6 For all remaining agents recommend $a^* = \max_{a \in \{1,2\}} \hat{\mu}_a^n$

To guarantee BIC, we use two parameters: the number of samples k collected in the sampling stage, and parameter θ inside the decision threshold. The k should be large high enough so that

when an agent sees a recommendation for arm 2 in the racing stage there is a significant probability that it is due to the fact that arm 2 has "won the race" rather than to exploration. The θ should be large enough so that the arm that has "won the race" would be much more appealing than the losing arm according to the agents' posteriors.

LEMMA A.2 (**BIC**). *Assume two arms. Fix an absolute constant $\tau \in (0,1)$ and let $\theta_\tau = \frac{4}{\tau} / \Pr(\mu_2 - \mu_1 \geq \tau)$. Algorithm 8 is BIC if Property (P4) holds, and the parameters satisfy $\theta \geq \theta_\tau$ and $k \geq \theta_\tau^2 \log T$.*

The regret bound does not depend on the parameters; it is obtained via standard techniques.

LEMMA A.3 (**Regret**). *Algorithm 8 with any parameters $k \in \mathbb{N}$ and $\theta \geq 1$ achieves ex-post regret*

$$R(T) \leq \frac{8 \log(T\theta)}{|\mu_1 - \mu_2|}. \tag{A.4}$$

Denoting $\Delta = |\mu_1 - \mu_2|$ and letting n^* be the duration of the initial sampling stage, the ex-post regret for each round t of the entire algorithm (both stages) is

$$R(t) \leq \min\left(n^*\Delta + \frac{8\log(T\theta)}{\Delta}, \ t\Delta\right) \leq \min\left(2n^*, \ \sqrt{8t\log(T\theta)}\right). \tag{A.5}$$

In the remainder of this section we prove Lemma A.1 and Lemma A.2. Denote $\hat{z}_n = \hat{\mu}_2^n - \hat{\mu}_1^n$ and let $c_n = \sqrt{\frac{\log(T\theta)}{n}}$ be the decision threshold for each phase $n \geq k$ in Algorithm 8. For ease of notation we will assume that even after the elimination at every iteration a sample of the eliminated arm is also drawn, but simply not revealed to the agent. Let $Z = \mu_2 - \mu_1$. According to Chernoff-Hoeffding Bound, the following event happens with high probability:

$$\mathcal{C} = \{\forall n \geq k : |Z - \hat{z}_n| < c_n\}. \tag{A.6}$$

We make this precise in the following claim:

CLAIM A.1. $\Pr(\neg\mathcal{C}|Z) \leq \frac{1}{T\theta}$.

Proof. By Chernoff-Hoeffding Bound and the union bound, for any Z, we have:

$$\Pr(\neg\mathcal{C}|Z) \leq \sum_{n \geq n^*} \Pr(|Z - \hat{z}_n| \geq c_n) \leq \sum_{n \geq n^*} e^{-2n \cdot c_n^2} \leq T \cdot e^{-2\log(T\theta)} \leq \frac{1}{T\theta}.$$

$\square$

Generally we will view $\neg\mathcal{C}$ as a "bad event", show that the expected "loss" from this event is negligible, and then assume that the "good event" $\mathcal{C}$ holds.

Proof of Lemma A.2. Let $n^* = \theta_\tau^2 \log T$. Fix phase $n \geq n^*$, and some agent p in this phase. We will show that $\mathbb{E}[Z|I_p = 2] \Pr(I_p = 2) \geq 0$.

By Claim A.1, we have $\Pr(\neg\mathcal{C}|Z) \leq \frac{1}{T\theta} \leq 1/\theta_\tau$. Therefore, since $Z \geq -1$, we can write:

$$\mathbb{E}[Z|I_p=2]\Pr(I_p=2) = \ \mathbb{E}[Z|I_p=2,\mathcal{C}]\Pr(I_p=2,\mathcal{C}) + \mathbb{E}[Z|I_p=2,\neg\mathcal{C}]\Pr(I_p=2,\neg\mathcal{C})$$

$$\geq \ \mathbb{E}[Z|I_p=2,\mathcal{C}]\Pr(I_p=2,\mathcal{C}) - \theta_\tau^{-1}.$$

