Can Probabilistic Feedback Drive User Impacts in Online Platforms?

Jessica Dai
U.C. Berkeley

Bailey Flanigan
Carnegie Mellon University

Nika Haghtalab
U.C. Berkeley

Meena Jagadeesan
U.C. Berkeley

Chara Podimata
Massachussetts Institute of Technology
Archimedes / Athena RC

Abstract

A common explanation for negative user impacts of content recommender systems is misalignment between the platform's objective and user welfare. In this work, we show that misalignment in the platform's objective is not the only potential cause of unintended impacts on users: even when the platform's objective is fully aligned with user welfare, the platform's learning algorithm can induce negative downstream impacts on users. The source of these user impacts is that different pieces of content may generate observable user reactions (feedback information) at different rates; these feedback rates may correlate with content properties, such as controversiality or demographic similarity of the creator, that affect the user experience. Since differences in feedback rates can impact how often the learning algorithm engages with different content, the learning algorithm may inadvertently promote content with certain such properties. Using the multi-armed bandit framework with probabilistic feedback, we examine the relationship between feedback rates and a learning algorithm's engagement with individual arms for different no-regret algorithms. We prove that no-regret algorithms can exhibit a wide range of dependencies: if the feedback rate of an arm increases, some no-regret algorithms engage with the arm more, some no-regret algorithms engage with the arm less, and other no-regret algorithms engage with the arm approximately the same number of times. From a platform design perspective, our results highlight the importance of looking beyond regret when measuring an algorithm's performance, and assessing the nature of a learning algorithm's engagement with different types of content as well as their resulting downstream impacts.[1]

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 Introduction

Recommendation platforms—which facilitate our consumption of news, music, social media, and many other forms of digital content—can harm users in unintended ways, as documented by researchers [Allcott et al., 2020], journalists [Wells et al., 2021], and regulators [Commission, 2022]. One prevailing explanation for these impacts has been *misalignment* between the platform's objective (e.g., platform profit or user engagement) and user welfare [Stray et al., 2021]. This raises the question: *Is aligning the platform's objective with user utility sufficient to avoid negative impacts on users?*

In this work, we show that even if the platform's objective *perfectly optimizes user utility*, the process by which the platform continually *learns* user preferences can induce unintended impacts on users. In this learning process, the platform's learning algorithm relies on observing users' reactions to content, such as whether a user clicked on a piece of content, pressed the like button, or retweeted it. Whether

[1] Authors listed alphabetically. See arXiv for the full version of this manuscript.

users react to content in observable ways can depend on the specifics of the content—e.g., the content could be controversial, provoking users to comment, or broadly relatable, prompting users to share it—in ways which are not captured by user utilities.[2] As a result, the approach by which the learning algorithm accounts for these differential rates of information gain can affect how often content with such properties (e.g., controversiality) is recommended. Unfortunately, the resulting impact on recommendations may inadvertently affect the overall user experience, as we describe in Examples 1 and 2.

We study the impact of the platform's learning algorithm within the *multi-armed bandits* framework with *probabilistic feedback*. In this model, each piece of content corresponds to an arm with a *loss*, which quantifies a fixed user's utility for the corresponding content and can vary over time. The platform's objective is regret minimization, and is aligned with maximizing user utility. To capture the fact that content may generate observable user data at different rates, each arm i has a fixed *feedback rate* f_i representing the probability of the algorithm observing a sample from that arm's loss distribution in a given round. The platform must then determine how to account for these differential rates of information gain in its learning algorithm—a choice which can significantly impact what content users see.

Rather than only focusing on regret, we study how often a bandit algorithm engages with individual arms, and how this depends upon the arm's feedback rate f_i. To quantify engagement with an arm, we introduce two measures: the *arm pull count* $\texttt{APC}_i$ (how often an algorithm pulls arm i in T rounds), and the *feedback observation count* $\texttt{FOC}_i$ (how often it sees feedback from arm i in T rounds).[3] To formalize how these measures vary with f_i, we introduce the notions of *feedback monotonicity* and *balance*. At a high level, an algorithm is positive (negative) feedback monotonic with respect to APC (FOC) if, when an arm's f_i increases, the algorithm weakly increases (decreases) $\texttt{APC}_i$ ($\texttt{FOC}_i$). An algorithm satisfies *balance* when such a change in i's feedback rate is guaranteed to have *no* effect on $\texttt{APC}_i$ ($\texttt{FOC}_i$).

The following examples illustrate how these types of feedback monotonicity properties can in turn affect downstream user experience on the platform. These effects transcend what is typically captured in individual utilities (how much a given user likes a given piece of content), instead constituting community-level, platform-level, and society-level impacts.

Example 1 (Own-group content and APC)**.** *For a given user, f_i may be higher for content that appears to be produced by own-group creators, e.g. creators who are demographically or ideologically similar to the user [Agan et al., 2023].[4] APC captures how often content is shown to users. If an algorithm induced positive monotonicity in APC, users may see own-group content disproportionately often, contributing to problems such as polarization and echo chambers.*

See Appendix A.1 for further examples.

1.1 Our contributions

We initiate the study of how a bandit algorithm's choice of arms to pull correlates with the probability of observing feedback for those arms. We introduce the measures APC (Def. 2) and FOC (Def. 3), which capture two aspects of how a bandit algorithm treats arms that can result in downstream impacts on users; feedback monotonicity and balance are the algorithmic properties we aim to analyze. We summarize our results in Table 1.

Our main technical finding is that no-regret algorithms for the probabilistic feedback setting can exhibit a range of behavior with respect to APC and FOC (Table 1). We illustrate this by constructing different families of no-regret algorithms with strikingly different monotonicity properties for both APC and FOC, where these differences are driven by how the algorithms respond to probabilistic feedback.

1. We present three black-box transformations ($\mathrm{BB}_{\mathrm{Divide}}, \mathrm{BB}_{\mathrm{Pull}}, \mathrm{BB}_{\mathrm{DA}}$) which convert a generic no-regret bandit algorithm for the *deterministic* feedback setting into a bandit algorithm for the *probabilistic* feedback setting (Sec. 3); each of these transformations has different consequences for APC and FOC.

2. We analyze these black-box transformations applied to concrete algorithms (UCB and AAE), and achieve both improved regret bounds and stricter monotonicity guarantees (Sec. 4.1, 4.2).

[2] We expect that user feedback rates are not intrinsically captured by user utility: for example, high-utility content may either induce a high feedback rate (e.g., if a user retweets controversial content that they agree with) or induce a low feedback rate (e.g., if the content is educational and does not provoke a response). Similarly, low-utility content may either incite response (e.g., if the user disagrees with controversial content) or be ignored (leading to low feedback rates).

[3] While these measures are linked through f_i, they can lead to different user impacts, so we consider both.

[4] The empirical study of Agan et al. [2023] is explicitly motivated by APC in the context of recommendations.

3. We give an algorithm which improves known regret bounds for adversarial losses, removing the dependence on the *minimum* feedback probability in Esposito et al. [2022] (Sec. 4.3).

Compared to regret, APC and FOC are finer-grained measures for the behavior of a bandit algorithm, so tightly analyzing how these properties change with f_i also requires finer-grained control than in typical regret analyses. To isolate the impact of modifying feedback probabilities, we use a coupling argument to explicitly compare the algorithm's behavior on two instances that are identical except for one f_i.

1.2 Related work

Our technical results build on the vast literature on multi-armed bandits (see Hazan et al. [2016] for a textbook treatment). Most relevant to our work is *multi-armed bandits with probabilistic feedback graphs* (e.g. Esposito et al. [2022]). This extends the framework of multi-armed bandits with feedback graphs [Alon et al., 2015], where at each round, when an arm is pulled, the loss of all of the neighbors of that arm is observed. In the probabilistic feedback setting, the graphs are drawn from a *distribution* at each time step. Recent work has studied regret guarantees for the probabilistic feedback graph setting for adversarial (e.g. Esposito et al. [2022], Ghari and Shen [2022]) and stochastic losses (e.g. Li et al. [2020], Cortes et al. [2020]). We study a special case of this framework where the graph is always (a union of) self-loops.

A handful of recent works have examined how the feedback observed by the bandit learner impacts the arm pull count APC. For example, Haupt and Narayanan [2022] study how the variance of the noise in the observations of arm rewards impacts APC for ϵ-Greedy in a 2-arm setting; in contrast, we vary the feedback probability that the reward is observed and study the behavior of more general algorithms and instances. Moreover, motivated by clickbait, Buening et al. [2023] also study how feedback probabilities impact APC, focusing on the K arms (content creators) strategically selecting feedback probabilities to optimize for APC. They design incentive-aware platform algorithms that optimize a utility function (that can take into account both clickthrough rates and rewards); in contrast, we consider no-regret platform algorithms that optimize only for arm reward, and analyze their impact in terms of monotonicity properties.

Separately, the measure APC has been studied in recent work that aims to achieve *fairness* across arms, with a focus on ensuring that higher mean reward arms are pulled more often than lower mean reward arms [Joseph et al., 2016]. Another notion of stronger constraints on arm pulls is *replicability* [Esfandiari et al., 2022], which seeks to ensure that an algorithm will pull arms in the same order across identical instances with high probability. Though related, this is distinct from our definition of APC and our goals of controlling monotonicity. Their algorithms employ a similar "block" approach as ours, though they give explicit algorithms rather than black-box transformations for generic algorithms. Discussion of further related work, including empirical evidence for probabilistic feedback, real-world interpretations of FOC and APC, and the societal impacts of recommender systems, can be found in Appendix A.2.

2 Model & Preliminaries

We model the interaction between the platform/learner and the user as a multi-armed bandit (MAB) that happens over T rounds. Each arm corresponds to a piece of content; "pulling" an arm corresponds to recommending that piece of content to the user. We say that an arm *"returns feedback"* if we observe its loss upon pulling it.

An *instance* of our problem is specified by $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$, where $\mathcal{A}$, $\mathcal{F}$, and $\mathcal{L}$ are defined as follows. Let $\mathcal{A} := [K]$ denote the set of K arms. Let $\mathcal{F} := [f_1, \dots, f_K]$ be the feedback probabilities for each arm, i.e., the value $f_i \in [0,1]$ denotes the probability with which arm i returns feedback when pulled. The probabilities $\mathcal{F}$ can be chosen arbitrarily by an adversary and are unknown to the learner, but remain fixed throughout the T rounds. In the *deterministic feedback setting*, $f_i = 1$ for all i; in the *probabilistic feedback setting*, f_i can be less than 1. Let $\mathcal{L}$ denote the process by which the losses $\ell_{i,t}$ are generated. Losses can be *adversarial* or *stochastic*. For adversarial losses, the sequence $\{\ell_{i,t}\}_{i\in[K],t\in[T]}$ can be chosen arbitrarily, but obliviously, i.e. before the start of the algorithm. For stochastic losses, each arm $i \in [K]$ has a loss distribution with mean $\bar{\ell}_i$ and variance 1 from which the per-round loss $\ell_{i,t}$ is sampled. For each arm i, we define $\Delta_i := \bar{\ell}_i - \min_j \bar{\ell}_j$ to be the difference between the mean loss of arm i and the mean loss of the optimal arm $\min_j \bar{\ell}_j$.

For an arm $i \in [K]$, the random variable $X_{i,t}$ corresponds to whether feedback will be observed if arm i is pulled at round t, i.e., $X_{i,t} \sim \texttt{Bern}(f_i)$. With some abuse of notation, we let $\ell_{i_\tau,\tau} \cdot X_{i_\tau,\tau}$ represent the observed loss at time τ, where $\ell_{i_\tau,\tau} \cdot X_{i_\tau,\tau} = \perp$ denotes lack of observation when $X_{i_\tau,\tau} = 0$ and $\ell_{i_\tau,\tau} \cdot X_{i_\tau,\tau} = \ell_{i_\tau,\tau}$ denotes the observed loss when

Algorithm		APC mono.		FOC mono.		Regret upper bound	
$\mathrm{BB}_{\mathrm{Divide}}(\textsc{Alg}, f^\star)$	Alg. 1	$\approx$	Thm. 3.2	$+$	Thm. 3.2	$R_{\textsc{Alg}}\,(Tf^\star/\ln(T)) \cdot \ln(T)/f^\star$	Thm. 3.1
$\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$	Alg. 2	$\approx/-$	Thm. 3.4	$\approx/+$	Thm. 3.4	$R_{\textsc{Alg}}(T) \cdot 1/\min_i f_i$	Thm. 3.3
$\mathrm{BB}_{\mathrm{DA}}(\textsc{Alg})$	Alg. 3	$\approx/+$	Thm. 3.6	$\approx/+$	Thm. 3.6	$R_{\textsc{Alg}}(T) \cdot 4\ln(T)/\min_j f_j$	Thm. 3.5
$\mathrm{BB}_{\mathrm{Pull}}(\mathrm{AAE})$	Alg. 7	$-^\diamond$	Thm. 4.2	$\approx^\diamond$	Thm. 4.2	$O\left(\ln(T) \cdot \sum_i 1/(\Delta_i f_i)\right)$	Thm. 4.1
$\mathrm{BB}_{\mathrm{DA}}(\mathrm{AAE})$	Alg. 9	$+^\diamond$	Thm. 4.3	$+$	Thm. 4.3	$O\left(\ln^2(T) \cdot \sum_i 1/(\Delta_i \min_j f_j)\right)$	Thm. 3.5
3-Phase EXP3	Alg. 4					$O\left(\sqrt{T\ln(K)\sum_{i\in[K]} 1/f_i}\right)$	Thm. 4.4

Table 1: $\textsc{Alg}$ is any no-regret bandit algorithm with regret $R_{\textsc{Alg}}$ in the deterministic feedback setting. $f^\star$ is a tunable parameter. AAE is active-arm elimination; UCB is the upper confidence bounds algorithm. In columns APC and FOC, $+$, $-$ indicate *strict* positive, negative feedback monotonicity. $\approx$ indicates *approximate* balance, differing across arms by up to a factor of $O(1/T)$. $\approx/+$ (resp. $\approx/-$) means that either approximate balance or positive (resp. negative) monotonicity may be achieved, depending on the underlying algorithm and problem instance. The superscript $\diamond$ indicates that the stated property holds only for suboptimal arms.

$X_{i_\tau,\tau} = 1$. Let $H_t = \{(i_\tau, \ell_{i_\tau,\tau} \cdot X_{i_\tau,\tau}, X_{i_\tau,\tau}\})\}_{\tau\in[t-1]}$ for some round t denote the *history* of play until round t, and $\mathcal{H}_t$ denote the family of all possible history trajectories until round t. An algorithm $\textsc{Alg} : \bigcup_{t=0}^{T} \mathcal{H}_t \to [K]$ produces a (possibly randomized) mapping from histories of play to arms to be chosen. We sometimes overload notation and write $\textsc{Alg}(t)$ to denote the mapping from H_t to i_t.

2.1 Measuring the behavior of an algorithm on an instance

We capture the behavior of an algorithm by the following three quantities. The first is the standard objective function in multi-armed bandits, an algorithm's *(pseudo-)regret*:[5]

Definition 1 (Regret)**.** *The (pseudo-)regret of an algorithm* $\textsc{Alg}$ *playing arm* $i_t \in [K]$ *at round* t *is defined as:*
$R_{\textsc{Alg}}(T) = \mathbb{E}[\sum_{t\in[T]} \ell_{i_t,t}] - \min_{j\in[K]} \mathbb{E}[\sum_{t\in[T]} \ell_{j,t}]$.

We are also interested in how much an algorithm engages with individual arms. To capture this, we define the quantities $\mathtt{FOC}_i$ and $\mathtt{APC}_i$ for arms $i \in [K]$.

Definition 2 (Arm Pull Count (APC))**.** *Given a problem instance* $\mathcal{I}$*, the arm pull count (*APC*) of an arm* i *over a run of an algorithm* $\textsc{Alg}$ *is equal to* $\mathtt{APC}_i(\mathcal{I}; \textsc{Alg}) = \mathbb{E}\big[\sum_{t\in[T]} \mathbb{1}[i_t = i]\big]$.

Definition 3 (Feedback Observation Count (FOC))**.** *Given a problem instance* $\mathcal{I}$*, the feedback observation count (*FOC*) of arm* i *over a run of an algorithm* $\textsc{Alg}$ *is equal to* $\mathtt{FOC}_i(\mathcal{I}; \textsc{Alg}) = \mathbb{E}\big[\sum_{t\in[T]} \mathbb{1}[i_t = i] \cdot X_{i_t,t}\big]$.

In all three definitions, the expectation is taken with respect to randomness in both the algorithm and the instance (i.e., loss distributions and feedback observations). When the instance $\mathcal{I}$ and algorithm $\textsc{Alg}$ are clear from context, we write $\mathtt{FOC}_i$ and $\mathtt{APC}_i$. A simple consequence of the definitions is that $\mathtt{FOC}_i$ and $\mathtt{APC}_i$ are related by a multiplicative factor of f_i.

[5] Throughout, we omit "pseudo" from the definition below for succinctness.

Lemma 2.1. *For any arm* i*, instance* $\mathcal{I}$*, and algorithm* $\textsc{Alg}$*, it holds that* $\mathtt{FOC}_i(\mathcal{I}) = f_i \cdot \mathtt{APC}_i(\mathcal{I})$.

We prove Lemma 2.1 in Appendix B; the result follows from noting that, at any time t, the realization of $X_{i_t,t}$ is independent of all history up to time t.

2.2 Feedback monotonicity and balance

Using FOC and APC, we formalize how an algorithm responds to feedback probabilities through *feedback monotonicity* and *balance*. We let $\widetilde{\mathcal{F}}(i)$ denote a set of feedback probabilities in which we have modified arm i's feedback rate, holding all else constant: that is, $f_i \neq \tilde{f}_i$, and $\forall j \neq i, f_j = \tilde{f}_j$. For an instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$, we use $\widetilde{\mathcal{I}}$ to notate the instance identical to $\mathcal{I}$ except for $\tilde{f}_i$, that is $\widetilde{\mathcal{I}} = \{\mathcal{A}, \widetilde{\mathcal{F}}(i), \mathcal{L}\}$. In our analysis, we let i be arbitrary, and only modify feedback for one arm $i \in [K]$ at a time, and analyze how $\mathtt{APC}_i$ and $\mathtt{FOC}_i$ would change if the feedback probabilities were $\widetilde{\mathcal{F}}(i)$ instead of $\mathcal{F}$. We formally define *monotonicity* and *balance* below.

Definition 4 (Feedback monotonicity.)**.** *An algorithm exhibits positive (resp. negative) feedback monotonicity wrt measure* $Q \in \{\mathtt{APC}, \mathtt{FOC}\}$ *if and only if for all* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$*, for all* $i \in \mathcal{A}$*, and for all pairs* $\tilde{f}_i, f_i \in [0, 1]$ *such that* $\tilde{f}_i > f_i$*, we have that* $Q_i(\widetilde{\mathcal{I}}) \geq Q_i(\mathcal{I})$ *(resp.* $Q_i(\widetilde{\mathcal{I}}) \leq Q_i(\mathcal{I})$*).*

Definition 5 (Balance)**.** *An algorithm is* balanced *with respect to a measure Q if for all pairs of $\widetilde{\mathcal{I}}, \mathcal{I}$, we have that $Q_i(\mathcal{I}) = Q_i(\widetilde{\mathcal{I}})$.*

The goal of our work is to examine the landscape of potential feedback monotonicity properties (positive, negative, balance) for each measure (`APC` and `FOC`). We note that not all combinations are achievable: in particular, the measure `FOC` cannot satisfy *balance* or *negative feedback monotonicity* across all instances and all arms.

Proposition 2.2. *Suppose that* Alg *has sublinear regret for stochastic losses in the probabilistic feedback setting. For any pair of feedback probabilities $\tilde{f}_i > f_i$, for sufficiently large T, there exists an instance $\mathcal{I}$ such that $\texttt{FOC}_i(\tilde{\mathcal{I}}) > \texttt{FOC}_i(\mathcal{I})$. In fact, $\texttt{FOC}_i(\tilde{\mathcal{I}}) - \texttt{FOC}_i(\mathcal{I}) > \frac{9}{10} \cdot T(\tilde{f}_i - f_i)$.*

We prove Proposition 2.2 in Appendix B. In the remaining sections, as we analyze what feedback monotonicities are achievable, we sometimes consider relaxed versions of the precise definitions (e.g., restricting to suboptimal arms), as we will make explicit in the theorem statements.

3 Algorithmic Transformations and Implications for `APC` and `FOC`

In order to understand how an algorithm behaves with respect to `APC` and `FOC`, we need to disentangle how it reacts to probabilistic feedback and how it incorporates feedback observations to make future decisions. To do this, we study *black-box (BB) transformations* of a generic no-regret algorithm Alg for the deterministic feedback setting into a no-regret algorithm BB(Alg) that accounts for probabilistic feedback. We call these transformations "black-box" as they require only query access to Alg.

We analyze three different black-box transformations, which exhibit distinct behavior with respect to regret, `APC`, and `FOC` (see Table 1). The first, $\text{BB}_{\text{Divide}}$, divides the time horizon T into equally-sized intervals and repeatedly pulls the same arm within each interval (Sec. 3.1). The second, BB_{Pull}, repeatedly pulls the same arm until feedback is observed (Sec. 3.2). The third, BB_{DA}, pulls each arm a pre-specified number of times, depending on the feedback probability f_i of that arm (Sec. 3.3).