Now we will upper bound the first term. We will split the first integral into cases for the value of Z. Observe that by the definition of c_n we have $c_{n^*} \leq \theta_\tau^{-1}$. Hence, conditional on $\mathcal{C}$, if $Z \geq \tau \geq 2c_{n^*}$, then we have definitely stopped and eliminated arm 1 at phase $n = k \geq n^*$, since $\hat{z}_n > Z - c_n \geq c_n$. Moreover, if $Z \leq -2c_n$, then by similar reasoning, we must have already eliminated arm 2, since $\hat{z}_n < c_n$. Thus in that case event $I_p = 2$ cannot occur. Moreover, if we ignore the case when $Z \in [0, \tau)$,

then the integral can only decrease, since the integrand is non-negative. Since we want to lower bound it, we will ignore this region. Hence:

$$
\begin{aligned}
\mathbb{E}[Z|I_p = 2, \mathcal{C}] \Pr\left(I_p = 2, \mathcal{C}\right) \geq{}& \tau \Pr\left(\mathcal{C}, Z \geq \tau\right) - 2 \cdot c_n \Pr\left(\mathcal{C}, -2c_n \leq Z \leq 0\right) \\
\geq{}& \tau \Pr\left(\mathcal{C}, Z \geq \tau\right) - 2 \cdot c_n \\
\geq{}& \tau \Pr\left(\mathcal{C}, Z \geq \tau\right) - \frac{\tau \cdot \Pr\left(Z \geq \tau\right)}{2}
\end{aligned}
$$

Using Chernoff-Hoeffding Bound, we can lower-bound the probability in the first summand of the above:

$$
\begin{aligned}
\Pr\left(\mathcal{C}, Z \geq \tau\right) ={}& \Pr\left(\mathcal{C}|Z \geq \tau\right)\Pr\left(Z \geq \tau\right) \geq \left(1 - \sum_{t \geq n^*} e^{-2 \cdot c_n^2}\right) \cdot \Pr\left(Z \geq \tau\right) \\
\geq{}& \left(1 - \tfrac{1}{T}\right) \cdot \Pr\left(Z \geq \tau\right) \geq \tfrac{3}{4}\Pr\left(Z \geq \tau\right).
\end{aligned}
$$

Combining all the above inequalities yields that $\mathbb{E}[Z|I_p = 2]\Pr\left(I_p = 2\right) \geq 0$. $\qquad\square$

Proof of Lemma A.3. The "Chernoff event" $\mathcal{C}$ has probability at most $1 - \frac{1}{T}$ by Claim A.1, so the expected ex-post regret conditional on $\neg C$ is at most 1. For the remainder of the analysis we will assume that $\mathcal{C}$ holds.

Recall that $\Delta = |\mu_1 - \mu_2|$. Observe that $\Delta \leq |\hat{z}_n| + c_n$. Therefore, for each phase n during the main loop we have $\Delta \leq 2 \cdot c_n$, so $n \leq \frac{4 \log(T\theta)}{\Delta^2}$. Thus, the main loop must end by phase $\frac{4 \log(T\theta)}{\Delta^2}$.

In each phase in the main loop, the algorithm collects regret Δ per round. Event $\mathcal{C}$ guarantees that the "winner" a^* is the best arm when and if the condition in the main loop becomes false, so no regret is accumulated after that. Hence, the total regret is $R(T) \leq \frac{8 \log(T\theta)}{\Delta}$, as claimed in Eq. A.4.

The corollary (A.5) is derived exactly as the corresponding corollary in Lemma 6.3 $\qquad\square$

A.3. A Chernoff bound for the proof of Lemma A.1

In this section we state and prove a version of Chernoff bound used in the proof of Lemma A.1.

LEMMA A.4. *Let $X \geq 1$ be a random variable. Let $x_1, \ldots, x_k \in [0, 1]$ be i.i.d. random variables with $\mathbb{E}[x_i|X] = X$. Let $\hat{x}_k = \frac{1}{k}\sum_{t=1}^{k} x_t$. Then for any $C, \zeta, \kappa \in (0, 1)$, if*

$$
k \geq \frac{-\log\left(\kappa \cdot (1 - \zeta) \cdot C \cdot \Pr\left(X \geq (1 + \zeta)C\right)\right)}{2\zeta^2 \cdot C^2},
$$

then:

$$
\mathbb{E}[X|\hat{x}_k \geq C] \cdot \Pr\left(\hat{x}_k \geq C\right) \geq (1 - \zeta) \cdot C \cdot (1 - \kappa - \kappa(1 - \zeta)) \cdot \Pr\left(X \geq (1 + \zeta)C\right).
$$

We first prove the following Lemma.