High-level approach. All three transformations use the high-level idea of dividing T into *blocks*, where the transformed algorithm BB(Alg) pulls the same arm for all rounds in the same block. Rounds of BB(Alg) are indexed $t \in [T]$. We index blocks (rounds of Alg) with ϕ, and let Φ be the total number of blocks (calls to Alg) in the evaluation of BB(Alg) up to T. Finally, S_ϕ denotes the set of all t indices that are within block ϕ. Then each transformation proceeds as follows. For each ϕ, we notate $i_\phi^{\text{ALG}} := \text{Alg}(\phi)$, i.e. the arm selected by Alg in its ϕth round. Then, BB(Alg) pulls i_ϕ^{ALG} for $t \in S_\phi$ and returns an observation $\ell_{i_\phi^{\text{ALG}},\phi}$ to Alg. Each transformation implements two steps differently: first, defining S_ϕ, and second, returning $\ell_{i_\phi^{\text{ALG}},\phi}$ to Alg.

3.1 $\text{BB}_{\text{Divide}}$: Transformation for balanced `APC` and positive `FOC`

The first black-box transformation that we construct, $\text{BB}_{\text{Divide}}$, generates algorithms that approximately balance `APC`. $\text{BB}_{\text{Divide}}$, formalized in Alg.1, separates T into equally sized blocks of size $B = \lceil 3 \ln T / f^\star \rceil$, where $f^\star \in (0, \min_i f_i]$ is a tunable parameter for trading-off regret and monotonicity.

In the context of the high-level approach described above, the set S_ϕ is taken to be the next B timesteps on BB(Alg)'s time horizon, i.e. $S_\phi = \{(\phi - 1) \cdot B + 1, \dots, \phi \cdot B\}$, and $\ell_{i_\phi^{\text{ALG}},\phi}$ is taken to be a uniform-at-random draw from the set of observations $\{\ell_{i_t,t} : X_{i_t,t} = 1, t \in S_\phi\}$.

Algorithm 1: BBDivide(Alg, $f^\star$)

Set the block size to $B = \lceil 3 \ln T / f^\star \rceil$ and initialize round count $t = 1$.
for *blocks* $\phi \in \{1, \dots, \Phi = \lfloor T/B \rfloor\}$ **do**
  Let $i_\phi^{\text{ALG}} = \text{Alg}(\phi)$ be the arm chosen by Alg on its ϕth timestep.
  Let $S_\phi = \{(\phi - 1) \cdot B + 1, \dots, \phi \cdot B\}$.
  Pull i_ϕ^{ALG} for rounds $t \in S_\phi$, i.e. $i_t = i_\phi^{\text{ALG}}, \forall t \in S_\phi$ and let $t \leftarrow t + 1$.
  if $\exists t \in S_\phi$ *s.t.* $X_{i_t,t} = 1$ *(i.e. there are observations)* **then**
    Return a random observation to Alg, i.e. $\ell_{i_\phi^{\text{ALG}},\phi} \sim \text{Unif}\{\ell_{i_t,t} : X_{i_t,t} = 1, t \in S_\phi\}$.
  else Return a loss of 1 to Alg, i.e. $\ell_{i_\phi^{\text{ALG}},\phi} = 1$
For remaining rounds, pull a random arm.

First, we show the following regret bound which holds for both stochastic and adversarial losses.

Theorem 3.1 (Regret $\text{BB}_{\text{Divide}}$)**.** *Let* Alg *be any algorithm for the deterministic feedback setting that achieves regret at most $R_{\text{ALG}}(T)$ when losses are adversarial (resp. stochastic). Then, for $f^\star \in (0, \min_i f_i]$ and adversarial (resp. stochastic) losses,*

$$R_{\text{BB}_{\text{Divide}}(\text{ALG}, f^\star)}(T) \leq \frac{3 \ln T}{f^\star} R_{\text{ALG}} \left(\frac{T f^\star}{3 \ln T} \right).$$

Theorem 3.1 indicates that the regret of $\text{BB}_{\text{Divide}}(\text{Alg})$ exceeds that of Alg by at most a factor of $3 \ln T/f^\star$. When Alg is specified, direct applications of Theorem 3.1 can improve the $f^\star$ dependence (Appendix C.1.1).

$\mathrm{BB}_{\mathrm{Divide}}$ *approximately balances* `APC` and is *positive feedback monotonic* with respect to `FOC`.

Theorem 3.2. *[Impact of* $\mathrm{BB}_{\mathrm{Divide}}$ *on* `APC` *and* `FOC` *] Fix an instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$*. Let* $\widetilde{f}_i \geq f_i$*, and let* $\widetilde{\mathcal{I}} = \{\mathcal{A}, \widetilde{\mathcal{F}}(i), \mathcal{L}\}$*. For any algorithm* Alg *for the deterministic feedback setting and for any* $f^\star \leq \min_i f_i$*, if* T *is sufficiently large, then the algorithm* $\mathrm{BB}_{\mathrm{Divide}}(\mathrm{Alg}, f^\star)$ *satisfies* $|\mathtt{APC}_i(\mathcal{I}) - \mathtt{APC}_i(\widetilde{\mathcal{I}})| \leq 1/T$ *and* $\mathtt{FOC}_i(\widetilde{\mathcal{I}}) > \mathtt{FOC}_i(\mathcal{I})$*.*

These monotonicity results, together with our regret bound, suggest that $f^\star$ may have opposite effects on regret and monotonicity. By Theorem 3.1, a higher value of $f^\star$ decreases the regret bound, and setting $f^\star$ to be close to $\min_i f_i$ is optimal.[6] Conversely, for monotonicity, while Theorem 3.2 set $\tilde{f}_i > f_i$, the reverse statements would also hold (i.e., we could have instead set $f_i > \tilde{f}_i \geq f^\star$).[7] As such, a higher value of $f^\star$ restricts the set of feedback probabilities under which the monotonicity results apply. We give full proofs in Appendix C.1.

3.2 $\mathrm{BB}_{\mathrm{Pull}}$: Transformation for negative `APC` and positive `FOC`

Our second transformation $\mathrm{BB}_{\mathrm{Pull}}$, formalized in Alg. 2, generates algorithms with *negative* monotonicity in `APC`. $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{Alg})$ will pull i_ϕ^{Alg} until feedback is observed for that arm, return the observation to Alg. In terms of the structure described at the beginning of the section, if block ϕ starts at time step t, S_ϕ is implicitly defined as the set of time steps until there is an observation: i.e., $S_\phi = \left\{t' \geq t \mid X_{i_\phi^{\mathrm{Alg}}, t''} = 0 \ \forall t'' < t'\right\}$. The loss passed to Alg is the observation made at the end of S_ϕ, i.e. $\ell_{i_\phi^{\mathrm{Alg}}, \phi} := \ell_{i_\phi^{\mathrm{Alg}}, \max\{t \mid t \in S_\phi\}}$.

First, we bound regret for stochastic losses.[8]

Theorem 3.3 (Regret $\mathrm{BB}_{\mathrm{Pull}}$)**.** *Let* Alg *be any algorithm for the deterministic feedback setting that achieves regret at most* $R_{\mathrm{Alg}}(T)$ *for stochastic losses. Then, for stochastic losses,* $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{Alg})$ *achieves regret at most* $R_{\mathrm{BB}_{\mathrm{Pull}}(\mathrm{Alg})}(T) \leq R_{\mathrm{Alg}}(T) \cdot \frac{1}{\min_i f_i}$.

[6] Via an estimation phase, we can estimate $\min_i f_i$ without asymptotically affecting the regret guarantees.

[7] The lower bound on $\tilde{f}_i$ ensures that there is still an observation in each block with high probability, despite the lower feedback probability.

[8] Theorem 3.3 requires stochastic losses, because regret analysis in Theorem 3.1 relies on the block size being fixed (and arm-independent).

Algorithm 2: $\mathrm{BB}_{\mathrm{PULL}}(\mathrm{Alg})$

Begin with $\phi = 1$ and $t = 1$.
while $t \leq T$ **do**
 Let $i_\phi^{\mathrm{Alg}} = \mathrm{Alg}(\phi)$ be the arm chosen by Alg on its ϕth timestep.
 while $X_{i_\phi^{\mathrm{Alg}}, t} = 0$ *and* $t \leq T$ **do**
  Pull i_ϕ^{Alg}, i.e. $i_t = i_\phi^{\mathrm{Alg}}$, and let $t \leftarrow t + 1$.
 Return $\ell_{i_t, t}$ to Alg, i.e. $\ell_{i_\phi^{\mathrm{Alg}}, \phi} = \ell_{i_t, t}$ and let $\phi \leftarrow \phi + 1$.

Theorem 3.3 shows that applying $\mathrm{BB}_{\mathrm{Pull}}$ increases regret by up to a $1/\min_i f_i$ factor, improving upon the regret for $\mathrm{BB}_{\mathrm{Divide}}$ (Theorem 3.1) by a $\ln T$ factor. We next formalize the monotonicity of $\mathrm{BB}_{\mathrm{Pull}}$. We show that although `APC` is negative monotonic, `FOC` maintains positive monotonicity, like in $\mathrm{BB}_{\mathrm{Divide}}$.

Theorem 3.4. *[Impact of* $\mathrm{BB}_{\mathrm{Pull}}$ *on* `APC` *and* `FOC` *] Fix an instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ *with stochastic losses. Let* $\widetilde{f}_i \geq f_i$*, and let* $\widetilde{\mathcal{I}} = \{\mathcal{A}, \widetilde{\mathcal{F}}(i), \mathcal{L}\}$*. For any algorithm* Alg *for the deterministic feedback setting, the algorithm* $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{Alg})$ *satisfies* $\mathtt{APC}_i(\mathcal{I}) \geq \mathtt{APC}_i(\widetilde{\mathcal{I}})$ *and* $\mathtt{FOC}_i(\mathcal{I}) \leq \mathtt{FOC}_i(\widetilde{\mathcal{I}})$*.*

Proof sketch. A coupling argument illustrates that the only source of difference in $\mathcal{I}$ and $\widetilde{\mathcal{I}}$ is in the number of times that Alg is called by $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{Alg})$. A higher f_i means that Alg can be called more times before T runs out on $\widetilde{\mathcal{I}}$, giving positive monotonicity in `FOC`. For `APC`, higher f_i means fewer pulls per observation. See Appendix C.2 for full proof. ∎

3.3 $\mathrm{BB}_{\mathrm{DA}}$: Transformation for positive `APC` and positive `FOC`

The third-black box transformation generates algorithms that are *positive* monotonic in `APC`. Given an algorithm Alg for the deterministic feedback setting, $\mathrm{BB}_{\mathrm{DA}}$, formalized in Alg. 3, combines conceptual ingredients from $\mathrm{BB}_{\mathrm{Divide}}$ and $\mathrm{BB}_{\mathrm{Pull}}$ (DA is short for DivideAdjusted). As in $\mathrm{BB}_{\mathrm{Divide}}$, block sizes are pre-specified, but are also arm-dependent, as in $\mathrm{BB}_{\mathrm{Pull}}$. To set block sizes B_i for each arm, we make the additional assumption that the algorithm designer knows the feedback probabilities apriori, and set $B_i = \lceil \frac{3 \ln T}{f^\star}(1 + f_i) \rceil$ for $f^\star \in (0, \min_j f_j]$. In terms of the high-level approach described at the beginning of the section, the set S_ϕ is taken to be $\left\{(\phi - 1) \cdot B_{i_\phi}^{\mathrm{Alg}} + 1, \ldots, \phi \cdot B_{i_\phi}^{\mathrm{Alg}}\right\}$, and $\ell_{i_\phi^{\mathrm{Alg}}, \phi}$ is taken to be a uniform-at-random draw from the set of observations $\{\ell_{i_t, t} : X_{i_t, t} = 1, t \in S_\phi\}$.

First, we show the following regret bound.

Algorithm 3: $\text{BB}_{\text{DA}}(\textsc{Alg}, f^\star)$

Begin with $\phi = 1$ and $t = 1$.
while $t \leq T$ **do**
  Let $i_\phi^{\textsc{Alg}} = \textsc{Alg}(\phi)$, $B_\phi = \lceil \frac{3 \ln T}{f^\star}(1 + f_{i_\phi^{\textsc{Alg}}}) \rceil$, and $S_\phi = \{t, t+1, \ldots, \min(t + B_\phi, T)\}$.
  for $t \in S_\phi$ **do**
    Pull $i_\phi^{\textsc{Alg}}$, i.e. $i_t = i_\phi^{\textsc{Alg}}$, and let $t \leftarrow t + 1$.
  if $\exists t \in S_\phi$ *s.t.* $X_{i_t,t} = 1$ *(i.e. there are observations)* **then**
    Return a random observation to $\textsc{Alg}$, i.e. $\ell_{i_\phi^{\textsc{Alg}},\phi} \sim \text{Unif}\{\ell_{i_t,t} : X_{i_t,t} = 1, t \in S_\phi\}$.
  else Return a loss of 1 to $\textsc{Alg}$, i.e. $\ell_{i_\phi^{\textsc{Alg}},\phi} = 1$.
Update $\phi \leftarrow \phi + 1$.

Theorem 3.5. *[Regret* BB_{DA}*] Let* $\textsc{Alg}$ *be any algorithm for the deterministic feedback setting that achieves regret at most* $R_{\textsc{Alg}}(T)$ *when the losses are stochastic. Then, for stochastic losses, for any* $f^\star \leq \min_i f_i$*, the algorithm* $\text{BB}_{\text{DA}}(\textsc{Alg}, f^\star)$ *achieves regret at most* $R_{\text{BB}_{\text{DA}}(\textsc{Alg})}(T) \leq \frac{6 \ln T}{f^\star} R_{\textsc{Alg}}\left(\frac{T f^\star}{3 \ln T}\right)$.

Since block size increases with f_i, $\text{BB}_{\text{DA}}(\textsc{Alg})$ pulls an arm more frequently when its feedback probability increases. More formally, increasing the feedback probability of an arm (approximately) *increases* the number of times it is pulled within any block where $\textsc{Alg}$ selects it. We show BB_{DA} exhibits positive monotonicity for both `APC` and `FOC`.

Theorem 3.6. *[Impact of* BB_{DA} *on* `APC` *and* `FOC` *] Fix an instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ *with stochastic losses. Let* $\tilde{f}_i \geq f_i$*, and let* $\widetilde{\mathcal{I}} = \{\mathcal{A}, \widetilde{\mathcal{F}}(i), \mathcal{L}\}$*. For any algorithm* $\textsc{Alg}$ *for the deterministic feedback setting and for any* $f^\star \leq \min_i f_i$*, the algorithm* $\text{BB}_{\text{DA}}(\textsc{Alg}, f^\star)$ *satisfies* $\texttt{APC}_i(\widetilde{\mathcal{I}}) \geq \texttt{APC}_i(\mathcal{I}) - 1/T$ *and* $\texttt{FOC}_i(\widetilde{\mathcal{I}}) \geq \frac{\tilde{f}_i}{f_i}\texttt{FOC}_i(\mathcal{I}) - \frac{\tilde{f}_i}{T} > \texttt{FOC}_i(\mathcal{I})$.

Theorem 3.6 also follows from a coupling argument; we defer proofs to Appendix C.3.

4 Finer-Grained Analyses of Monotonicity and Regret

While the black-box transformations in Section 3 provided a clean way to analyze the behavior of `FOC` and `APC`, the regret bounds obtained for those transformations unfortunately scaled with the *minimum* feedback probability $1/\min_{i \in [K]} f_i$ of any arm, and the monotonicity analysis did not differentiate between strict monotonicity and balance. In this section, we introduce four concrete algorithms—which are variants of EXP3 [Auer et al., 2002b], UCB [Auer et al., 2002a], and AAE (Active Arm Elimination [Even-Dar et al., 2002])—that have improved monotonicity and/or regret guarantees.

In Section 4.1, we show that applications of BB_{Pull} to AAE and UCB can also achieve improved regret bounds that scale with the average feedback probability $\sum_{i \in [K]} {}^{1}/_{f_i}$ across arms rather than the minimum feedback probability ${}^{K}/_{(\min_{i \in [K]} f_i)}$. Section 4.2 shows that more explicit analyses of BB_{Pull} and BB_{DA} applied to AAE enjoy stronger monotonicity guarantees than what is implied by naive applications of Theorems 3.4 and 3.6. Finally, in Section 4.3, we move beyond black-box transformations and present a variant of EXP3, which also achieves regret that scales with $\sum_{i \in [K]} {}^{1}/_{f_i}$ in the adversarial case, but which lacks clean monotonicity properties.

4.1 Improved regret guarantees

Consider BB_{Pull} applied to AAE and UCB.[9] First, we show that these algorithms achieve improved regret bounds compared to a naive application of Theorem 3.1.

Theorem 4.1. *On any stochastic instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{I}\}$*,* $\text{BB}_{\text{Pull}}(\text{AAE})$ *(presented in Algorithm 7) and* $\text{BB}_{\text{Pull}}(\text{UCB})$ *(presented in Algorithm 8) have regret bound of* $O\left(\sqrt{T \ln(T) \sum_{i \in [K]} 1/f_i}\right)$ *and an instance-dependent regret bound of* $O\left(\sum_{i \in [K] | \Delta_i > 0} \frac{\ln T}{\Delta_i f_i}\right)$.

To show this result, we extend analyses of AAE and UCB. We defer a full proof to D.1 for $\text{BB}_{\text{Pull}}(\text{AAE})$ and to D.2 for $\text{BB}_{\text{Pull}}(\text{UCB})$.

Theorem 4.1 converts the dependence on the minimum feedback probability $\min_j f_j$ from Theorem 3.3 into a finer-grained dependence on the per-arm feedback probabilities f_j. In particular, in the instance-dependent regret bounds of Theorem 4.1, the *"effective"* gap $\Delta_i f_i$ can be small either if the arm is close to optimal or if the feedback probability is small. In contrast, the regret bound of $O(\sum_{i \in [K]} \frac{\ln T}{\Delta_i \min_j f_j})$ given by applying Theorem 3.3 directly has an effective gap $\Delta_i \min_j f_j$ that can be small even if $\min_j f_j$ is small. Similarly, the instance-independent regret bounds in Theorem 4.1, in comparison to the instant-independent regret bound of $O(\sqrt{T(\ln T)\frac{K}{\min_i f_i}})$, also replace the dependence on $K/\min_i f_i$ with $\sum_{j \in [K]} 1/f_j$.

Interestingly, the improvement in regret relies on the specifics of BB_{Pull}: we do not expect it to be possible to obtain a similar improvement in regret for

[9] See Appendix D for formal statements of each algorithm.

$\text{BB}_{\text{Divide}}$ or BB_{DA} applied to AAE or UCB. Intuitively, this is because BB_{Pull} does not pull any arm more than necessary to observe feedback, while $\text{BB}_{\text{Divide}}$ and BB_{DA} must pull all arms (including sub-optimal ones) a prespecified number of times.

4.2 Stricter monotonicity guarantees

When the transformations BB_{Pull} and BB_{DA} are applied to AAE, we show stronger monotonicity properties for *suboptimal* arms (any arm i where $\bar{\ell}_i > \min_{j\in[K]} \bar{\ell}_j$). $\text{BB}_{\text{Pull}}(\text{AAE})$ achieves *strict negative* monotonicity in `APC` and *approximate balance* in `FOC` (Thm. 4.2), while $\text{BB}_{\text{DA}}(\text{AAE})$ achieves *strict positive* monotonicity in `APC` (Thm. 4.3).[10]

We start by analyzing $\text{BB}_{\text{Pull}}(\text{AAE})$ (formalized in Algorithm 7). We show that for suboptimal arms, $\text{BB}_{\text{Pull}}(\text{AAE})$ is approximately balanced for `FOC` as long as T is sufficiently large, implying (with Lemma 2.1) that `APC` strictly decreases in f_i.

Theorem 4.2. *Fix a stochastic instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$. *Let* i *be such that* $\bar{\ell}_i > \min_{j\in[K]} \bar{\ell}_j$. *Let* $\widetilde{f}_i > f_i$, *and let* $\widetilde{\mathcal{I}} = \{\mathcal{A}, \widetilde{\mathcal{F}}(i), \mathcal{L}\}$. *For sufficiently large* T, $\text{BB}_{\text{Pull}}(\text{AAE})$ *satisfies* $|\texttt{FOC}_i(\mathcal{I}) - \texttt{FOC}_i(\widetilde{\mathcal{I}})| \le 1/T$ *and* $\texttt{APC}_i(\widetilde{\mathcal{I}}) < \texttt{APC}_i(\mathcal{I})$.

Thm. 4.2 strengthens the monotonicity properties of BB_{Pull}: $\text{BB}_{\text{Pull}}(\text{AAE})$ satisfies *strict* (rather than *weak*) negative monotonicity in `APC`, and *approximate balance* (rather than *weak positive* monotonicity) in `FOC`.[11] The proof of Thm. 4.2 leverages a modified version of the coupling argument from the proof of Thm. 3.4. We defer a full proof to Sec. D.1.

To achieve strictly positive feedback monotonicity in `APC`, we turn to $\text{BB}_{\text{DA}}(\text{AAE})$ (formalized in Algorithm 9). The monotonicity properties of $\text{BB}_{\text{DA}}(\text{AAE})$ are given in Theorem 4.2.

Theorem 4.3. *Fix a stochastic instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$. *Let* i *be such that* $\bar{\ell}_i > \min_{j\in[K]} \bar{\ell}_j$. *Let* $\widetilde{f}_i > f_i$, *and let* $\widetilde{\mathcal{I}} = \{\mathcal{A}, \mathcal{F}(i), \mathcal{L}\}$. *For any* $f^\star \le \min_i f_i$ *and sufficiently large* T, $\text{BB}_{\text{DA}}(\text{AAE}, f^\star)$ *satisfies* $\texttt{APC}_i(\widetilde{\mathcal{I}}) > \texttt{APC}_i(\mathcal{I})$ *and* $\texttt{FOC}_i(\widetilde{\mathcal{I}}) > \texttt{FOC}_i(\mathcal{I})$.

[10] For both of these results, we focus on suboptimal arms for technical reasons (for large T, AAE eventually only pulls the optimal arm, which would equalize `APC`). Despite this restriction, we expect that the qualitative impacts of monotonicity still arise even if monotonicity holds for all arms but the optimal arm.

[11] At first glance, it would appear that this result contradicts Prop. 2.2, because we show balance is possible for `FOC` (for suboptimal arms). However, Prop. 2.2 only shows that balance is not possible across *all* arms (in particular, the optimal arm necessarily exhibits positive feedback monotonicity).

We defer a full proof to Appendix D.3.

4.3 Improving regret for adversarial losses

Next, we consider the adversarial setting and aim for regret that scales with $\sum_{i\in[K]} 1/f_i$, to match the stochastic result of Thm. 4.1. Like in the stochastic setting, our black-box transforms fail to achieve this improved regret dependence: the regret analysis from Thm. 3.1 for $\text{BB}_{\text{Divide}}(\text{EXP3})$ results in regret that unavoidably scales with $\sqrt{K/\min_i f_i}$,[12] because $\min_i f_i$ is explicitly used for determining the block size in $\text{BB}_{\text{Divide}}$, and our regret guarantees in Section 3 for BB_{Pull} and BB_{DA} are restricted to stochastic losses. Moreover, directly using standard EXP3 incurs linear regret (Prop. E.1).

We move beyond the black-box framework and construct 3-Phase EXP3 (Algorithm 4), an algorithm that achieves improved regret bounds.

Algorithm 4: 3-Phase EXP3

Phase 1: Set $N = \lceil 8\log(TK) \rceil$.
for *arms* $i \in [K]$ **do**
  Pull arm i until a reward is observed N times.
  Set P_i^{LR} to be the total number of rounds taken by the previous step divided by N.
Phase 2: for *arms* $i \in [K]$ **do**
  Pull arm i until a reward is observed.
  Set P_i^{E} to be the number of rounds taken by the previous step.
Let t_0 indicate the current round (after the completion of phase 1 and 2).
Let $\pi_{i,t_0} = 1/K$ for all $i \in [K]$.
Phase 3: Set $\eta = \sqrt{\frac{\log K}{T\sum_{i\in[K]} P_i^{LR}}}$.
for *rounds* $t = t_0, \dots, T$ **do**
  Pull an arm i_t with probability $\pi_{i_t,t}$.
  Update estimator: $\widehat{\ell}_{i,t} = \frac{\ell_{i,t}\cdot X_{i,t}}{\pi_{i,t}} P_i^E, \forall i \in [K]$.
  Update weights:
  $w_{i,t+1} = w_{i,t} \cdot \exp(-\eta \widehat{\ell}_{i,t}), \forall i \in [K]$.
  Update probability distribution:
  $\pi_{i,t+1} = \frac{w_{i,t+1}}{\sum_{j\in[K]} w_{j,t+1}}, \forall i \in [K]$.

Theorem 4.4. *Let* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{I}\}$ *be an adversarial instance such that* $\ell_{i,t} \in [0,1]$ *for all arms* i *and all time steps* $1 \le t \le T$. *For an oblivious adversary and unknown* f_i *values, Algorithm 4 incurs regret* $R(T) \le O\left(\sqrt{T\ln(K)\sum_{i\in[K]} 1/f_i}\right)$.

The intuition for Theorem 4.4 follows. If f_i's were known, a natural way to create an unbiased loss es-

[12] See Cor. C.1. Note this is still better than scaling with $K/\min_i f_i$, the naive implication of Thm. 3.1.

timator would have been $\hat{\ell}_{i,t} = (\ell_{i,t}/\pi_{i,t}) \cdot (X_{i,t}/f_i)$. The first two phases of the algorithm adjust the algorithm to account for *unknown* f_i. In particular, P_i^E is a low-variance unbiased estimator of $1/f_i$ and $P_i^{LR} = \Theta(1/f_i)$ with high probability. With these estimates, we adjust the second moment analysis of EXP3 while incurring only a constant overhead in the regret. We defer the full proof to Appendix E.3.

The regret from Theorem 4.4 also outperforms regret bounds from existing work on multi-armed bandits with probabilistic feedback (e.g. Esposito et al. [2022]). In particular, the feedback structure in our setting corresponds to a simple feedback graph consisting of a union of K self-loops (one for each arm) with probability f_i associated with the self-loop for arm i. Esposito et al. [2022] show a regret bound of $\widetilde{O}(\sqrt{TK/\min_i f_i})$ (with some additional optimizations when $\min_i f_i$ is very small). Their algorithm is very similar to $\text{BB}_{\text{Divide}}$ applied to EXP3 and uses a similar approach of splitting the time horizon into blocks. Esposito et al. [2022] provide an algorithm that achieves a regret bound of $\widetilde{O}(\sqrt{T \cdot \sum_{i\in[K]} 1/f_i})$, but only under the additional assumption, unsatisfied in our setting, that the full feedback graph is observed at every round. In comparison, our bound achieves a more fine-grained dependence on the feedback probabilities f_i.

However, it seems that improved regret for Algorithm 4 comes at the cost of clean monotonicity properties in `FOC` and `APC`. In fact, in Appendix E.2, we construct two simple instances that differ only in their loss functions. Even for a simplified version of 3-Phase EXP3 where f_i are known, the algorithm exhibits strictly positive monotonicity for `APC` in one instance and strictly negative monotonicity in the other. Examining 3-Phase EXP3 further, we can see that this may be due to its loss estimator directly incorporating estimates for f_i. Unlike algorithms generated with the black-box reductions in Section 3, 3-phase EXP3 does not permit a clean separation between how it reacts to probabilistic feedback and how it incorporates loss observations to make future decisions. This entanglement may make monotonicity unavoidably instance-dependent.

5 Discussion

In this work, we illustrate how the learning algorithm can inadvertently lead to downstream impacts on users even when the objective is perfectly aligned with user welfare. In particular, we show that the ways in which the algorithm handles heterogeneous rates of user reaction across different types of content can inadvertently impact the user experience. To study this, we provide a framework to investigate how the learning algorithm's engagement with individual arms depends on the feedback rates of the arms. We analyze the monotonicity of the arm pull count `APC` and the feedback observation count `FOC` in the feedback rates across the space of no-regret algorithms. From a platform design perspective, our results highlight the importance of measuring the feedback monotonicity of a learning algorithm as well as the resulting downstream impacts on users.

To achieve some of these monotonicities, our algorithms often require discarding information. This is an interesting parallel to the literature on robust bandits: While it is common to discard bandit observations that are produced by adversarial or non-myopic agents (e.g. Lykouris et al. [2018], Haghtalab et al. [2022], Gupta et al. [2019]), our work discards information not because the information is untrustworthy but because we aim to avoid undesirable downstream impact. On the other hand, with probabilistic feedback, information is already hard to come by; we may want to do better with the information we do have access to. An interesting open question, therefore, is about whether it might be possible to interpolate between monotonicity properties and how efficiently information is used, and whether such an approach can also improve regret.

Ultimately, which feedback monotonicities the platform may hope to induce will inherently depend on which downstream effects are desirable or concerning. This may differ across areas of the content space. Among content that generates constructive discussion, we may want *positive* feedback monotonicity in `FOC`, as this would elicit more beneficial user interaction. Meanwhile, among the kinds of content described in Example 1, *negative* monotonicity or balance may be preferred. As the algorithms we give in this work affect monotonicity over *all* content, more practical future directions may explore finer-grained control over monotonicity in different subsets of the content space.

Finally, though our theoretical analysis focuses on *monotonicity*, in real-world settings, more general *correlations* between feedback and `APC` or `FOC` may also be of concern to the platform. In Appendix F, we give a simulation study of correlations induced by common bandit algorithms. Combined with our rigorous monotonicity results, these simulations provide a bridge towards better understanding how probabilistic feedback can shape the impacts of a learning algorithm on users.

Acknowledgments

We thank anonymous reviewers for feedback on this work. JD is supported by an NSF GRFP and the AI Policy Hub at UC Berkeley. BF is supported by a Fannie and John Hertz Fellowship and an NSF GRFP. NH is supported in part by the National Science Foundation under grant CCF-2145898, a C3.AI Digital Transformation Institute grant, and a Google Research Scholars award. MJ is supported in part by an Open Philanthropy AI Fellowship. This work was partially done while the authors were visiting the Simons Institute for the Theory of Computing.

References

Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research*, 24(4):956–975, 2013.

Amanda Y Agan, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. Automating automaticity: How the context of human choice affects the extent of algorithmic bias. Technical report, National Bureau of Economic Research, 2023.

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. The welfare effects of social media. *American Economic Review*, 110(3):629–676, 2020.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 23–35. JMLR.org, 2015.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32 (1):48–77, 2002b.

Omer Ben-Porat and Moshe Tennenholtz. A game-theoretic approach to recommendation systems with strategic content providers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1118–1128, 2018.

Omer Ben-Porat, Itay Rosenberg, and Moshe Tennenholtz. Content provider dynamics and coordination in recommendation ecosystems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Thomas Kleine Buening, Aadirupa Saha, Christos Dimitrakakis, and Haifeng Xu. Bandits meet mechanism design to combat clickbait in online recommendation. *arXiv preprint arXiv:2311.15647*, 2023.

Dongmei Cao, Maureen Meadows, Donna Wong, and Senmao Xia. Understanding consumers' social media engagement behaviour: An examination of the moderation effect of social media context. *Journal of Business Research*, 122:835–846, 2021.

Micah D. Carroll, Anca D. Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and penalizing induced preference shifts in recommender systems. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2686–2708. PMLR, 2022.

The European Commission. Digital services act: EU's landmark rules for online platforms enter into force. 2022.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Ningshan Zhang. Online learning with dependent stochastic feedback graphs. In *International Conference on Machine Learning*, pages 2154–2163. PMLR, 2020.

Sarah Dean and Jamie Morgenstern. Preference dynamics under personalized recommendations. In David M. Pennock, Ilya Segal, and Sven Seuken, editors, *EC '22: The 23rd ACM Conference on Economics and Computation, Boulder, CO, USA, July 11 - 15, 2022*, pages 795–816. ACM, 2022.

Sarah Dean, Sarah Rich, and Benjamin Recht. Recommendations and user agency: the reachability of collaboratively-filtered information. In *Conference on Fairness, Accountability, and Transparency (FAT* '20)*, pages 436–445. ACM, 2020.

Michael D. Ekstrand and Martijn C. Willemsen. Behaviorism is not enough: Better recommendations through listening to users. In Shilad Sen, Werner Geyer, Jill Freyne, and Pablo Castells, editors, *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016*, pages 221–224. ACM, 2016.

Hossein Esfandiari, Alkis Kalavasis, Amin Karbasi, Andreas Krause, Vahab Mirrokni, and Grigoris Velegkas. Replicable bandits. In *The Eleventh International Conference on Learning Representations*, 2022.

Emmanuel Esposito, Federico Fusco, Dirk van der Hoeven, and Nicolò Cesa-Bianchi. Learning on the edge: Online learning with stochastic feedback graphs. *Advances in Neural Information Processing Systems*, 35:34776–34788, 2022.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270, 2002.

Emilio Ferrara and Zeyao Yang. Measuring emotional contagion in social media. *PloS one*, 10 (11):e0142390, 2015.

Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80:298–320, 2016.

Pouya M Ghari and Yanning Shen. Online learning with probabilistic feedback. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4183–4187. IEEE, 2022.

Wenshuo Guo, Karl Krauth, Michael I. Jordan, and Nikhil Garg. The stereotyping problem in collaboratively filtered recommender systems. In *ACM Conference on Equity and Access in Algorithms (EAAMO)*, pages 6:1–6:10, 2021.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

Nika Haghtalab, Thodoris Lykouris, Sloan Nietert, and Alexander Wei. Learning in stackelberg games with non-myopic agents. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 917–918, 2022.

Andreas Haupt and Aroon Narayanan. Risk aversion in learning algorithms and an application to recommendation systems. *arXiv preprint arXiv:2205.04619*, 2022.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Peter Henderson, Ben Chugg, Brandon Anderson, and Daniel E Ho. Beyond ads: Sequential decision-making algorithms in law and public policy. In *Proceedings of the 2022 Symposium on Computer Science and Law*, pages 87–100, 2022.

Jiri Hron, Karl Krauth, Michael I. Jordan, Niki Kilbertus, and Sarah Dean. Modeling content creator incentives on algorithm-curated platforms. *CoRR*, abs/2206.13102, 2022.

Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. Supply-side equilibria in recommender systems. *CoRR*, abs/2206.13489, 2022.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. 2005.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 781–789, 2017.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. In David M. Pennock, Ilya Segal, and Sven Seuken, editors, *EC '22: The 23rd ACM Conference on Economics and Computation, Boulder, CO, USA, July 11 - 15, 2022*, page 29. ACM, 2022.

Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National academy of Sciences of the United States of America*, 111(24):8788, 2014.

Shuai Li, Wei Chen, Zheng Wen, and Kwong-Sak Leung. Stochastic online learning with probabilistic graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4675–4682, 2020.

Yiyi Li and Ying Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2020.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

Smitha Milli, Luca Belli, and Moritz Hardt. From optimizing engagement to measuring value. In Madeleine Clare Elish, William Isaac, and Richard S. Zemel, editors, *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 714–722. ACM, 2021.

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141, 2020.

Kevin Roose. The making of a youtube radical. *The New York Times*, 2019.

Tobias Schnabel, Paul N Bennett, and Thorsten Joachims. Shaping feedback data in recommender systems with interventions based on information foraging theory. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 546–554, 2019.

Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. What are you optimizing for? aligning recommender systems with human values. *CoRR*, abs/2107.10939, 2021.

Jonathan Stray, Alon Y. Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Lex Beattie, Michael D. Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy X. Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. Building human values into recommender systems: An interdisciplinary synthesis. *CoRR*, abs/2207.10192, 2022.

Georgia Wells, Jeff Horwitz, and Deepa Seetharaman. Facebook knows instagram is toxic for teen girls, company documents show. *The Wall Street Journal*, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes — in appendix]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes – in appendix]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Supplemental Materials for Section 1

A.1 Additional Motivating Examples

Here, we provide additional elaboration for why varied feedback rates may cause downstream impacts on users, motivating our study of `APC` and `FOC`.

Example 2 (Incendiary content and `FOC`.)**.** *Observable feedback often occurs in the form of "retweets," and high f_i can be associated with highly controversial or incendiary content.* `FOC` *captures observable engagement metrics. If an algorithm induced positive monotonicity in* `FOC`*, creators may be incentivized to optimize for* `FOC` *by creating more incendiary content; this would increase the incendiariness of the overall landscape of content available on the platform. Moreover, since retweets by users about incendiary content are visible to other users, positive monotonicity in* `FOC` *may also create a toxic environment on the platform and impact the overall user experience.*

Example 3 (Clickbait and `APC`)**.** *Observable feedback often occurs in the form of "clicks" or "likes/dislikes;" a high feedback rate is correlated with how "clickbaity" the content is. Creators may be incentivized to optimize for* `APC`*, which captures objectives like view count. A creator can easily increase clickbaitiness (e.g. by changing content title or video thumbnail) without affecting the content's true utility. Thus, if the algorithm induces positive correlations between* `APC` *and feedback rate, creators may seek to make cosmetic changes without necessarily improving content quality. In the absence of positive correlation between* `APC` *and feedback rates, creators would be unable to rely on cheap strategies to generate engagement; instead, they would need to actually improve the quality of their content in order to increase the likelihood that it is shown to users.*

More generally, even absent creator incentives, we can see that different relationships between feedback rate and `APC`/`FOC` can result in qualitative differences in user experience.

Example 4 (Recommended topics)**.** *Certain content topics, such as political commentary or news, may naturally correspond to higher feedback rates than other topics, such as scientific or educational material. If the algorithm induces positive correlations between APC and feedback rate, then this would result in more political content shown to users; if it induces negative correlations, more educational content would be shown to users. Both have significant consequences for the overall qualitative user experience on the platform. (Of course it is possible for the platform to manually up- or down-weight content of various topics to control the content balance shown to users; however, we are interested in understanding possibly-unexpected changes that arise as a consequence of the learning algorithm itself.) Adding the possibility of creator incentives in this setting only amplifies these effects.*

Finally, we would like to highlight that understanding how algorithms behave with respect to APC and FOC under probabilistic feedback settings is of general interest in many applications where bandits are used to model sequential decisionmaking settings, even beyond content recommendation in online platforms.

Example 5 (Advertising)**.** *In online advertising, retailers trying to place ads via a centralized platform (such as Google) can decide whether to pay the platform per-click, or per-conversion. We can think of the number of times an ad is shown as* `APC`*, and the number of times an ad is clicked as* `FOC`*. If the retailers choose pay-per-conversion, the resulting data provided to the platform can be viewed as having a lower feedback probability than the data that would have been provided for pay-per-click: this is because a conversion only happens a subset of the time that a click happens. Retailers may want to maximize the number of times that their ad is shown, which is captured by* `APC`*. Whether an algorithm induces positive or negative correlations with* `APC` *and feedback rate could affect which of these payment models advertisers decide to select.*

Example 6 (Audits and public policy)**.** *Because it is costly to ensure that every single person or organization complies exactly with established standards or laws, governments often instead prefer to conduct audits, where some people or organizations are selected for an audit.*[13] *In this model, we can think of an arm as the person or organization to be audited; the pull of an arm as an audit; and feedback observation as whether the government will be able to get the ground truth "yes/no" for whether the law was violated. Why might some arms have lower or higher feedback probabilities? There may be some other reasons/features of the*

[13] See Henderson et al. [2022] for a more extensive discussion of bandits applied in similar contexts.

arm that affect feedback probability: for example, it may be harder to observe feedback for small businesses (vs bigger ones that have structured accounting departments), or non-English-speaking businesses. This reasoning also gives intuition for why it may be undesirable to pull low or high feedback arms more often (i.e. why monotonicity in APC *may be problematic) — for example, perhaps this means that in the long run, minority-owned businesses or smaller companies are audited disproportionately more than larger ones.*

A.2 Additional Related Work

Empirical evidence for probabilistic feedback. The idea that recommendation platforms may not observe all user "utilities" at all times is well-studied. While the intuition that expressed preferences may not be a full picture of their true preferences underpins an entire subfield of behavioral economics, we note several works here that study the problem applied to recommendation systems through a more algorithmic lens. In particular, probabilistic feedback often occurs for reasons that cannot be fully explained by quality of the content itself, which motivates our idea that f_i should be studied separately from utilities. For example, Schnabel et al. [2019] show that probabilistic feedback can arise from interface design choices; Joachims et al. [2005] uses eye-tracking to show that clickthrough (i.e. feedback) rates depend on factors like ranking position and the set of other content that is shown, while Joachims et al. [2017] applies this intuition to develop recommendation algorithms that are sensitive to the impact of ranking position on feedback rates; Li and Xie [2020] show that advertisements with images induce more user engagement than advertisements with text only, and that various attributes of images (e.g. colorfulness, professional versus amateur photography, human face, image-text fit) can also affect feedback rates; and Cao et al. [2021] find similar results in the context of fashion social media marketing, with both media richness and trustworthiness of marketing content as factors that affect feedback rates.

Real-world interpretations of FOC and APC. Many (though of course not all) of the commonly-discussed harms of recommendation systems and online platforms can be formalized in terms of FOC and APC. For example, the setting described in Wells et al. [2021]—harm to teen girls on Instagram—harm arises due to repeated exposure (APC) to particular types of content; in the setting described in Roose [2019]—radicalization on Youtube—the harm is due to "rabbit holes" that arise due to a combination of APC and FOC. In fact, though Roose [2019] is a general-audience reported case study, the more rigorous evaluation of Ribeiro et al. [2020] also examines both APC and FOC in the context of evaluating the role of algorithms in radicalization. Similarly, emotional contagion experiments (e.g. Ferrara and Yang [2015] on Twitter, Kramer et al. [2014] on Facebook) often find that exposure to (APC) content with emotional valence (either positive or negative) also affects the emotional valence of users' downstream posts.

Of particular note is Agan et al. [2023], which is the most closely-related empirical work to our knowledge. This recent work is an empirical study explicitly motivated by the harms of APC in recommendation systems, and correlations that may arise due to a learning algorithm's treatment of information; this work motivates our Example 1. In particular, they model f_i as related to "own-group" content, e.g. demographic similarity of the creator, and are concerned about algorithmic bias in the sense of over-representing content from "own-group" creators. They show that under this model, standard learning algorithms do in fact induce correlations between "own-group" content (i.e. f_i) and how often it is shown (i.e. APC). This work can be seen as an empirical validation that our theoretical framework may be concretely applicable.