LEMMA A.5. *Let $x_1, \ldots, x_k$ be i.i.d. random variables with $\mathbb{E}[x_i|X] = X$, and $X \geq -1$. Let $\hat{x}_k = \frac{1}{n}\sum_{t=1}^{k} x_t$. Then for any $C > 0$, $\epsilon \leq C$ and $\delta > 0$:*

$$
\mathbb{E}[X|\hat{x}_k \geq C] \cdot \Pr\left(\hat{x}_k \leq C\right) \geq (C - \epsilon) \cdot \left(1 - e^{-2 \cdot \delta^2 \cdot k}\right) \cdot \Pr\left(X \geq C + \delta\right) - e^{-2 \cdot \epsilon^2 \cdot k} \tag{A.7}
$$

Proof. Let $\mathcal{E}_{C-\epsilon}$ be the event $X \geq C - \epsilon$ and $\hat{\mathcal{E}}_C$ be the event that $\hat{x}_k \geq C$. By the Chernoff bound observe that:

$$
\Pr\left(\neg\mathcal{E}_{C-\epsilon}, \hat{\mathcal{E}}_C\right) \leq \Pr\left(|X - \hat{x}_k| > \epsilon\right) \leq e^{-2 \cdot \epsilon^2 \cdot k} \tag{A.8}
$$

Moreover, again by Chernoff bounds and a simple factorization of the probability:

$$
\begin{aligned}
\Pr\left(\mathcal{E}_{C-\epsilon}, \hat{\mathcal{E}}_C\right) &\geq \Pr\left(\mathcal{E}_{C+\delta}, \hat{\mathcal{E}}_C\right) = \Pr\left(\hat{\mathcal{E}}_C \mid \mathcal{E}_{C+\delta}\right) \cdot \Pr\left(\mathcal{E}_{C+\delta}\right) \\
&= \left(1 - \Pr\left(\hat{x}_k < C \mid X \geq C+\delta\right)\right) \cdot \Pr\left(X \geq C+\delta\right) \\
&\geq \left(1 - \Pr\left(\hat{x}_k - X < \delta \mid X \geq C+\delta\right)\right) \cdot \Pr\left(X \geq C+\delta\right) \\
&\geq \left(1 - e^{-2 \cdot \delta^2 \cdot k}\right) \cdot \Pr\left(X \geq C+\delta\right)
\end{aligned}
$$

We can now break apart the conditional expectation in two cases and use the latter lower and upper bounds on the probabilities of each case:

$$
\begin{aligned}
\mathbb{E}[X|\hat{\mathcal{E}}_C]\Pr\left(\hat{\mathcal{E}}_C\right) &= \mathbb{E}[X|\mathcal{E}_{C-\epsilon}, \hat{\mathcal{E}}_C]\Pr\left(\mathcal{E}_{C-\epsilon}, \hat{\mathcal{E}}_C\right) + \mathbb{E}[X|\neg\mathcal{E}_{C-\epsilon}, \hat{\mathcal{E}}_C] \cdot \Pr\left(\neg\mathcal{E}_{C-\epsilon}, \hat{\mathcal{E}}_C\right) \\
&\geq (C-\epsilon)\left(1 - e^{-2 \cdot \delta^2 \cdot k}\right) \cdot \Pr\left(X \geq C+\delta\right) - 1 \cdot e^{-2 \cdot \epsilon^2 \cdot k}
\end{aligned}
$$

Where we also used that $X \geq -1$. This concludes the proof of the lemma. $\qquad\square$

    Proof of Lemma A.4.   Now, if $k \geq \frac{-\log(\kappa \cdot (1-\zeta) \cdot C \cdot \Pr(X \geq (1+\zeta)C))}{2\zeta^2 \cdot C^2}$ and appplying the above Lemma for $\delta = \epsilon = \zeta \cdot C$ we get:

$$
\begin{aligned}
\mathbb{E}[X|\hat{\mathcal{E}}_C]\Pr\left(\hat{\mathcal{E}}_C\right) &\geq (1-\zeta)C(1 - \kappa \cdot (1-\zeta))\Pr\left(X \geq (1+\zeta)C\right) - \kappa(1-\zeta)C\Pr\left(X \geq (1+\zeta)C\right) \\
&\geq (1-\zeta)C(1 - \kappa - \kappa(1-\zeta))\Pr\left(X \geq (1+\zeta)C\right)
\end{aligned}
$$

Which concludes the proof of the Lemma. $\qquad\square$