Societal impacts of recommender systems. This research thread has broadly investigated misalignment between recommendations and user utility. One proposed source of misalignment is potential discrepancies between metrics derived from observed behavior (e.g. engagement) and user utility (e.g. Ekstrand and Willemsen [2016], Milli et al. [2021], Kleinberg et al. [2022]). Another source of misalignment that has recently been studied is how recommendations can shape user preferences over time [Adomavicius et al., 2013, Carroll et al., 2022, Dean and Morgenstern, 2022]. Furthermore, approaches for bringing human values in recommender system design have been investigated [Stray et al., 2021, 2022]. Several other societal impacts of recommender systems have been studied including the emergence of filter bubbles [Flaxman et al., 2016], stereotyping [Guo et al., 2021], the ability of users to reach different content [Dean et al., 2020], and content creator incentives induced by the recommendation algorithm [Ben-Porat and Tennenholtz, 2018, Ben-Porat et al., 2020, Jagadeesan et al., 2022, Hron et al., 2022].

B Supplemental Materials for Section 2

We prove Lemma 2.1, restated below for convenience.

Lemma 2.1. *For any arm i, instance $\mathcal{I}$, and algorithm* ALG, *it holds that* $\texttt{FOC}_i(\mathcal{I}) = f_i \cdot \texttt{APC}_i(\mathcal{I})$.

Proof of Lemma 2.1. For an arm $i \in [K]$, recall that we have defined $X_{i,t}$ to be the random variable corresponding to whether feedback is returned at round t, if arm i were to be pulled, i.e., $X_{i,t} \sim \texttt{Bern}(f_i)$, and H_t to be the history of algorithm ALG until round t: $H_t = \{(i_\tau, \ell_{i_\tau,\tau} \cdot \mathbb{1}\{X_{i_\tau,\tau}\}, X_{i_\tau,\tau})\}_{\tau\in[t-1]}$. Then, for any arm $i \in [K]$, by the definition of $\texttt{FOC}_i$ we have that:

$$\begin{aligned}
\texttt{FOC}_i(\mathcal{I}) &= \mathbb{E}_{H_t,\ell_{i_t,t},X_{i_t,t}}\left[\sum_{t\in[T]} \mathbb{1}\left[i_t = i\right] \cdot X_{i_t,t}\right] && \text{(Definition 3)}\\
&= \mathbb{E}_{H_t}\left[\mathbb{E}_{\ell_{i_t,t},X_{i_t,t}}\left[\sum_{t\in[T]} \mathbb{1}\left[i_t = i\right] \cdot X_{i_t,t}\middle| H_t\right]\right] && \text{(law of total expectation)}\\
&= f_i \cdot \mathbb{E}_{\ell_{i_t,t},X_{i_t,t}}\left[\sum_{t\in[T]} \mathbb{1}\left[i_t = i\right]\right] && \text{(note that } \mathbb{E}[X_{i,t} \mid H_t] = \mathbb{E}[X_{i_t,t}] = f_i)\\
&= f_i \cdot \texttt{APC}_i(\mathcal{I})
\end{aligned}$$

where the third equality is due to the fact that conditioning on H_t, the arm drawn by ALG i_t is independent of whether $X_{i_t,t}$ is 0 or 1. ∎

We prove Proposition 2.2, restated below for convenience.

Proposition 2.2. *Suppose that* ALG *has sublinear regret for stochastic losses in the probabilistic feedback setting. For any pair of feedback probabilities $\tilde{f}_i > f_i$, for sufficiently large T, there exists an instance $\mathcal{I}$ such that $\texttt{FOC}_i(\tilde{\mathcal{I}}) > \texttt{FOC}_i(\mathcal{I})$. In fact, $\texttt{FOC}_i(\tilde{\mathcal{I}}) - \texttt{FOC}_i(\mathcal{I}) > \frac{9}{10} \cdot T(\tilde{f}_i - f_i)$.*

Proof of Proposition 2.2. By Lemma 2.1, we have that $\texttt{FOC}_i(\widetilde{\mathcal{I}}) - \texttt{FOC}_i(\mathcal{I}) = \tilde{f}_i * \texttt{APC}_i(\widetilde{\mathcal{I}}) - f_i * \texttt{APC}_i(\mathcal{I})$, and since $f_i < \widetilde{f}_i$, we can write $f_i = c \cdot \widetilde{f}_i$ for some $c < 1$.

Let i be the optimal arm. A no-regret algorithm will pull i $T - o(T)$ times on both $\mathcal{I}$ and $\widetilde{\mathcal{I}}$, so that for a fixed T, there exists some $\alpha > 0$ where $\texttt{APC}_i(\widetilde{\mathcal{I}}) > T \cdot c^\alpha$.

Now, we have

$$\begin{aligned}
\tilde{f}_i * \texttt{APC}_i(\widetilde{\mathcal{I}}) - f_i * \texttt{APC}_i(\mathcal{I}) &> \tilde{f}_i \cdot c^\alpha \cdot T - f_i \cdot \texttt{APC}_i(\mathcal{I})\\
&\geq \tilde{f}_i \cdot c^\alpha \cdot T - f_i \cdot T\\
&= T \cdot \widetilde{f}_i \cdot (c^\alpha - c).
\end{aligned}$$

It remains to show that

$$\begin{aligned}
T \cdot \widetilde{f}_i \cdot (c^\alpha - c) > \frac{9}{10} \cdot T(\widetilde{f}_i - f_i) &= \frac{9}{10} \cdot T \cdot \widetilde{f}_i \cdot (1 - c)\\
\iff c^\alpha - c &> \frac{9}{10} - \frac{9}{10}c\\
\iff c^\alpha - \frac{c}{10} &> 9/10.
\end{aligned}$$

Taking T to be sufficiently large guarantees that α is sufficiently small for the above inequality to hold.

∎

C Supplemental Materials for Section 3

In this section, we provide proofs of regret and montonicity guarantees for our black-box transformations.

C.1 Proofs for Section 3.1: $\mathrm{BB}_{\mathrm{Divide}}$

Recall that $\mathrm{BB}_{\mathrm{Divide}}$, formalized in Algorithm 1, divides time horizon into equally-sized blocks of size $B = {}^{3\ln T}/_{f^\star}$. We analyze its regret and monotonicity properties below.

C.1.1 Corollaries of Theorem 3.5

Finally, we present several corollaries which give the regret of $\mathrm{BB}_{\mathrm{Divide}}$ applied to standard bandit algorithms.

Corollary C.1. *For fixed $f^\star \in (0, \min_i f_i]$, transformation $\mathrm{BB}_{\mathrm{Divide}}$ applied to standard EXP3 incurs the following regret:*

$$R_{\mathrm{BB}_{\mathrm{Divide}}(EXP3, f^\star))}(T) \leq O\left(\sqrt{\frac{1}{f^\star} \cdot TK\ln(T)\ln(K)}\right).$$

This follows from Theorem 3.1, along with the known result of Auer et al. [2002b] that the regret for standard EXP3 is $R_{EXP3}(T) \leq O(\sqrt{TK\ln K})$.

Corollary C.2. *For fixed $f^\star \in (0, \min_i f_i]$, transformation $\mathrm{BB}_{\mathrm{Divide}}$ applied to standard UCB incurs the following regret:*

$$R_{\mathrm{BB}_{\mathrm{Divide}}(UCB, f^\star))}(T) \leq O\left(\sum_{i \in [K]} \frac{\ln^2(T)}{\Delta_i \cdot f^\star}\right).$$

This follows from Theorem 3.1, along the known result of Auer et al. [2002a] that the instance-dependent regret for standard UCB is $R_{UCB}(T) \leq O\left(\ln T \cdot \left(\sum_{i\in[K]} \frac{1}{\Delta_i}\right)\right)$, where $\Delta_i = \bar{\ell}_i - \min_j \bar{\ell}_j$.

Corollary C.3. *For fixed $f^\star \in (0, \min_i f_i]$, transformation $\mathrm{BB}_{\mathrm{Divide}}$ applied to standard AAE incurs the following regret:*

$$R_{\mathrm{BB}_{\mathrm{Divide}}(AAE, f^\star))}(T) \leq O\left(\sum_{i \in [K]} \frac{\ln^2(T)}{\Delta_i \cdot f^\star}\right).$$

This follows from Theorem 3.1, along the known result by Even-Dar et al. [2002] that the instance-dependent regret for standard AAE is $R_{AAE}(T) \leq O\left(\ln T \cdot \left(\sum_{i\in[K]} \frac{1}{\Delta_i}\right)\right)$, where $\Delta_i = \bar{\ell}_i - \min_j \bar{\ell}_j$.

C.1.2 Regret of $\mathrm{BB}_{\mathrm{Divide}}$: Proof of Theorem 3.1

We prove Theorem 3.1 and give some applications to concrete algorithms. For convenience, we restate the regret bound of $\mathrm{BB}_{\mathrm{Divide}}$ below.

Theorem 3.1 (Regret $\mathrm{BB}_{\mathrm{Divide}}$). *Let* Alg *be any algorithm for the deterministic feedback setting that achieves regret at most $R_{\mathrm{ALG}}(T)$ when losses are adversarial (resp. stochastic). Then, for $f^\star \in (0, \min_i f_i]$ and adversarial (resp. stochastic) losses,*

$R_{\mathrm{BB}_{\mathrm{Divide}}(\mathrm{ALG}, f^\star)}(T) \leq \frac{3\ln T}{f^\star} R_{\mathrm{ALG}}\left(\frac{Tf^\star}{3\ln T}\right).$

To analyze the regret of $\mathrm{BB}_{\mathrm{Divide}}$ applied to a generic algorithm Alg, we will use the following lemma, which lower bounds the likelihood of seeing a sample from the true loss distribution in every block.

Lemma C.4. *Fix an $f^\star \in (0, \min_i f_i]$, and divide the time horizon T into blocks of size $B = \frac{3\ln T}{f^\star}$ and let $\Phi = \lfloor T/B \rfloor$, as in Algorithm 1. Suppose then that for each block $\phi \in \{1, 2, \dots, \Phi\}$, we play the same arm i_ϕ for every round in block ϕ. Let E be the "clean event" that at least one feedback observation occurs in* each *block ϕ, i.e., that for all blocks ϕ, $\exists t \in S_\phi : X_{i_t,t} = 1$. Then, $\Pr[E] \geq 1 - 1/T^2$.*

Proof. Let E_ϕ be the event that at least one feedback observation occurred in block ϕ, i.e., $\exists t \in S_\phi : X_{i_t,t} = 1$. Since for any arm i, $\Pr[X_{i,t} = 1] = f_i$, then for arm i_ϕ, we have that

$$\Pr[\neg E_\phi] = (1 - f_{i_\phi})^B \leq (1 - f^\star)^B \leq \exp(-f^\star B) = 1/T^3.$$

Union bounding over all $\lfloor T/B \rfloor$ blocks, we conclude that

$$\Pr[\neg E] \leq \sum_{\phi \in [\Phi]} \Pr[\neg E_\phi] \leq 1/T^2. \qquad \blacksquare$$

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1: Regret $\mathrm{BB}_{\mathrm{Divide}}$. Throughout the proof we will use $f^\star \in (0, \min_i f_i]$.

First, observe that Line 9 of Algorithm 1 (i.e., the last $T - B\Phi$ steps of the time horizon) contributes $O(\ln T)$ regret, because $T - B\Phi < T - B\frac{T}{B} + B \leq B = \frac{3 \ln T}{f^\star}$. The rest of the proof thus analyzes the regret incurred in the first $B\Phi$ time steps. Now, we divide up these rounds into Φ blocks of size B, and let E be the "clean event" that at least one feedback observation occurs in each block $\phi \in \{1, \ldots, \Phi\}$. By Lemma C.4, we have that $\Pr[E] \geq 1 - \frac{1}{T^2}$. The event that E does not occur contributes at most $O(1)$ to the expected regret, so we can condition on E for the remainder of the analysis.

Next, fix an instance with stochastic feedback $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ over T rounds. Now, we define corresponding instance with deterministic feedback $\mathcal{I}' = \{\mathcal{A}, (1, \ldots, 1), \mathcal{L}'\}$ over $\Phi = \lfloor T/B \rfloor$ time steps, where $\mathcal{L}'$ denotes the process generating the following sequence of Φ losses $\ell'_{i,1}, \ldots, \ell'_{i,\Phi}$ for all $i \in \mathcal{A}$: For all $i \in \mathcal{A}$ and $\phi \in \{1, \ldots, \Phi\}$, $\ell'_{i,\phi} \sim \mathrm{Unif}\{\ell_{i,s} : s \in S_\phi\}$, i.e., the loss is sampled uniformly from the loss functions of the original instance within block ϕ. Now, we show that the (pseudo)regret of $\textsc{Alg}$ on instance $\mathcal{I}'$ over Φ rounds is the same as that of $\mathrm{BB}_{\mathrm{Divide}}(\textsc{Alg}, f^\star)$ on instance $\mathcal{I}$ over T rounds. By the definition of the regret, the regret of $\textsc{Alg}$ on the instance $\mathcal{I}'$ is equal to

$$\mathbb{E}\left[\sum_{\phi=1}^{\Phi} \ell'_{i_\phi,\phi}\right] - \min_i \mathbb{E}\left[\sum_{\phi=1}^{\Phi} \ell'_{i,\phi}\right], \tag{1}$$

where the randomness of the first expectation is due to the randomness of the algorithm $\textsc{Alg}$ and $\mathcal{L}'$. The regret of $\mathrm{BB}_{\mathrm{Divide}}(\textsc{Alg})$ on $\mathcal{I}$ is equal to:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell_{i_t,t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{i,t}\right] &= \mathbb{E}\left[\sum_{\phi=1}^{\Phi} \sum_{t \in S_\phi} \ell_{i_\phi,t}\right] - \min_i \sum_{\phi=1}^{\Phi} \sum_{t \in S_\phi} \mathbb{E}[\ell_{i,t}] \\
&= B \cdot \left(\mathbb{E}\left[\sum_{\phi=1}^{\Phi} \underbrace{\frac{1}{B} \sum_{t \in S_\phi} \ell_{i_\phi,t}}_{(A)}\right] - \min_i \sum_{\phi=1}^{\Phi} \underbrace{\frac{1}{B} \sum_{t \in S_\phi} \mathbb{E}[\ell_{i,t}]}_{(B)} \right)
\end{aligned}$$

where randomness in the expectation is due to the randomness of $\mathrm{BB}_{\mathrm{Divide}}(\textsc{Alg})$, the randomness of feedback observations, and the randomness of the loss functions. Notice that (A) is equal to the $\mathbb{E}[\ell'_{i_\phi,\phi}]$ and (B) is equal to $\mathbb{E}[\ell'_{i,\phi}]$. Moreover, the observation at the end of the ϕth block, which is passed as the loss to $\textsc{Alg}$ for its ϕth timestep, is also distributed according to $\mathrm{Unif}\left\{\ell_{i_\phi,t} : t \in S_\phi\right\}$.

We thus see that:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell_{i_t,t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{i,t}\right] &= B \cdot \left(\mathbb{E}\left[\sum_{\phi=1}^{\Phi} \mathbb{E}[\ell'_{i_\phi,\phi}]\right] - \min_i \sum_{\phi=1}^{\Phi} \mathbb{E}[\ell'_{i,\phi}] \right) \\
&= B \cdot \left(\mathbb{E}_{\mathcal{L}'}\left[\sum_{\phi=1}^{\Phi} \ell'_{i_\phi,\phi}\right] - \min_i \mathbb{E}_{\mathcal{L}'}\left[\sum_{\phi=1}^{\Phi} \ell'_{i,\phi}\right] \right).
\end{aligned}$$

This expression corresponds exactly to B times the regret of $\textsc{Alg}$ on $\mathcal{I}'$ (see (1)), and thus can be upper bounded by

$$B \cdot \left(\mathbb{E}_{\mathcal{L}'}\left[\sum_{\phi=1}^{\Phi} \ell'_{i_\phi,\phi}\right] - \min_i \mathbb{E}_{\mathcal{L}'}\left[\sum_{\phi=1}^{\Phi} \ell'_{i,\phi}\right]\right) \leq B \cdot R_{\textsc{Alg}}(\Phi)$$
$$\leq B \cdot R_{\textsc{Alg}}(T/B)$$
$$\leq \frac{3\ln T}{f^\star} \cdot R_{\textsc{Alg}}\left(\frac{Tf^\star}{3\ln T}\right).$$

The $\frac{3\ln T}{f^\star}$ error term from the last $T - B\Phi$ steps can be absorbed into this regret term, since $R_{\textsc{Alg}}\left(\frac{Tf^\star}{3\ln T}\right) \geq 1$.

■

C.1.3 Monotonicity of $\mathrm{BB}_{\mathrm{Divide}}$: Proof of Theorem 3.2

Next, we analyze the monotonicity properties of $\mathrm{BB}_{\mathrm{Divide}}$. For convenience, we restate Theorem 3.2 below.

Theorem 3.2. *[Impact of $\mathrm{BB}_{\mathrm{Divide}}$ on `APC` and `FOC`] Fix an instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$. Let $\widetilde{f}_i \geq f_i$, and let $\widetilde{\mathcal{I}} = \{\mathcal{A}, \widetilde{\mathcal{F}}(i), \mathcal{L}\}$. For any algorithm $\textsc{Alg}$ for the deterministic feedback setting and for any $f^\star \leq \min_i f_i$, if T is sufficiently large, then the algorithm $\mathrm{BB}_{\mathrm{Divide}}(\textsc{Alg}, f^\star)$ satisfies $|\mathtt{APC}_i(\mathcal{I}) - \mathtt{APC}_i(\widetilde{\mathcal{I}})| \leq 1/T$ and $\mathtt{FOC}_i(\widetilde{\mathcal{I}}) > \mathtt{FOC}_i(\mathcal{I})$.*

The intuition for the `APC` statement is that since $\mathrm{BB}_{\mathrm{Divide}}$ effectively treats each block as one round of $\textsc{Alg}$, equalizing the block sizes will naturally balance `APC`. Once $\textsc{Alg}$ decides to pull an arm i, $\mathrm{BB}_{\mathrm{Divide}}(\textsc{Alg})$ will pull it B times regardless of its feedback probability. This result relies on $f^\star$ being sufficiently small to ensure that there is an observation in every block. The `FOC` statement follows from an application of Lemma 2.1. We formalize this below.

Proof of Theorem 3.2. We first analyze `APC`. Let E be the "clean event" that at least one feedback observation occurs in each block $1 \leq \phi \leq \Phi$. By Lemma C.4, we know that $\mathbb{P}[E] \geq 1 - \frac{1}{T^2}$. Conditioning on the clean event E, we see that $\mathtt{APC}_i(\mathcal{I}) = \mathtt{APC}_i(\tilde{\mathcal{I}})$ by the construction of the algorithm, since in every block where $\textsc{Alg}$ selects i, $\mathrm{BB}_{\mathrm{Divide}}(\textsc{Alg})$ will pull i exactly B times. The event that E does not occur contributes at most $1/T$ to `APC`.

We next analyze `FOC`. From the proof of the `APC` statement, we have that $\mathtt{APC}_i(\tilde{\mathcal{I}}) \geq \mathtt{APC}_i(\mathcal{I}) - 1/T$. Applying Lemma 2.1, which states that $\mathtt{FOC}_i = f_i \cdot \mathtt{APC}_i$, we have

$$\mathtt{FOC}_i(\tilde{\mathcal{I}}) = \tilde{f}_i \cdot \mathtt{APC}_i(\tilde{\mathcal{I}}) \geq \tilde{f}_i \cdot \mathtt{APC}_i(\mathcal{I}) - \tilde{f}_i/T = \frac{\tilde{f}_i}{f_i} \cdot \mathtt{FOC}_i(\mathcal{I}) - \tilde{f}_i/T \geq \mathtt{FOC}_i(\mathcal{I}) - \tilde{f}_i/T,$$

where the last inequality is because $\tilde{f}_i \geq f_i$.

For strict inequality, notice that it suffices to show that $\frac{\tilde{f}_i}{f_i} \cdot \mathtt{FOC}_i(\mathcal{I}) - \frac{\tilde{f}_i}{T} > \mathtt{FOC}_i(\mathcal{I})$. As long as $\textsc{Alg}$ pulls i at least once, this will hold for sufficiently large values of T. ■

C.2 Proofs for Section 3.2: $\mathrm{BB}_{\mathrm{Pull}}$

In this section, we provide proofs for the regret and monotonicity of algorithms transformed by $\mathrm{BB}_{\mathrm{Pull}}$. Before doing so explicitly, we first introduce a simulated version of $\mathrm{BB}_{\mathrm{Pull}}$, as well as Lemmas C.5 and C.6, which will help us compare transformed algorithms on similar instances.

C.2.1 Construction of a simulated version of $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{Alg})$

We first introduce Algorithm 5, a simulated version of $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$, which will be easier to analyze but behaves the same way as $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$. First, let us define the following random variables. (Recall that ϕ indexes losses for the time horizon of $\textsc{Alg}$, Φ is the total number of times $\textsc{Alg}$ is called by $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$, and $\Phi \leq T$ because $\textsc{Alg}$ can be called at most T times.)

- *Losses:* For each round $\phi \in [\Phi]$ of ALG and each arm $j \in [K]$, $\ell'_{j,\phi}$ is the placeholder for the loss passed to ALG if ALG were to observed the loss of arm j at time Φ. More formally, $\ell'_{j,\phi} := \ell_{j,t}$ the loss for arm j at a time step t that corresponds to the last time step in block ϕ of $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$. Since we are in the stochastic loss setting, $\ell'_{j,\phi}$ is a random variable drawn from the distribution of arm j (with mean $\bar{\ell}_j$) independently across ϕ and j. We note that these losses are only observed up to timestep Φ (which is a random variable less than T) and only for the specific arms pulled by the algorithm.
- *Feedback realizations:* For all $j \in [K]$ and $\phi \in [T]$, let $Q_{j,\phi} \sim \mathrm{Geom}(f_j)$ for $\phi \in [T]$ be a random variable distributed according to the geometric distribution with parameter equal to the feedback probability of arm j. This will represent the number of Bernoulli trials needed to observe a success. (These random variables are also fully independent across values of j and ϕ.)
- *Algorithm randomness:* Let b be randomness of ALG that will be used across time steps $1 \le \phi \le T$. Let $\textsc{Alg}_b$ denote ALG initialized with randomness b.

We are now ready to present the simulated version of $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$, described in Algorithm 5.

Algorithm 5: Simulated version of $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$

Input: A sequence of positive integers $Q_{j,\phi}$ for $\phi \in [T]$ and $j \in [K]$.

```
Initialize t = 1 and φ = 1.
while t ≤ T do
    Let i_φ^ALG = ALG(φ) be the output of ALG at timestep φ.
    for min(T − t, Q_{i_φ^ALG, φ}) iterations do
        Pull i_t := i_φ^ALG and let t ← t + 1.
    if t < T then
        Observe ℓ_{i_t, t} and return ℓ'_{i_φ^ALG, φ} := ℓ_{i_t, t} to ALG.
        Let φ ← φ + 1.
```

Note that the random variables $Q_{j_\phi,\phi}$ actually now capture the *block size* of the transformed algorithm B_ϕ (which, for $\mathrm{BB}_{\mathrm{Pull}}$, is a random variable). For clarity, we will use Q rather than B in the remaining analyses.

We first argue that, given two instances $\mathcal{I}, \tilde{\mathcal{I}}$ which are identical except for $\tilde{f}_i \ge f_i$, the sequences of arms that Algorithm 5 pulls are distributed identically across the instances. We formalize this in the following lemma.

Lemma C.5. *Let $Q_{j,\phi}$ and $\widetilde{Q}_{j,\phi}$ for $j \in [K]$ and $\phi = 1, \ldots,$ be an infinitely-long sequence of arbitrary positive integers. Let Φ^* be any positive integer and $T = \max\{\sum_{\phi\in[\Phi^*]} \sum_{j\in[K]} Q_{j,\phi}, \sum_{\phi\in[\Phi^*]} \sum_{j\in[K]} \widetilde{Q}_{j,\phi}\}$ be the time horizon. Let $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ be a stochastic instance with time horizon T; let $\tilde{f}_i \ge f_i$ and $\widetilde{\mathcal{I}} = \{\mathcal{A}, \tilde{\mathcal{F}}(i), \mathcal{L}\}$. Run Algorithm 5 with parameters $\{Q_{j,\phi}\}_{j\in[K],\phi\in[T]}$ on $\mathcal{I}$ and run Algorithm 5 with parameters $\left\{\widetilde{Q}_{j,\phi}\right\}_{j\in[K],\phi\in[T]}$ on $\widetilde{\mathcal{I}}$. Let $i_\phi^{\textsc{Alg}}$ and $\tilde{i}_\phi^{\textsc{Alg}}$ denote the arms pulled in the description of Algorithm 5 for the two instances, respectively. Then, the following two vector valued random variables are identically distributed: $(i_1^{\textsc{Alg}}, \ldots, i_{\Phi^*}^{\textsc{Alg}})$ and $(\tilde{i}_1^{\textsc{Alg}}, \ldots, \tilde{i}_{\Phi^*}^{\textsc{Alg}})$.*

The intuitive interpretation of Lemma C.5 is very natural: if we have two set of arms with identical loss distributions and run $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$ on them, we expect to see that the sequence of arms recommended by ALG is distributed identically across the two instances, even if we can't guarantee that the exact same arm is picked at every timestep on each instance. We provide a formal proof below.

Proof of Lemma C.5. Let $\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}$ denote possible loss sequences observed on $\mathcal{I}$ up to some $\psi \le \Phi^*$ and $\{\tilde{\ell}'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}$ denote possible loss sequences observed on $\widetilde{\mathcal{I}}$ up to the same ψ. Let us fix the bit of randomness b used for ALG on $\mathcal{I}$ to be the same as the bit of randomness used for ALG on $\widetilde{\mathcal{I}}$. Because of the way we have set $T = \max\{\sum_{\phi\in[\Phi^*]} \sum_{j\in[K]} Q_{j,\phi}, \sum_{\phi\in[\Phi^*]} \sum_{j\in[K]} \widetilde{Q}_{j,\phi}\}$, we are guaranteed that blocks

$\phi = 1, \ldots, \psi$ will have been reached on both $\widetilde{\mathcal{I}}$ and $\mathcal{I}$. Conditioned on b, let $F_b : [0,1]^{K\times\psi} \to [K]^\psi$ be the mapping from all $\ell'_{j,\phi}$ for $\phi \leq \psi$, to the sequence of arms it would have pulled correspondingly, that is,

$$F_b\left(\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right) = (i_1^{\textsc{Alg}}, i_2^{\textsc{Alg}}, \ldots, i_\psi^{\textsc{Alg}}).$$

Note that F_b does not depend on the feedback probabilities f_i or the random variables $Q_{i,\phi}$, because Alg is fully oblivious to these quantities. For any b, F_b is fully deterministic. Therefore, the distribution of $(i_1^{\textsc{Alg}}, i_2^{\textsc{Alg}}, \ldots, i_\psi^{\textsc{Alg}})$ is fully specified by the distributions of $\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}$, and the distribution of $(\tilde{i}_1^{\textsc{Alg}}, \tilde{i}_2^{\textsc{Alg}}, \ldots, \tilde{i}_\psi^{\textsc{Alg}})$ is fully specified by the distributions of $\{\tilde{\ell}'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}$.

Since the loss sequences are distributed identically across instances, we have that

$$\begin{aligned}
&\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]} \overset{d}{=} \{\tilde{\ell}'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\\
\implies\ & F_b\left(\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right) \overset{d}{=} F_b\left(\{\tilde{\ell}'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right)\\
\implies\ & (i_1^{\textsc{Alg}}, i_2^{\textsc{Alg}}, \ldots, i_\psi^{\textsc{Alg}}) \overset{d}{=} (\tilde{i}_1^{\textsc{Alg}}, \tilde{i}_2^{\textsc{Alg}}, \ldots, \tilde{i}_\psi^{\textsc{Alg}}),
\end{aligned}$$

where $\overset{d}{=}$ denotes identically distributed relationship. Finally, because this holds conditionally over any arbitrary b, we can integrate over all possible random bits b to establish the claim. ■

To use Algorithm 5 in our proofs, we need to argue that it makes decisions that are distributed identically to those of Algorithm 2. We formalize this below:

Lemma C.6 (Distribution of arms pulled by simulated algorithm)**.** *Fix an instance $\mathcal{I}$. Let $\{i_t^{orig}\}_{t\in[T]}$ be a sequence of random variables that represents the arms selected by Algorithm 2 on $\mathcal{I}$ over the time horizon T, and $\{i_t^{sim}\}_{t\in[T]}$ be a sequence of random variables that represents the arms selected by Algorithm 5 on an identical instance $\mathcal{I}$. Then the sequence $\{i_t^{orig}\}_{t\in[T]}$ is distributed identically to $\{i_t^{sim}\}_{t\in[T]}$.*

The key difference between Algorithm 5 and Algorithm 2 is that the number of times Algorithm 5 pulls $i_\phi^{\textsc{Alg}}$ is determined by the random variable $Q_{i_\phi,\phi}$, rather than by the first time feedback is observed in Algorithm 2. However, $Q_{i_\phi,\phi}$ is distributed identically to the number of times feedback will be observed, so the simulated version should overall produce the same distribution of outputs. We formalize this intuition below.

Proof of Lemma C.6. This proof will proceed in three main steps. First, we argue that the sequence of arms selected by Alg for either Algorithm 5 or Algorithm 2 are identically distributed. Second, we relate the $X_{j,t}$ used by Algorithm 2 to the ϕ timescale. Third, we show by induction that feedback observations are identically distributed on Algorithm 5 and Algorithm 2. Finally, we argue that the sequences of arms selected by Algorithm 5 and Algorithm 2 are identically distributed.

Step 1: Coupling arm pulls $i_\phi^{\textsc{Alg},\mathrm{orig}} = i_\phi^{\textsc{Alg},\mathrm{sim}}$.

Fix a sequence of random variables $Q_{j,\phi} \sim \mathrm{Geom}(f_j)$ for $\phi \in [T]$ and $j \in [K]$ used to run Algorithm 5. Let $T^* = \sum_{\phi\in[T]} \max_{j\in[K]} Q_{j,\phi}$; then fix a sequence of random variables $X_{j,t} \sim \mathrm{Bern}(f_j)$ for $t \in [T^*]$ and $j \in [K]$ that determine feedback observations in Algorithm 2.

We run Algorithm 2 and Algorithm 5 on identical copies of $\mathcal{I}$ for T^* rounds; we distinguish each copy by $\mathcal{I}^{\mathrm{orig}}$ for Algorithm 2 and $\mathcal{I}^{\mathrm{sim}}$ for Algorithm 5. We set T^* in this way to guarantee that timestep ϕ will be reached on both $\mathcal{I}^{\mathrm{orig}}$ and $\mathcal{I}^{\mathrm{sim}}$; we will handle truncation in the final step.

Recall that Algorithm 2 and Algorithm 5 both make calls to the same underlying Alg. Let b be the bit of randomness used for Alg in Algorithm 2 and Algorithm 5. Now, conditioning on b, let $F_b : [0,1]^{K\times\psi} \to [K]^\psi$ be, as defined before, the mapping from the sequence of possible losses that Alg may have observed for any arm at any time $\phi \leq \psi$, to the sequence of arms it would have pulled corresponding to those losses, that is,

$$F_b\left(\{\ell_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right) = (i_1^{\textsc{Alg}}, i_2^{\textsc{Alg}}, \ldots, i_\psi^{\textsc{Alg}}).$$

Note that F_b does not depend on the feedback probabilities f_i or the feedback observations $Q_{i,\phi}$ or $X_{i,t}$, because Alg is fully oblivious to these quantities. For any b, F_b is fully deterministic. Furthermore, the simulated and real algorithms use $\textsc{Alg}$ with the same bit of randomness, so $F_b^{\text{orig}} = F_b^{\text{sim}}$, and the arms selected by $\textsc{Alg}$ for either Algorithm 2 or Algorithm 5 are fully specified by the distributions of the losses for each arm. Then, for any $\psi \le T$, we have that

$$\begin{aligned}
&\{\ell_{j,\phi}^{\text{orig}}\}_{j\in[K],\phi\in[\psi]} \overset{d}{=} \{\ell_{j,\phi}^{\text{sim}}\}_{j\in[K],\phi\in[\psi]} && \textit{because } \mathcal{I}^{\text{orig}} = \mathcal{I}^{\text{sim}}\\
\implies\ & F_b\left(\{\ell_{j,\phi}^{\text{orig}}\}_{j\in[K],\phi\in[\psi]}\right) \overset{d}{=} F_b\left(\{\ell_{j,\phi}^{\text{sim}}\}_{j\in[K],\phi\in[\psi]}\right) && \textit{because } F_b^{\text{orig}} = F_b^{\text{sim}}\\
\implies\ & \{i_\phi^{\textsc{Alg},\text{orig}}\}_{\phi\in[\psi]} \overset{d}{=} \{i_\phi^{\textsc{Alg},\text{sim}}\}_{\phi\in[\psi]} && \textit{by definition of } F_b.
\end{aligned}$$

We end this step by creating a coupling between arms selected by $\textsc{Alg}$ so that $i_\phi^{\textsc{Alg},\text{orig}} = i_\phi^{\textsc{Alg},\text{sim}}$ for all $\phi \in [T]$. For the rest of the analysis, we condition on this particular sequence.

Step 2: Coupling the block lengths. Define $Q'_\phi : [K]^\phi \times \{0,1\}^{K\times T^*} \to \mathbb{N}$ to be the function that maps $\{i_{\phi'}^{\textsc{Alg},\text{orig}}\}_{\phi'\le\phi}$, the sequence of arms selected by $\textsc{Alg}$ up to ϕ, and the sequence of Bernoullis $X_{j,t}$ used to run Algorithm 2, to the number of times $i_\phi^{\textsc{Alg},\text{orig}}$ needs to be pulled (on t timescale) before feedback is observed.[14] In some abuse of notation, we let $Q'_\phi := Q'_\phi\left(\{i_{\phi'}^{\textsc{Alg},\text{orig}}\}_{\phi'\le\phi}, \{X_{j,t}\}_{j\in[K],t\ge1}\right)$ be shorthand for the number of times $i_\phi^{\textsc{Alg},\text{orig}}$ must be pulled until a feedback observation. Now, Q'_ϕ is fully determined by the history of $\textsc{Alg}$ arm pulls and the sequence of $X_{j,t}$. (Note that Q'_ϕ needs to depend on the $X_{j,t}$'s as well as the *history* of $\textsc{Alg}$'s selections. This is because even if we know $i_\phi^{\textsc{Alg}}$, we do not know which t indices of the $X_{i_\phi^{\textsc{Alg}},t}$ sequence determine whether we make an observation or not. We can only relate t to ϕ correctly if we know exactly which arms were pulled in previous $\phi' < \phi$, *and* their corresponding feedback observations.)

Next, we will show that for all $\phi \in [T]$, conditioned on fixing Q'_ψ and $Q_{i_\psi,\psi}$ such that $Q_{i_\psi,\psi} = Q'_\psi$ for all $\psi < \phi$, it holds that $Q_{i_\phi,\phi} \overset{d}{=} Q'_\phi$.

Recall that we have fixed a coupling between arms selected by $\textsc{Alg}$ so that $i_\phi^{\textsc{Alg},\text{orig}} = i_\phi^{\textsc{Alg},\text{sim}}$ for all $\phi \in [T]$. For ease of presentation, we refer to these arms as i_ϕ simply.

First note that for any ϕ, $Q_{i_\phi,\phi} \sim \text{Geom}(f_{i_\phi})$ by definition; furthermore, these are independent across all ϕ. To complete our claim, it suffices to show that $Q'_\phi \sim \text{Geom}(f_{i_\phi})$ conditioned on fixing Q'_ψ and $Q_{i_\psi,\psi}$ such that $Q_{i_\psi,\psi} = Q'_\psi$ for all $\psi < \phi$.

Recall that $Q'_\phi := Q'_\phi\left(\{i_{\phi'}\}_{\phi'\le\phi}, \{X_{j,t}\}_{j\in[K],t\ge1}\right)$ is the shorthand for the number of times i_ϕ must be pulled until a feedback observation. Let t_ϕ be the first time steps t that belongs to block ϕ. Note that t_ϕ is a deterministic function of the fixed variables $\{Q'_\psi\}_{\psi<\phi} = \{Q_{i_\psi,\psi}\}_{\psi<\phi}$. Furthermore, $X_{i_\phi,t}$s for $t \ge t_\phi$ are independent of $\{Q'_\psi\}_{\psi<\phi}$ and Q'_ϕ is a function of i_ϕ that only depends on $X_{i_\phi,t}$ for $t \ge t_\phi$. Moreover, $X_{i_\phi,t}$ for $t \ge t_\phi$ are Bernoulli random variables that are independent of t_ϕ. Therefore, $Q'_\phi \sim \text{Geom}(f_{i_\phi})$ conditioned on the past. That is, for all $\phi \in [T]$, it holds that $Q_{i_\phi,\phi} \overset{d}{=} Q'_\phi$ conditioned on fixing Q'_ψ and $Q_{i_\psi,\psi}$ such that $Q_{i_\psi,\psi} = Q'_\psi$ for all $\psi < \phi$.

We end this step by taking an adaptive coupling over the realizations of Q'_ψ and Q_j, ψ, such that for all $\phi \in [T]$, $Q_{i_\phi,\phi} = Q'_\phi$.

Step 3: Arms selected by Algorithm 2 and Algorithm 5 are identically distributed. Finally, we are ready to prove the main claim. Let us condition on the coupled sequence of arms pulled by $\textsc{Alg}$ from Step 1, $\{i_\phi^{\textsc{Alg}}\}_{\phi\in[T]}$, and the coupled feedback observations $\left(\{Q_{i_\phi,\phi}\}_{\phi\in[T]}, \{Q'_\phi\}_{\phi\in[T]}\right)$ from Step 2.

For Algorithm 2, the random variables $X_{j,t}$ and the sequence $\{i_\phi^{\textsc{Alg},\text{orig}}\}_{\phi\in[T]}$ fully specifies the arms pulled by 2, i.e. the sequence $\{i_t^{\text{orig}}\}_{t\in[T]}$. For Algorithm 5, the random variables $X_{j,\phi}$ and the sequence $\{i_\phi^{\textsc{Alg},\text{sim}}\}_{\phi\in[T]}$ fully specifies the arms pulled by 5, i.e. the sequence $\{i_t^{\text{sim}}\}_{t\in[T]}$. Conditioned on this coupling

[14] We note that while Algorithm 2 only takes $X_{j,t}$ variables into account for arm j that was pulled at time t, these random variables can be defined for all arms at all time steps without changing the behavior of Algorithm 2.

$\left(\{Q_{j,\phi}\}_{\phi\in[T]}, \{X_{j,t}\}_{t\geq 1}\right)$, therefore, we have that

$$\begin{aligned}
&\{i_\phi^{\textsc{Alg},\text{orig}}\}_{\phi\in[\psi]} \stackrel{d}{=} \{i_\phi^{\textsc{Alg},\text{sim}}\}_{\phi\in[\psi]} \\
\implies &\{i_t^{\text{orig}}\}_{t\in[T]} \stackrel{d}{=} \{i_t^{\text{sim}}\}_{t\in[T]} \qquad\qquad \textit{from conditioning on coupling,}
\end{aligned}$$

i.e. that the distribution of arms pulled by Algorithm 2 and Algorithm 5 are identically distributed, conditioned on the coupling. Truncating ψ to T for both algorithms also preserves the identical distribution. The main claim of the lemma follows by another application of the law of total probability.

■

C.2.2 Regret of BB_{Pull}: Proof of Theorem 3.3

We prove Theorem 3.3, restated below.

Theorem 3.3 (Regret BB_{Pull}). *Let* Alg *be any algorithm for the deterministic feedback setting that achieves regret at most* $R_{\textsc{Alg}}(T)$ *for stochastic losses. Then, for stochastic losses,* $\text{BB}_{\text{Pull}}(\textsc{Alg})$ *achieves regret at most* $R_{\text{BB}_{\text{Pull}}(\textsc{Alg})}(T) \leq R_{\textsc{Alg}}(T) \cdot \frac{1}{\min_i f_i}$.

The intuition is that in expectation, the number of times that an arm is pulled in $\text{BB}_{\text{Pull}}(\textsc{Alg})$ before feedback is observed is at most $1/\min_i f_i$. This means that we can upper bound the regret of $\text{BB}_{\text{Pull}}(\textsc{Alg})$ as $1/\min_i f_i$ times the regret of Alg.

Proof of Theorem 3.3 (Regret of BB_{Pull}*).* Recall that the regret guarantees for BB_{Pull} apply only to stochastic losses. To relate the regret of $\text{BB}_{\text{Pull}}(\textsc{Alg})$ to the regret of Alg, we consider the outputs of Alg while $\text{BB}_{\text{Pull}}(\textsc{Alg})$ is evaluated. Recall that Φ is the number of times that Alg is called. Note that the simulated version of $\text{BB}_{\text{Pull}}(\textsc{Alg})$, Algorithm 5, is run with the set of random variables $Q_{j,\phi}$ for $j\in[K]$ and $\phi\in[T]$, such that $Q_{j,\phi}\sim\text{Geom}(f_j)$, independently. Here, $Q_{i_\phi^{\textsc{Alg}},\phi}$ denotes the number of times of arm $i_\phi^{\textsc{Alg}}$ is pulled until feedback is observed. Recall that $\bar{\ell}_i$ denotes the mean loss of arm i and let $i^* = \arg\min_i \bar{\ell}_i$ be the arm with optimal expected loss. The (pseudo-)regret of $\text{BB}_{\text{Pull}}(\textsc{Alg})$ can be expressed as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\bar{\ell}_{i_t}\right] - \min_i \sum_{t=1}^{T}\bar{\ell}_i &= \mathbb{E}\left[\sum_{\phi=1}^{\Phi}\sum_{i\in\mathcal{A}} \mathbb{1}[i_\phi^{\textsc{Alg}} = i]\cdot Q_{i,\phi}\cdot\left(\bar{\ell}_i - \bar{\ell}_{i^*}\right)\right] \\
&\leq \mathbb{E}\left[\sum_{\phi=1}^{T}\sum_{i\in\mathcal{A}} \mathbb{1}[i_\phi^{\textsc{Alg}} = i]\cdot Q_{i,\phi}\cdot\left(\bar{\ell}_i - \bar{\ell}_{i^*}\right)\right] \\
&= \mathbb{E}\left[\sum_{\phi=1}^{T}\sum_{i\in\mathcal{A}} \mathbb{1}[i_\phi^{\textsc{Alg}} = i]\cdot \mathbb{E}[Q_{i,\phi}]\cdot\left(\bar{\ell}_i - \bar{\ell}_{i^*}\right)\right] \\
&\leq \frac{1}{\min_i f_i}\underbrace{\mathbb{E}\left[\sum_{\phi=1}^{T}\sum_{i\in\mathcal{A}} \mathbb{1}[i_\phi^{\textsc{Alg}} = i]\cdot\left(\bar{\ell}_i - \bar{\ell}_{i^*}\right)\right]}_{(1)},
\end{aligned}$$

where the second transition follows by noting that, as described above, Algorithm 5 is run with well-defined variables $Q_{i,\phi}\geq 0$ for all $\phi\leq T$ and $\bar{\ell}_i - \bar{\ell}_{i^*}\geq 0$ for all $i\in\mathcal{A}$, so that we can extend the summation to $\phi\in(\Phi,T]$. In the third transition, the outer expectation is over Alg and the inner expectation is over the feedback observations. And the last transition uses $\mathbb{E}[Q_{i,\phi}] = \frac{1}{f_i}\leq\frac{1}{\min_i f_i}$.

To relate (1) to the regret of Alg, we observe that in $\text{BB}_{\text{Pull}}(\textsc{Alg})$, the algorithm Alg also receives stochastic losses with mean $\bar{\ell}_i$ when it pulls $i_\phi^{\textsc{Alg}} = i$ that are identically distributed as in the original instance $\mathcal{I}$. This means that (1) is exactly equal to the regret of Alg in an instance with stochastic losses over T time steps. This completes the proof. ■

C.2.3 Monotonicity of $\mathrm{BB}_{\mathrm{Pull}}$: Proof of Theorem 3.4

Here, we formalize the coupling argument which will allow us to show positive feedback monotonicity in FOC and negative feedback monotonicity in APC for $\mathrm{BB}_{\mathrm{Pull}}$ applied to an underlying algorithm $\textsc{Alg}$. A very similar approach will be used to prove Theorems 3.6, 4.2, and 4.3 in the following sections, though those arguments will require a slightly more complex conditioning step.

For reference, we restate the result below.

Theorem 3.4. *[Impact of* $\mathrm{BB}_{\mathrm{Pull}}$ *on* APC *and* FOC *] Fix an instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ *with stochastic losses. Let* $\widetilde{f}_i \geq f_i$*, and let* $\widetilde{\mathcal{I}} = \{\mathcal{A}, \widetilde{\mathcal{F}}(i), \mathcal{L}\}$*. For any algorithm* $\textsc{Alg}$ *for the deterministic feedback setting, the algorithm* $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$ *satisfies* $\mathtt{APC}_i(\mathcal{I}) \geq \mathtt{APC}_i(\widetilde{\mathcal{I}})$*and* $\mathtt{FOC}_i(\mathcal{I}) \leq \mathtt{FOC}_i(\widetilde{\mathcal{I}})$*.*

We are now ready to proceed with the main coupling argument.

Proof of Theorem 3.4 (Monotonicity of $\mathrm{BB}_{\mathrm{Pull}}$*).* Fix an instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ with stochastic losses. Let $\tilde{f}_i \geq f_i$, and let $\widetilde{\mathcal{I}} = \{\mathcal{A}, \mathcal{F}(i), \mathcal{L}\}$. We will denote the time horizon of the transformed algorithm on $\mathcal{I}$ as Φ, as before, and the time horizon of the transformed algorithm on $\widetilde{\mathcal{I}}$ as $\widetilde{\Phi}$. We will analyze $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$ by comparing the behavior of Algorithm 5 on $\mathcal{I}$ and on $\widetilde{\mathcal{I}}$ in three steps as follows:

1. We construct a probability coupling between the sequence of random variables $Q_{j,\phi}$ and $\widetilde{Q}_{j,\phi}$ for $j \in [K]$ and $\phi = 1, \ldots, \infty$. This coupling ensures that $Q_{i,\phi} \geq \widetilde{Q}_{i,\phi}$ for arm i and $Q_{j,\phi} = \widetilde{Q}_{j,\phi}$ for all other arms $j \neq i$, for all ϕ.[15]

2. We call Algorithm 5 on $\mathcal{I}$ and $\widetilde{\mathcal{I}}$ with $Q_{j,\phi}$ and $\widetilde{Q}_{j,\phi}$ for $j \in [K]$ and $\phi = 1, \ldots, \infty$, respectively. Using Lemma C.5, we argue that for any Φ^*, $(i_1^{\textsc{Alg}}, \ldots, i_{\Phi^*}^{\textsc{Alg}}) \overset{d}{=} (\tilde{i}_1^{\textsc{Alg}}, \ldots, \tilde{i}_{\Phi^*}^{\textsc{Alg}})$; then, we couple the arm pulls on each instance so $(i_1^{\textsc{Alg}}, \ldots, i_{\Phi^*}^{\textsc{Alg}}) = (\tilde{i}_1^{\textsc{Alg}}, \ldots, \tilde{i}_{\Phi^*}^{\textsc{Alg}})$.

3. By this step, random variables $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, $i_\phi^{\textsc{Alg}}$, and $\tilde{i}_\phi^{\textsc{Alg}}$ are fixed according to the above coupling. As a final step, we modify step 2 so that Algorithm 5 terminates after T rounds. In this case, $\textsc{Alg}$ may be called a different number of times, $\Phi \leq \widetilde{\Phi}$, on instance $\mathcal{I}$ and $\widetilde{\mathcal{I}}$. We handle this by showing that this impacts the monotonicity in the claimed direction.

Step 1: Coupling realizations of feedback observations. Note that for $\tilde{f}_i > f_i$, the distribution of $\widetilde{Q}_{i,\phi}$ is stochastically dominated by $Q_{i,\phi}$. That is, as the feedback probability increases, we need fewer pulls to observe feedback when that arm is pulled. Therefore, there is a joint probability distribution over $(Q_{j,\phi}, \widetilde{Q}_{j,\phi})$ such that for all ϕ, with probability 1 the following hold: $Q_{i,\phi} \geq \widetilde{Q}_{i,\phi}$ and for all $j \neq i$, $Q_{j,\phi} = \widetilde{Q}_{j,\phi}$. This also gives us a coupling, that is a joint distribution, over $\left(\{Q_{j,\phi}\}_{j\in[K],\phi\in\{1,\ldots,\infty\}}, \{\widetilde{Q}_{j,\phi}\}_{j\in[K],\phi\in\{1,\ldots,\infty\}}\right)$ that meets the aforementioned property. (See Footnote 15 about dealing with infinitely long sequences.)

Step 2: Coupling arms pulled by Alg across instances $\mathcal{I}$ and $\widetilde{\mathcal{I}}$. We next consider Algorithm 5 on two instances $\mathcal{I}$ and $\widetilde{\mathcal{I}}$ using the coupled sequence of random variables $Q_{j,\phi}$ and $\widetilde{Q}_{j,\phi}$ for $j \in [K]$ and $\phi = 1, \ldots, \infty$, respectively, as coupled in in Step 1. Conditioned on these sequences, we now apply Lemma C.5. Note that the preconditions of this lemma are met for any Φ^*, so we have that $(i_1^{\textsc{Alg}}, \ldots, i_{\Phi^*}^{\textsc{Alg}})$ and $(\tilde{i}_1^{\textsc{Alg}}, \ldots, \tilde{i}_{\Phi^*}^{\textsc{Alg}})$ are identically distributed. This allows us to consider a joint probability distribution over $(i_1^{\textsc{Alg}}, \ldots, i_{\Phi^*}^{\textsc{Alg}}, \tilde{i}_1^{\textsc{Alg}}, \ldots, \tilde{i}_{\Phi^*}^{\textsc{Alg}})$ such that $i_\phi^{\textsc{Alg}} = \tilde{i}_\phi^{\textsc{Alg}}$ for all $\phi \in [\Phi^*]$.

Step 3: Handling different stopping times. We now have random variables $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, $i_\phi^{\textsc{Alg}}$, $\tilde{i}_\phi^{\textsc{Alg}}$ are all fixed and for all $\phi = 1, \cdots, \infty$ satisfy $i_\phi^{\textsc{Alg}} = \tilde{i}_\phi^{\textsc{Alg}}$, $Q_{i,\phi} \geq \widetilde{Q}_{i,\phi}$, and $Q_{j,\phi} = \widetilde{Q}_{j,\phi}$ for $j \neq i$.

We next consider the actual performance of Algorithm 5 on instances $\mathcal{I}$ and $\widetilde{\mathcal{I}}$ over T timesteps. Note that this is exactly the same as history of arms played by Algorithm 5 on $\mathcal{I}$ and $\widetilde{\mathcal{I}}$, respectively, in Step 2 of the analysis, except that the algorithm now terminates at time T. Therefore, the number of rounds $\textsc{Alg}$ is called in each of these two instances may be different. Notice that Φ and $\widetilde{\Phi}$ are deterministic variables, since the arms pulled and the number of rounds until an observation is made are all fixed. It is not hard to see that

[15] Constructing an infinitely long sequence is only for convenience in using Lemma C.5; we only consume at most T of these random variables in any algorithm for analysis.

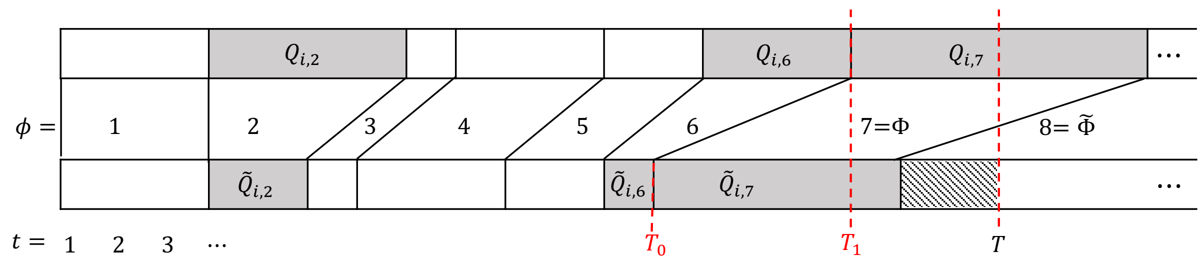

Figure 1: Timelines of $\text{BB}_{\text{Pull}}(\textsc{Alg})$ on instances $\mathcal{I}$ (top row) and $\widetilde{\mathcal{I}}$ (bottom row) are demonstrated. Each time step $t \in [T]$ maps to a block number in $\mathcal{I}$ that is no more than its block number in $\widetilde{\mathcal{I}}$. The total number of times $\textsc{Alg}$ is called in instance $\mathcal{I}$, Φ, and the number of times it is called in $\widetilde{\mathcal{I}}$, $\widetilde{\Phi}$, satisfy $\Phi \leq \widetilde{\Phi}$.

$\Phi \leq \widetilde{\Phi}$. This is perhaps best seen by considering Figure 1. We note that the time horizon of $\text{BB}_{\text{Pull}}(\textsc{Alg})$ for the two instances can be thought of as two sequence of blocks $[\Phi]$ and $[\widetilde{\Phi}]$. For each $\phi \leq \min\{\Phi, \widetilde{\Phi}\}$, $i_\phi^{\textsc{Alg}} = \tilde{i}_\phi^{\textsc{Alg}}$. Therefore, the only case where $Q_{i_\phi^{\textsc{Alg}},\phi} \neq \widetilde{Q}_{i_\phi^{\textsc{Alg}},\phi}$ is when $i_\phi^{\textsc{Alg}} = i$; these are shown by gray blocks in Figure 1. In this case $Q_{i_\phi^{\textsc{Alg}},\phi} \geq \widetilde{Q}_{i_\phi^{\textsc{Alg}},\phi}$ by the coupling we designed above. In all other blocks, where $i_\phi^{\textsc{Alg}} \neq i$, we have that $Q_{i_\phi^{\textsc{Alg}},\phi} = \widetilde{Q}_{i_\phi^{\textsc{Alg}},\phi}$. Let $\phi(t)$ (resp. $\widetilde{\phi}(t)$) be the function that maps timesteps on the timescale indexed by t to timesteps on $\textsc{Alg}$'s timescale on $\mathcal{I}$ (resp. $\widetilde{\mathcal{I}}$). We can now see that every time step $t \in [T]$ maps to blocks $\phi(t)$ and $\widetilde{\phi}(t)$ in instances $\mathcal{I}$ and $\widetilde{\mathcal{I}}$, respectively, such that $\widetilde{\phi}(t) \geq \phi(t)$. This implies that $\Phi \leq \widetilde{\Phi}$, because $\phi(T) \leq \widetilde{\phi}(T)$.

Notation for Analyzing `FOC` and `APC`. The remainder of the proof boils down to analyzing `FOC` and `APC` on $\mathcal{I}$ and $\widetilde{\mathcal{I}}$. We use the coupling thus far with the property that $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, $i_\phi^{\textsc{Alg}}$, $\tilde{i}_\phi^{\textsc{Alg}}$ are all fixed and for all $\phi = 1, \cdots, \infty$ satisfy $i_\phi^{\textsc{Alg}} = \tilde{i}_\phi^{\textsc{Alg}}$, $Q_{i,\phi} \geq \widetilde{Q}_{i,\phi}$, and $Q_{j,\phi} = \widetilde{Q}_{j,\phi}$ for $j \neq i$. Figure 1 provides an intuitive proof of the desired claims.

To formalize these claims, we introduce the following additional notation. Given a range $R \subseteq [T]$, let $\texttt{FOC}_i^R(\mathcal{I})$ be the number of times feedback is observed on arm i in timesteps in R on $\widetilde{\mathcal{I}}$, and let $\texttt{FOC}_i^R(\widetilde{\mathcal{I}})$ be the number of times feedback is observed on arm i in timesteps in R on $\tilde{\mathcal{I}}$. Similarly, let $\texttt{APC}_i^R(\mathcal{I})$ be the number of times arm i is pulled in timesteps in R on $\mathcal{I}$, and let $\texttt{APC}_i^R(\widetilde{\mathcal{I}})$ be the number of times arm i is pulled in timesteps in R on $\widetilde{\mathcal{I}}$. Since we have conditioned on $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, $i_\phi^{\textsc{Alg}}$, $\tilde{i}_\phi^{\textsc{Alg}}$, we see that at this point $\texttt{FOC}_i^R(\mathcal{I})$, $\texttt{FOC}_i^R(\widetilde{\mathcal{I}})$, $\texttt{APC}_i^R(\mathcal{I})$, and $\texttt{APC}_i^R(\widetilde{\mathcal{I}})$ are all deterministic.

Since we will analyze the last time block separately, we let $T_1 = \sum_{\phi \in [\Phi-1]} Q_{i_\phi^{\textsc{Alg}},\phi} \leq T$ be the time step referring to the penultimate block of $\text{BB}_{\text{Pull}}(\textsc{Alg})$ on $\mathcal{I}$. We let T_0 be the corresponding time step on instance $\widetilde{\mathcal{I}}$ defined by $T_0 = \sum_{\phi \in [\Phi-1]} \widetilde{Q}_{i_\phi^{\textsc{Alg}},\phi}$ (note that the expression sums over $\phi \in [\Phi - 1]$, and not over $\phi \in [\tilde{\Phi} - 1]$). By definition, it holds that $T_0 \leq T_1$.

Analyzing `FOC`. We first prove that $\texttt{FOC}_i^{[T]}(\tilde{\mathcal{I}}) - \texttt{FOC}_i^{[T]}(\mathcal{I}) \geq 0$. First, we observe that:

$$\texttt{FOC}_i^{[T_1]}(\mathcal{I}) = \sum_{\phi \in [\Phi-1]} \mathbb{1}[i_\phi^{\textsc{Alg}} = i] = \texttt{FOC}_i^{[T_0]}(\tilde{\mathcal{I}}).$$

It thus suffices to show that:

$$\texttt{FOC}_i^{[T]}(\mathcal{I}) - \texttt{FOC}_i^{[T_1]}(\mathcal{I}) \leq \texttt{FOC}_i^{[T]}(\tilde{\mathcal{I}}) - \texttt{FOC}_i^{[T_0]}(\tilde{\mathcal{I}}).$$

For ease of exposition, we now consider two cases.

- Case 1: The Φth (last) block of $\mathcal{I}$ pulls $j \neq i$, i.e., $i_\Phi^{\textsc{Alg}} \neq i$. In this case, we have that:

$$\texttt{FOC}_i^{[T]}(\mathcal{I}) - \texttt{FOC}_i^{[T_1]}(\mathcal{I}) = 0 \leq \texttt{FOC}_i^{[T]}(\mathcal{I}) - \texttt{FOC}_i^{[T_0]}(\tilde{\mathcal{I}}).$$

- Case 2: i was pulled in the Φth block of $\mathcal{I}$, i.e., $i_\Phi^{\textsc{Alg}} = i$. In this case, we have that:

$$\texttt{FOC}_i^{[T]}(\mathcal{I}) - \texttt{FOC}_i^{[T_1]}(\mathcal{I}) = \mathbb{1}\left[Q_{i,\Phi} \le T - T_1\right] \le \mathbb{1}\left[\tilde{Q}_{i,\Phi} \le T - T_1\right] \le \mathbb{1}\left[\tilde{Q}_{i,\Phi} \le T - T_0\right] \le \texttt{FOC}_i^{[T]}(\tilde{\mathcal{I}}) - \texttt{FOC}_i^{[T_0]}(\tilde{\mathcal{I}}),$$

  as desired.

These two cases prove that $\texttt{FOC}_i^{[T]}(\tilde{\mathcal{I}}) - \texttt{FOC}_i^{[T]}(\mathcal{I}) \ge 0$.

Taking an expectation over $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, $i_\phi^{\textsc{Alg}}$, $\tilde{i}_\phi^{\textsc{Alg}}$, we see that:

$$\texttt{FOC}_i(\tilde{\mathcal{I}}) - \texttt{FOC}_i(\mathcal{I}) = \mathbb{E}\left[\texttt{FOC}_i^{[T]}(\tilde{\mathcal{I}}) - \texttt{FOC}_i^{[T]}(\mathcal{I})\right] \ge 0.$$

Analyzing APC. We first prove that $\texttt{APC}_i^{[T]}(\tilde{\mathcal{I}}) - \texttt{APC}_i^{[T]}(\mathcal{I}) \le 0$. We claim that

$$T_1 - T_0 = \sum_{\phi \in [\Phi - 1]} \mathbb{1}(i_\phi^{\textsc{Alg}} = i)\left(Q_{i,\phi} - \widetilde{Q}_{i,\phi}\right). \tag{2}$$

This is due to the fact that, as discussed above, the only case where $Q_{i_\phi^{\textsc{Alg}},\phi} \ne \widetilde{Q}_{i_\phi^{\textsc{Alg}},\phi}$ is when $i_\phi^{\textsc{Alg}} = i$ (these are shown by gray blocks in Figure 1) in which case $Q_{i_\phi^{\textsc{Alg}},\phi} \ge \widetilde{Q}_{i_\phi^{\textsc{Alg}},\phi}$. In all other cases, $Q_{i_\phi^{\textsc{Alg}},\phi} = \widetilde{Q}_{i_\phi^{\textsc{Alg}},\phi}$. Equation (2) implies that

$$\texttt{APC}_i^{[T_1]}(\mathcal{I}) - \texttt{APC}_i^{[T_0]}(\widetilde{\mathcal{I}}) = T_1 - T_0. \tag{3}$$

For ease of exposition, we now consider two cases.

- Case 1: Φth block of $\mathcal{I}$ pulls $j \ne i$, i.e., $i_\Phi^{\textsc{Alg}} \ne i$. In this case, we have that $\texttt{APC}_i^{[T]}(\mathcal{I}) = \texttt{APC}_i^{[T_1]}(\mathcal{I})$. Moreover, within the last $T - T_0$ timesteps of $\widetilde{\mathcal{I}}$ at least $T - T_1$ are dedicated to pulling arm $j \ne i$ in the Φth block of $\widetilde{\mathcal{I}}$. Thus,

$$\texttt{APC}_i^{[T]}(\widetilde{\mathcal{I}}) \le \texttt{APC}_i^{[T_0]}(\widetilde{\mathcal{I}}) + T - T_0 - (T - T_1) = \texttt{APC}_i^{[T_0]}(\widetilde{\mathcal{I}}) + T_1 - T_0 = \texttt{APC}_i^{[T_1]}(\mathcal{I}) = \texttt{APC}_i^{[T]}(\mathcal{I}),$$

  where the second to last equality is by Equation (3).

- Case 2: i was pulled in the Φth block of $\mathcal{I}$, i.e., $i_\Phi^{\textsc{Alg}} = i$. In this case, we have that $\texttt{APC}_i^{[T]}(\mathcal{I}) = \texttt{APC}_i^{[T_1]}(\mathcal{I}) + T - T_1$. Furthermore,

$$\texttt{APC}_i^{[T]}(\widetilde{\mathcal{I}}) \le \texttt{APC}_i^{[T_0]}(\widetilde{\mathcal{I}}) + T - T_0 = \texttt{APC}_i^{[T_1]}(\mathcal{I}) + T - T_1 = \texttt{APC}_i^{[T]}(\mathcal{I}),$$

  where the second equation is by Equation (3).

These two cases prove that $\texttt{APC}_i^{[T]}(\tilde{\mathcal{I}}) - \texttt{APC}_i^{[T]}(\mathcal{I}) \le 0$. Taking an expectation over $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, $i_\phi^{\textsc{Alg}}$, $\tilde{i}_\phi^{\textsc{Alg}}$, we see that:

$$\texttt{APC}_i(\tilde{\mathcal{I}}) - \texttt{APC}_i(\mathcal{I}) = \mathbb{E}\left[\texttt{APC}_i^{[T]}(\widetilde{\mathcal{I}}) - \texttt{APC}_i^{[T]}(\mathcal{I})\right] \le 0.$$

This completes the proof. ■

C.3 Proofs for Section 3.3: $\mathrm{BB}_{\mathrm{DA}}$

To analyze $\mathrm{BB}_{\mathrm{DA}}$, we will combine the approaches of our analyses for $\mathrm{BB}_{\mathrm{Pull}}$ and $\mathrm{BB}_{\mathrm{Divide}}$. For regret, we will analyze the the per-block regret; for monotonicity, we will make a coupling argument. For both we will analyze a simulated version of $\mathrm{BB}_{\mathrm{DA}}$, which we present in the following section.

We restate the algorithm below to clarify the dependence on the input $f^\star \in (0, \min_i f_i]$.

Algorithm 3: $\text{BB}\textsc{da}(\textsc{Alg}, f^\star)$

Begin with $\phi = 1$ and $t = 1$.
while $t \leq T$ **do**
  Let $i_\phi^{\textsc{Alg}} = \textsc{Alg}(\phi)$, $B_\phi = \lceil \frac{3 \ln T}{f^\star}(1 + f_{i_\phi^{\textsc{Alg}}}) \rceil$, and $S_\phi = \{t, t+1, \ldots, \min(t + B_\phi, T)\}$.
  for $t \in S_\phi$ **do**
    Pull $i_\phi^{\textsc{Alg}}$, i.e. $i_t = i_\phi^{\textsc{Alg}}$, and let $t \leftarrow t + 1$.
  if $\exists t \in S_\phi$ *s.t.* $X_{i_t,t} = 1$ *(i.e. there are observations)* **then**
    Return a random observation to $\textsc{Alg}$, i.e. $\ell_{i_\phi^{\textsc{Alg}},\phi} \sim \text{Unif}\{\ell_{i_t,t} : X_{i_t,t} = 1, t \in S_\phi\}$.
  else Return a loss of 1 to $\textsc{Alg}$, i.e. $\ell_{i_\phi^{\textsc{Alg}},\phi} = 1$.
Update $\phi \leftarrow \phi + 1$.

C.3.1 Constructing a simulated version of BB_{DA}

As before, we construct a simulated version of $\text{BB}_{\text{DA}}(\textsc{Alg})$. Again, we will define a sequence of random variables that determine how $\text{BB}_{\text{DA}}(\textsc{Alg})$ will proceed on $\mathcal{I}$ and $\widetilde{\mathcal{I}}$, and simulate a statistically indistinguishable version of $\text{BB}_{\text{DA}}(\textsc{Alg})$ in Algorithm 6. Again, we will index the time horizon with $\textsc{Alg}$ with ϕ.

- *Losses:* For each round $\phi \in [\Phi]$ of $\textsc{Alg}$ and each arm $j \in [K]$, $\ell'_{j,\phi}$ is the placeholder for the loss passed to $\textsc{Alg}$ if $\textsc{Alg}$ were to observe the loss of arm j at time ϕ. Since we are in the stochastic loss setting, $\ell'_{j,\phi}$ is a random variable drawn from the distribution of arm j (with mean $\bar{\ell}_j$) independently across ϕ and j, if at least one observation is realized in block ϕ, and $\ell'_{j,\phi} = 1$ otherwise. We note that these losses are only observed up to timestep Φ (which is a random variable less than T) and only for the specific arms pulled by the algorithm.
- *Feedback probabilities:* for each arm $j \in [K]$ and $\phi \in [T]$, let $U_{j,\phi} \sim \text{Bern}\left(1 - (1 - f_j)^{B_j}\right)$ denote the indicator variable for whether feedback will be observed in block ϕ, where $B_j = \lceil \frac{3 \ln(T)}{f^\star}(1 + f_j) \rceil$, for $f^\star \in (0, \min_i f_i]$.

Algorithm 6: Simulated version of $\text{BB}_{\text{DA}}(\textsc{Alg}, f^\star)$

Input: A sequence of integers in $\{0, 1\}$, $U_{j,\phi}$ for $\phi \in [T]$ and $j \in [K]$; $f^\star \in (0, \min_i f_i]$
Initialize $\phi = 1$.
For each arm $j \in [K]$, set $B_j = \lceil \frac{3 \ln(T)}{f^\star}(1 + f_j) \rceil$.
while $t \leq T$ **do**
  Let $i_\phi^{\textsc{Alg}} = \textsc{Alg}(\phi)$ be the output of $\textsc{Alg}$ at timestep ϕ.
  Let $S_\phi = \{t, t+1, \ldots, \min(t + B_{i_\phi^{\textsc{Alg}}}, T)\}$.
  for $t' \in S_\phi$ **do**
    Pull $i_\phi^{\textsc{Alg}}$, i.e. $i_{t'} = i_\phi^{\textsc{Alg}}$, and let $t \leftarrow t + 1$.
  if $U_{i_\phi^{\textsc{Alg}},\phi} = 1$ **then**
    Observe and return $\ell'_{i_\phi^{\textsc{Alg}},\phi} := \ell_{i_t,t}$ to $\textsc{Alg}$.
  else
    Return $\ell'_{i_\phi^{\textsc{Alg}},\phi} = 1$ to $\textsc{Alg}$.
  Let $\phi \leftarrow \phi + 1$.

Lemma C.7. *For each arm $j \in [K]$, set $B_j = \lceil \frac{3 \ln(T)}{\min_i f_i}(1 + f_j) \rceil$. Let Φ^* be any positive integer and $T = \Phi^* \cdot \max_j B_j$ be the time horizon. Let $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ be a stochastic instance with time horizon T; let $\tilde{f}_i \geq f_i$ and $\widetilde{\mathcal{I}} = \{\mathcal{A}, \tilde{\mathcal{F}}(i), \mathcal{L}\}$. Let $U_{j,\phi} = \widetilde{U}_{j,\phi}$ for $j \in [K]$ and $\phi \in [T]$. Run Algorithm 6 with parameters $\{U_{j,\phi}\}_{j \in [K], \phi \in [T]}$ on $\mathcal{I}$ and run Algorithm 6 with parameters $\left\{\widetilde{U}_{j,\phi}\right\}_{j \in [K], \phi \in [T]}$ on $\widetilde{\mathcal{I}}$. Let $i_\phi^{\textsc{Alg}}$ and $\tilde{i}_\phi^{\textsc{Alg}}$ denote the arms pulled in the description of Algorithm 6 for the two instances, respectively. Then, the following two vector valued random variables are identically distributed: $(i_1^{\textsc{Alg}}, \ldots, i_{\Phi^*}^{\textsc{Alg}})$ and $(\tilde{i}_1^{\textsc{Alg}}, \ldots, \tilde{i}_{\Phi^*}^{\textsc{Alg}})$.*

Proof. Let $\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}$ denote possible loss sequences observed on $\mathcal{I}$ up to some $\psi \leq \Phi^*$ and $\{\tilde{\ell}'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}$ denote possible loss sequences observed on $\widetilde{\mathcal{I}}$ up to the same ψ. Let us fix the bit of randomness b used for Alg on $\mathcal{I}$ to be the same as the bit of randomness used for Alg on $\widetilde{\mathcal{I}}$. Because we have set $T = \Phi^* \cdot \max_j B_j$, we are guaranteed that blocks $\phi = 1, \dots, \psi$ will have been reached on both $\tilde{\mathcal{I}}$ and $\mathcal{I}$. Conditioned on b, let $F_b : [0,1]^{K\times\psi} \to [K]^\psi$ be the mapping from all $\ell'_{j,\phi}$ for $\phi \leq \psi$, to the sequence of arms Alg would have pulled corresponding to those losses, that is,

$$F_b\left(\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right) = (i_1^{\text{ALG}}, i_2^{\text{ALG}}, \dots, i_\psi^{\text{ALG}}).$$

Note that F_b does not depend on the feedback probabilities f_i, because Alg is fully oblivious to these quantities. For any b, F_b is fully deterministic. Therefore, the distribution of $(i_1^{\text{ALG}}, i_2^{\text{ALG}}, \dots, i_\psi^{\text{ALG}})$ is fully specified by the distributions of $\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}$, and the distribution of $(\tilde{i}_1^{\text{ALG}}, \tilde{i}_2^{\text{ALG}}, \dots, \tilde{i}_\psi^{\text{ALG}})$ is fully specified by the distributions of $\{\tilde{\ell}'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}$.

In our specification of Algorithm 6, the sequences of losses passed to Alg are determined not only by the underlying loss distributions for each arm selected i_t, but also by the random variables $U_{j,\phi}$ which determine whether Alg will observe $\ell_{i_t,t}$ (which is actually sampled from the distribution of the selected arm i_t), or a loss of 1. Conditioning on $U_{j,\phi} = \widetilde{U}_{j,\phi}$ for all j and ϕ gives us that the loss sequences are distributed identically across instances. Therefore, we have that

$$\begin{aligned}
&\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]} \overset{d}{=} \{\tilde{\ell}'_{j,\phi}\}_{j\in[K],\phi\in[\psi]} \\
\implies\ & F_b\left(\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right) \overset{d}{=} F_b\left(\{\tilde{\ell}'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right) \\
\implies\ & (i_1^{\text{ALG}}, i_2^{\text{ALG}}, \dots, i_\psi^{\text{ALG}}) \overset{d}{=} (\tilde{i}_1^{\text{ALG}}, \tilde{i}_2^{\text{ALG}}, \dots, \tilde{i}_\psi^{\text{ALG}}),
\end{aligned}$$

where $\overset{d}{=}$ denotes identically distributed relationship. Finally, because this holds conditionally over any arbitrary b, we can integrate over all possible random bits b to establish the claim. ■

Lemma C.8. *Fix an instance $\mathcal{I}$. Let $\{i_t^{orig}\}_{t\in[T]}$ be a sequence of random variables that represents the arms selected by Algorithm 3 on $\mathcal{I}$ over the time horizon T, and $\{i_t^{sim}\}_{t\in[T]}$ be a sequence of random variables that represents the arms selected by Algorithm 6 on an identical instance $\mathcal{I}$. Then the sequence $\{i_t^{orig}\}_{t\in[T]}$ is distributed identically to $\{i_t^{sim}\}_{t\in[T]}$.*

The intuition for this lemma is similar to the proof of Lemma C.6; here, we argue that the likelihood that no feedback is observed at any block ϕ is identically distributed for both Algorithm 3 and Algorithm 6, and that taking one sample from the loss distribution (as Algorithm 6 does) is the same as taking a uniform sample out of several possible observations (as Algorithm 3 does).

Proof. We run Algorithm 3 and Algorithm 6 on identical copies of $\mathcal{I}$; we distinguish each copy by $\mathcal{I}^{\text{orig}}$ for Algorithm 3 and $\mathcal{I}^{\text{sim}}$ for Algorithm 6. In the first step, we introduce F_b which formalizes that arm selected by Alg given the random variables $\ell'_{j,\phi}$ defined earlier. We use this in the second step to show that arms selected by Alg are distributed the same across the two algorithms. In the last step, we use the fact that the block sizes are of equal lengths across the two algorithms to show that arms pulled by Algorithm 3 and Algorithm 6 are distributed the same.

Step 1: Formalize Alg arm selection. Recall that Algorithm 3 and Algorithm 6 both make calls to the same underlying Alg. Let $\ell'_{j,\phi}$ be, as defined earlier, the placeholder for losses passed to Alg, if Alg were to observe the loss of arm j at time ϕ. Let b be the bit of randomness used for Alg in Algorithm 3 and Algorithm 6. Now, conditioning on b, let $F_b : [0,1]^{K\times\psi} \to [K]^\psi$ be the mapping from all $\ell'_{j,\phi}$s up to time $\phi \leq \psi$, to the sequence of arms Alg would have pulled corresponding to those losses, that is,

$$F_b\left(\{\ell_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right) = (i_1^{\text{ALG}}, i_2^{\text{ALG}}, \dots, i_\psi^{\text{ALG}}).$$

Note that F_b does not depend on the feedback probabilities f_i or the feedback observations $Q_{i,\phi}$ or $X_{i,t}$, because Alg is fully oblivious to these quantities. For any b, F_b is fully deterministic. Furthermore, the simulated and real algorithms use ALG with the same bit of randomness, so $F_b^{\text{orig}} = F_b^{\text{sim}}$, and the arms selected by ALG for either Algorithm 3 and Algorithm 6 are fully specified by the distributions of the losses for each arm.

Step 2: Arms selected by Alg are distributed the same. We first establish that $\{\ell'_{j,\phi}\}_{j\in[K],\phi\in[\psi]}$ are identically distributed.

Recall that $\ell'_{j,\phi}$ are placeholders for losses of all arms j and round ϕ of ALG (although ALG only takes into account the random variables for arms it pulls).

By our specification of Algorithm 6, given j and ϕ, the event that $\ell'_{j,\phi}$ is drawn from the distribution of arm j is determined by $U_{j,\phi} \sim \text{Bern}(1-(1-f_j)^{B_j})$ and has probability probability $1-(1-f_j)^{B_j}$. And, with probability $(1-f_j)^{B_j}$, $\ell'_{j,\phi} = 1$.

For Algorithm 3, note that j will be pulled exactly B_j times in each block. The likelihood that at least at one of these round a loss is generated from arm j is exactly $1-(1-f_j)^{B_j}$. Note that in this case, $\ell'_{j,\phi}$ is drawn uniformly from the realized losses, which is equivalent to being drawn from the loss of arm j. And, with probability $(1-f_j)^{B_j}$, $\ell'_{j,\phi}$ is deterministically set to 1. Note that the realizations of $U_{j,\phi}$ and $X_{j,t}$ are all independent across ϕ, t, and K, so we have that

$$\{\ell'^{\text{orig}}_{j,\phi}\}_{j\in[K],\phi\in[\psi]} \overset{d}{=} \{\ell'^{\text{sim}}_{j,\phi}\}_{j\in[K],\phi\in[\psi]}.$$

Since F_b is a deterministic map, we have that

$$F_b\left(\{\ell'^{\text{orig}}_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right) \overset{d}{=} F_b\left(\{\ell'^{\text{sim}}_{j,\phi}\}_{j\in[K],\phi\in[\psi]}\right)$$
$$\{i_\phi^{\text{ALG,orig}}\}_{\phi\in[\psi]} \overset{d}{=} \{i_\phi^{\text{ALG,sim}}\}_{\phi\in[\psi]}$$

Step 3: Arms selected by Algorithm 3 and Algorithm 6 are identically distributed. Note that by the specification of each algorithm, for every i_ϕ selected by ALG, Algorithm 3 and Algorithm 6 will pull i_ϕ exactly B_{i_ϕ} times. Having steps 1 and 2 for $\psi > T$, gives us

$$\{i_\phi^{\text{ALG,orig}}\}_{\phi\in[\psi]} \overset{d}{=} \{i_\phi^{\text{ALG,sim}}\}_{\phi\in[\psi]}$$
$$\{i_t^{\text{orig}}\}_{t\in[T]} \overset{d}{=} \{i_t^{\text{sim}}\}_{t\in[T]}.$$

Applying the law of total expectation over possible random bits b proves the claim.

■

C.3.2 Regret of BB_{DA}: Proof of Theorem 3.5

First, we prove Theorem 3.5. The intuition is that we can bound the size of any block by $\max B_i \le \frac{6\ln T}{f^\star}$. Since B_i is sufficiently large, with high probability, there will be at least one observation in each block. Since the losses are stochastic, we can upper bound the regret of $\text{BB}_{\text{DA}}(\text{ALG})$ as $\max_j B_j \cdot R_{\text{ALG}}(T)$ as desired.

We restate the regret result of Theorem 3.5 below.

Theorem 3.5. *[Regret BB_{DA}] Let ALG be any algorithm for the deterministic feedback setting that achieves regret at most $R_{\text{ALG}}(T)$ when the losses are stochastic. Then, for stochastic losses, for any $f^\star \le \min_i f_i$, the algorithm $\text{BB}_{\text{DA}}(\text{ALG}, f^\star)$ achieves regret at most $R_{\text{BB}_{\text{DA}}(\text{ALG})}(T) \le \frac{6\ln T}{f^\star} R_{\text{ALG}}\left(\frac{Tf^\star}{3\ln T}\right)$.*

The argument requires Lemma C.9, which ensures that at least one observation from the true loss distribution is made in every block (note that this is very similar to the statement and proof of Lemma C.4, except that the block size B is no longer fixed).

Lemma C.9. *Fix an $f^\star \in (0, \min_i f_i]$, and let $\Phi \le T$. Divide the time horizon T into blocks of size $B_\phi \ge \frac{3\ln T}{f^\star}$ for $\phi \in \Phi$. Suppose then that for each block $\phi \in \{1, 2, \ldots, \Phi\}$, we play the same arm i_ϕ a total of B_ϕ times, i.e. for every round in block ϕ, as in Algorithms 3 and 6. Let E be the "*clean event*" that at least one feedback observation occurs in* each *block ϕ, i.e., that for all blocks ϕ, $\exists t \in S_\phi : X_{i_t,t} = 1$. Then, $\Pr[E] \ge 1 - 1/T^2$.*

Proof. Let E_ϕ be the event that at least one feedback observation occurred in block ϕ, i.e., $\exists t \in S_\phi : X_{i_t,t} = 1$. Since for any arm i, $\Pr[X_{i,t} = 1] = f_i$, then for arm i_ϕ, we have that

$$\Pr[\neg E_\phi] = (1 - f_{i_\phi})^B \leq (1 - f^\star)^B \leq \exp(-f^\star B) \leq 1/T^3.$$

Union bounding over all $\Phi \leq T$ blocks, we conclude that

$$\Pr[\neg E] \leq \sum_{\phi \in [\Phi]} \Pr[\neg E_\phi] \leq 1/T^2. \qquad \blacksquare$$

Proof of Theorem 3.5 (Regret of $\mathrm{BB}_{\mathrm{DA}}$*).* This argument proceeds almost identically to the proof of Theorem 3.1, the regret bound on $\mathrm{BB}_{\mathrm{Divide}}$, in the stochastic case. Recall that $f^\star \in (0, \min_j f_j]$. For notational convenience, let $B = \frac{3 \cdot \ln(T)}{f^\star}$. First, note that it must be the case that the size of any block B_i is bounded as follows, because $1 \leq 1 + f_i \leq 2$:

$$B \leq B_i \leq 2B.$$

Then, we will have at most $\lfloor T/B \rfloor$ blocks, and each block will incur at most $2B$ regret. We use Lemma C.9 to argue that we will see at least one feedback observation in each block with probability $1 - 1/T^2$; conditioned on this occurring, using the above bounds on the number of blocks and the size of each block, the (pseudo-)regret of $\mathrm{BB}_{\mathrm{DA}}(\textsc{Alg})$ can be expressed as

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \bar{\ell}_{i_t}\right] - \min_i \sum_{t=1}^{T} \bar{\ell}_i &= \mathbb{E}\left[\sum_{\phi=1}^{\Phi} \sum_{i \in \mathcal{A}} \mathbb{1}[i_\phi^{\textsc{Alg}} = i] \cdot B_{i_\phi} \cdot \left(\bar{\ell}_i - \bar{\ell}_{i^*}\right)\right] \\
&\leq \mathbb{E}\left[\sum_{\phi=1}^{\lfloor T/B \rfloor} \sum_{i \in \mathcal{A}} \mathbb{1}[i_\phi^{\textsc{Alg}} = i] \cdot B_{i_\phi} \cdot \left(\bar{\ell}_i - \bar{\ell}_{i^*}\right)\right] \\
&\leq \mathbb{E}\left[\sum_{\phi=1}^{\lfloor T/B \rfloor} \sum_{i \in \mathcal{A}} \mathbb{1}[i_\phi^{\textsc{Alg}} = i] \cdot \max_j B_j \cdot \left(\bar{\ell}_i - \bar{\ell}_{i^*}\right)\right] \\
&= \frac{6 \ln(T)}{f^\star} \cdot \underbrace{\mathbb{E}\left[\sum_{\phi=1}^{\lfloor T/B \rfloor} \sum_{i \in \mathcal{A}} \mathbb{1}[i_\phi^{\textsc{Alg}} = i] \cdot \left(\bar{\ell}_i - \bar{\ell}_{i^*}\right)\right]}_{(1)},
\end{aligned}$$

where the second transition follows by noting that, as described above, $\Phi \leq \lfloor T/B \rfloor$, and B_j is well defined for all $j \in [K]$, regardless of timestep, so that we can extend the summation to $\phi \in (\Phi, \lfloor T/B \rfloor]$. The last transition uses $\max_j B_j = \frac{6 \ln(T)}{f^\star}$, a deterministic quantity. To relate (1) to the regret of $\textsc{Alg}$, we observe that in $\mathrm{BB}_{\mathrm{DA}}(\textsc{Alg})$, the algorithm $\textsc{Alg}$ also receives stochastic losses with mean $\bar{\ell}_i$ when it pulls $i_\phi^{\textsc{Alg}} = i$ that are identically distributed as in the original instance $\mathcal{I}$. This means that (1) is exactly equal to the regret of $\textsc{Alg}$ in an instance with stochastic losses over $\frac{Tf^\star}{3 \ln T}$ time steps. This completes the proof.

$\blacksquare$

C.3.3 Monotonicity of $\mathrm{BB}_{\mathrm{DA}}$: Proof of Theorem 3.6

We now prove Theorem 3.6. While the regret proof followed the regret proof for $\mathrm{BB}_{\mathrm{Divide}}$, the monotonicity proof will parallel the coupling argument we made for $\mathrm{BB}_{\mathrm{Pull}}$, with two key differences. First, the size of each block is now deterministic rather than a random variable; this makes analyzing each block easier, but requires a slightly different approach to formalizing the realization of randomness because the randomness is now in the *selection* of observations. Second, higher feedback probabilities will correspond to *larger* blocks by construction, which changes the direction of monotonicity in `APC` as desired.

The analogous result for FOC follows directly from Lemma 2.1. Intuitively, recall that in general, higher f_i implies higher $\texttt{FOC}_i$ for the same number of arm pulls, by definition; therefore, if $\texttt{APC}_i$ is positive monotonic, $\texttt{FOC}_i$ must be as well.

For reference, we restate Theorem 3.6 below.

Theorem 3.6. *[Impact of* $\mathrm{BB}_{\mathrm{DA}}$ *on* APC *and* FOC *] Fix an instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ *with stochastic losses. Let* $\tilde{f}_i \geq f_i$*, and let* $\widetilde{\mathcal{I}} = \{\mathcal{A}, \widetilde{\mathcal{F}}(i), \mathcal{L}\}$*. For any algorithm* ALG *for the deterministic feedback setting and for any* $f^\star \leq \min_i f_i$*, the algorithm* $\mathrm{BB}_{\mathrm{DA}}(\mathrm{ALG}, f^\star)$ *satisfies* $\texttt{APC}_i(\widetilde{\mathcal{I}}) \geq \texttt{APC}_i(\mathcal{I}) - 1/T$ *and* $\texttt{FOC}_i(\widetilde{\mathcal{I}}) \geq \frac{\tilde{f}_i}{f_i}\texttt{FOC}_i(\mathcal{I}) - \frac{\tilde{f}_i}{T} > \texttt{FOC}_i(\mathcal{I})$.

Proof of Theorem 3.6 (Monotonicity of $\mathrm{BB}_{\mathrm{DA}}$*).* Again, we fix an instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ with stochastic losses. Let $\tilde{f}_i \geq f_i$, and let $\tilde{\mathcal{I}} = \{\mathcal{A}, \mathcal{F}(i), \mathcal{L}\}$. The four-step argument proceeds as follows:

1. We first condition on the event that $U_{j,\phi} = \widetilde{U}_{j,\phi} = 1$ for all $j \in [K], \phi \in [T]$.
2. We call Algorithm 6 on $\mathcal{I}$ and $\tilde{\mathcal{I}}$, passing in $U_{j,\phi}$ and $\widetilde{U}_{j,\phi}$, respectively. We use Lemma C.7 to argue that for any Φ^*, $(i_1^{\mathrm{ALG}}, \ldots, i_{\Phi^*}^{\mathrm{ALG}}) \stackrel{d}{=} (\tilde{i}_1^{\mathrm{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\mathrm{ALG}})$, then couple the arm pulls on each instance so $(i_1^{\mathrm{ALG}}, \ldots, i_{\Phi^*}^{\mathrm{ALG}}) = (\tilde{i}_1^{\mathrm{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\mathrm{ALG}})$.
3. By this step, i_ϕ^{ALG} and $\tilde{i}_\phi^{\mathrm{ALG}}$ are fixed up to Φ^* by the above coupling. Now, we truncate the run Algorithm 6 to T rounds on each instance. In this case, ALG may be called a different number of times, $\Phi \geq \widetilde{\Phi}$, on instance $\mathcal{I}$ and $\widetilde{\mathcal{I}}$. This impacts the monotonicity of APC in the claimed direction.
4. Finally, we handle the conditioning from Step 1, using Lemma C.4 to argue that the event that an observation is not observed in at least one block ϕ contributes at most $1/T$ to $\texttt{APC}_i(\mathcal{I})$.

Step 1: Condition on feedback observations. First, let E be the event that $U_{j,\phi} = \widetilde{U}_{j,\phi} = 1$ for all $j \in [K], \phi \in [T]$. By Lemma C.4, $\Pr[E] \geq 1 - 1/T^2$. Then, for any $\phi > T$, we let $U_{j,\phi}$ and $\widetilde{U}_{j,\phi}$ take on arbitrary values in $\{0, 1\}$. We condition on E for Steps 2-4.

Step 2: Run Algorithm 6 and couple arms pulled by Alg across $\mathcal{I}$ and $\widetilde{\mathcal{I}}$. We next consider Algorithm 6 on $\mathcal{I}$ and $\tilde{\mathcal{I}}$ using the sequences $U_{j,\phi}$ and $\widetilde{U}_{j,\phi}$ for $j \in [K]$ and $\phi = 1, \ldots, \infty$, respectively. We can now apply Lemma C.7, letting $\Phi^* = T$, so that $(i_1^{\mathrm{ALG}}, \ldots, i_{\Phi^*}^{\mathrm{ALG}})$ and $(\tilde{i}_1^{\mathrm{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\mathrm{ALG}})$ are identically distributed. This allows us to consider a joint probability distribution over $(i_1^{\mathrm{ALG}}, \ldots, i_{\Phi^*}^{\mathrm{ALG}}, \tilde{i}_1^{\mathrm{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\mathrm{ALG}})$ such that $i_\phi^{\mathrm{ALG}} = \tilde{i}_\phi^{\mathrm{ALG}}$ for all $\phi \in [\Phi^*]$.

Step 3: Handle stopping times. We condition on the coupling thus far with the property that $U_{j,\phi}$, $\widetilde{U}_{j,\phi}$, i_ϕ^{ALG}, $\tilde{i}_\phi^{\mathrm{ALG}}$ are all fixed and satisfy $i_\phi^{\mathrm{ALG}} = \tilde{i}_\phi^{\mathrm{ALG}}$ and $U_{j,\phi} = \tilde{U}_{j,\phi} = 1$.

This step can be thought of as the inverse of Step 3 of the proof of Theorem 3.4. As in that step, Φ and $\tilde{\Phi}$ are deterministic. This time, however, we now have that $\Phi \geq \widetilde{\Phi}$; see Figure 2 for an illustration. Intuitively, on $\widetilde{\mathcal{I}}$, the block sizes when i is pulled are *larger* than on $\mathcal{I}$, so $\mathrm{BB}_{\mathrm{DA}}(\mathrm{ALG})$ moves through the ϕ-indexed timescale more slowly on $\widetilde{\mathcal{I}}$.

Let $B_\phi := B_{i_\phi^{\mathrm{ALG}}}$ denote the size of block ϕ on $\mathcal{I}$ and $\widetilde{B}_\phi := B_{\tilde{i}_\phi^{\mathrm{ALG}}}$ denote the size of block ϕ on $\tilde{\mathcal{I}}$. For each $\phi \leq \min(\Phi, \widetilde{\Phi})$, we know that $i_\phi^{\mathrm{ALG}} = \tilde{i}_\phi^{\mathrm{ALG}}$. Therefore, the only case where $B_\phi \neq \widetilde{B}_\phi$ is when $i_\phi^{\mathrm{ALG}} = i$; these are illustrated by gray blocks in Figure 2, in which case $B_\phi \leq \widetilde{B}_\phi$, by definition. Let $\phi(t)$ (resp. $\widetilde{\phi}(t)$) be the function that maps timesteps on the timescale indexed by t to timesteps on ALG's timescale on $\mathcal{I}$ (resp. $\widetilde{\mathcal{I}}$). Every time step $t \in [T]$ maps to blocks $\phi(t)$ and $\widetilde{\phi}(t)$ in instances $\mathcal{I}$ and $\widetilde{\mathcal{I}}$, respectively, such that $\widetilde{\phi}(t) \leq \phi(t)$. This implies that $\Phi \leq \widetilde{\Phi}$.

Notation for Analyzing APC. We are now ready to analyze APC on $\mathcal{I}$ and $\widetilde{\mathcal{I}}$. To formalize our analysis, we introduce the following additional notation (following the proof of Theorem 3.4). Given a range $R \subseteq [T]$, let $\texttt{APC}_i^R(\mathcal{I})$ be the number of times arm i is pulled in timesteps in R on $\mathcal{I}$, and let $\texttt{APC}_i^R(\widetilde{\mathcal{I}})$ be the number of times arm i is pulled in timesteps in R on $\widetilde{\mathcal{I}}$. Since we have conditioned on $U_{j,\phi}$, $\widetilde{U}_{j,\phi}$, i_ϕ^{ALG}, $\tilde{i}_\phi^{\mathrm{ALG}}$, we see that at this point $\texttt{APC}_i^R(\mathcal{I})$ and $\texttt{APC}_i^R(\widetilde{\mathcal{I}})$ are both deterministic.

Since we will analyze the last time block separately, we let $T_1 = \sum_{\phi \in [\widetilde{\Phi}-1]} B_\phi \leq T$ be the time step referring to the end of the penultimate block of $\mathrm{BB}_{\mathrm{DA}}(\textsc{Alg})$ on $\widetilde{\mathcal{I}}$. Let T_0 be the analogous time on $\mathcal{I}$, so that $T_0 = \sum_{\phi \in [\widetilde{\Phi}-1]} \widetilde{B}_\phi \leq T_1$.

Analyzing APC. We first prove that $\mathtt{APC}_i^{[T]}(\mathcal{I}) \leq \mathtt{APC}_i^{[T]}(\widetilde{\mathcal{I}})$. Because the only case where $B_\phi \neq \widetilde{B}_\phi$ is when $i_\phi^{\textsc{Alg}} = i$, we have that

$$\begin{aligned}
T_1 - T_0 &= \sum_{\phi \in [\widetilde{\Phi}-1]} \mathbb{1}(i_\phi^{\textsc{Alg}} = i) \cdot \left(\widetilde{B}_{i_\phi} - B_{i_\phi}\right). \\
&= \mathtt{APC}_i^{[T_1]}(\widetilde{\mathcal{I}}) - \mathtt{APC}_i^{[T_0]}(\mathcal{I}).
\end{aligned}$$

Now, consider two cases.

- Case 1: i was pulled in the $\widetilde{\Phi}$th block of $\widetilde{\mathcal{I}}$, i.e. $i_{\widetilde{\Phi}}^{\textsc{Alg}} = i$. Then,

$$\begin{aligned}
\mathtt{APC}_i^{[T]}(\mathcal{I}) &\leq \mathtt{APC}_i^{[T_0]}(\mathcal{I}) + T - T_0 \\
&= \mathtt{APC}_i^{[T_1]}(\widetilde{\mathcal{I}}) + T - T_1 \\
&= \mathtt{APC}_i^{[T]}(\widetilde{\mathcal{I}}).
\end{aligned}$$

- Case 2: Some other arm was pulled in the $\widetilde{\Phi}$th block of $\widetilde{\mathcal{I}}$, i.e. $i_{\widetilde{\Phi}}^{\textsc{Alg}} \neq i$. Then, we know that $\mathtt{APC}_i(\widetilde{\mathcal{I}}) = \mathtt{APC}_i^{[T_1]}(\widetilde{\mathcal{I}})$. Moreover, within the last $T - T_0$ timesteps of $\mathcal{I}$, at least $T - T_1$ of them are dedicated to pulling arm $j \neq i$ in the $\widetilde{\Phi}$th block of $\widetilde{\mathcal{I}}$. Then,

$$\begin{aligned}
\mathtt{APC}_i^{[T]}(\mathcal{I}) &\leq \mathtt{APC}_i^{[T_0]}(\mathcal{I}) + T - T_0 - (T - T_1) \\
&= \mathtt{APC}_i^{[T_0]}(\mathcal{I}) + T_1 - T_0 \\
&= \mathtt{APC}_i^{[T_1]}(\widetilde{\mathcal{I}}) = \mathtt{APC}_i^{[T]}(\widetilde{\mathcal{I}}).
\end{aligned}$$

Combining these two cases gives us that

$$\mathtt{APC}_i^{[T]}(\mathcal{I}) \leq \mathtt{APC}_i^{[T]}(\widetilde{\mathcal{I}}).$$

We can apply the law of total expectation over the sequences $U_{j,\phi}$, $\widetilde{U}_{j,\phi}$, $i_\phi^{\textsc{Alg}}$, $\tilde{i}_\phi^{\textsc{Alg}}$. Let $\mathtt{APC}_i(\mathcal{I} \mid E)$ notate the metric $\mathtt{APC}_i$ on instance $\mathcal{I}$ conditioned on the clean event E. We see that:

$$\mathtt{APC}_i(\widetilde{\mathcal{I}} \mid E) - \mathtt{APC}_i(\mathcal{I} \mid E) = \mathbb{E}\left[\mathtt{APC}_i^{[T]}(\widetilde{\mathcal{I}}) - \mathtt{APC}_i^{[T]}(\mathcal{I}) \mid E\right] \geq 0.$$

This means that:

$$\mathtt{APC}_i(\widetilde{\mathcal{I}} \mid E) \geq \mathtt{APC}_i(\mathcal{I} \mid E).$$

Step 4: Handle conditioning on feedback observations. Finally, recall that up to this point, we are still conditioning on E from Step 1, i.e. that we see feedback in every block on each instance. By Lemma C.4, $\Pr[\neg E] \leq 1/T^2$. In the worst case, we pull i for every $t \in [T]$ on $\mathcal{I}$, which gives $\mathtt{APC}_i(\mathcal{I} \mid \neg E) \leq T$. To relate this to $\mathtt{APC}_i(\mathcal{I})$ overall, we can see that

$$\begin{aligned}
\mathtt{APC}_i(\mathcal{I}) &= \mathtt{APC}_i(\mathcal{I} \mid E) \cdot \Pr[E] + \mathtt{APC}_i(\mathcal{I} \mid \neg E) \cdot \Pr[\neg E] \\
&\leq \mathtt{APC}_i(\mathcal{I} \mid E) + T \cdot 1/T^2 \\
\implies \mathtt{APC}_i(\mathcal{I}) - 1/T &\leq \mathtt{APC}_i(\mathcal{I} \mid E).
\end{aligned}$$

Combining this with the result from Step 3, we have that

$$\mathtt{APC}_i(\widetilde{\mathcal{I}}) \geq \mathtt{APC}_i(\widetilde{\mathcal{I}} \mid E) \geq \mathtt{APC}_i(\mathcal{I} \mid E) \geq \mathtt{APC}_i(\mathcal{I}) - 1/T.$$

Analyzing FOC. Applying Lemma 2.1 gives us $\mathtt{FOC}_i(\tilde{\mathcal{I}}) \cdot f_i \geq \mathtt{FOC}_i(\mathcal{I}) \cdot \tilde{f}_i - \tilde{f}_i/T$, and the result follows from dividing both sides by f_i. ■

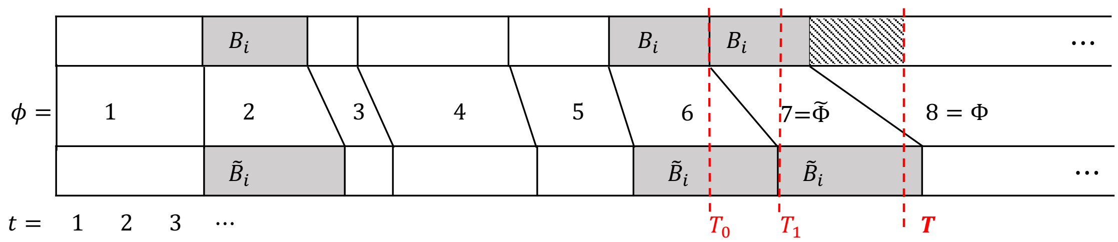

Figure 2: Timelines of $\mathrm{BB}_{\mathrm{DA}}(\textsc{Alg})$ on instances $\mathcal{I}$ (top row) and $\widetilde{\mathcal{I}}$ (bottom row) are demonstrated. Each time step $t \in [T]$ maps to a block number in $\mathcal{I}$ that is no *less* than its block number in $\widetilde{\mathcal{I}}$. The total number of times $\textsc{Alg}$ is called in instance $\mathcal{I}$, Φ, and the number of times it is called in $\widetilde{\mathcal{I}}$, $\widetilde{\Phi}$, satisfy $\Phi \geq \widetilde{\Phi}$. Note that this is similar to Figure 1, except that the direction of monotonicity has switched and that the size of B_i and $\widetilde{B}_i$ is deterministic in each instance.

D Supplemental Materials for Sections 4.1 and 4.2

In this section, we analyze $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{AAE})$ (Appendix D.1), $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{UCB})$ (Appendix D.2), and $\mathrm{BB}_{\mathrm{DA}}(\mathrm{AAE})$ (Appendix D.3).

We first formalize each below.

Algorithm 7: $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{AAE})$

Maintain active set A; start with $A := [K]$.
Initialize phase $s = 1$ and $t = 1$.
while $t \leq T$ **do**
  for *arm* $i \in A$ **do**
    Let $R_{i,s} = \emptyset$.
    while $|R_{i,s}| \leq 8 \ln T \cdot 2^{2s}$ *and* $t \leq T$ **do**
      if $X_{i,t} = 1$ **then**
        Append $R_{i,s} \leftarrow R_{i,s} \cup \{t\}$.
      Pull $i_t = i$, and increment $t \leftarrow t + 1$.
    Calculate the mean $\mu_s(i) := -\frac{1}{|R_{i,s}|} \sum_{t' \in R_{i,s}} \ell_{i,t'}$ of the negative of all observations.[16]
    Set $\mathrm{LCB}_s(i) = \mu_s(i) - 2^{-s}$ and $\mathrm{UCB}_s(i) = \mu_s(i) + 2^{-s}$.
  For any arm $i \in A$ where $\exists j \in A$ such that $\mathrm{LCB}_s(j) > \mathrm{UCB}_s(i)$, remove i from A.
  Increment $s \leftarrow s + 1$.

D.1 Analysis of $\mathbf{BB_{Pull}}$ applied to AAE

We prove monotonicity properties and regret bounds for $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{AAE})$ (Algorithm 7), where AAE denotes the standard Active Arm Elimination algorithm.

D.1.1 A simulated version of $\mathbf{BB_{Pull}}$(AAE)

We consider the simulated version of $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{AAE})$ given by Algorithm 5 applied to AAE. For convenience, we explicitly state this algorithm below (Algorithm 10).

Let us define the same random variables as those used in Algorithm 5, restated for convenience. (Recall that ϕ indexes losses for the time horizon of $\textsc{Alg}$, Φ is the total number of times $\textsc{Alg}$ is called by $\mathrm{BB}_{\mathrm{Pull}}(\textsc{Alg})$, and $\Phi \leq T$ because $\textsc{Alg}$ can be called at most T times.)

- *Losses:* For each round $\phi \in [\Phi]$ of $\textsc{Alg} = \mathrm{AAE}$ and each arm $j \in [K]$, let $\ell'_{j,\phi} := \ell_{j,t}$ be the loss for arm j at a time step t that corresponds to the last time step in block ϕ of $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{AAE})$. Since we are in

[16]The negative is introduced to convert losses into utilities.

Algorithm 8: $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{UCB})$

Initialize number of pulls $n_i = 0$ for all $i \in [K]$.
Initialize empirical mean $\mu(i) = 0$ for all $i \in [K]$.
Initialize $t = 1$.
while $t \leq T$ **do**
    if $n_i = 0$ *for any arm* $i \in [K]$ **then**
        Let i_t be the arm with the smallest index such that $n_{i_t} = 0$.
    else
        For every arm $i \in [K]$, compute $\mathrm{UCB}(i) = \mu(i) + \sqrt{\frac{6 \ln T}{n_i}}$.
        Let $i_t = \mathrm{argmax}_{j \in [K]} \mathrm{UCB}(j)$.
    Pull arm i_t.
    if $X_{i,t} = 1$ **then**
        Update the empirical mean $\mu(i) \leftarrow \frac{n_{i_t} \cdot \mu(i)}{n_{i_t}+1} - \frac{\ell_{i_t,t}}{n_{i_t}+1}$.
        Increment $n_{i_t} \leftarrow n_{i_t} + 1$.
    Increment $t \leftarrow t + 1$.

Algorithm 9: $\mathrm{BB}_{\mathrm{DA}}(\mathrm{AAE}, f^\star)$

Maintain active set of A; start with $A := [K]$.
For arm $i \in [K]$, set $B_i = \lceil (1 + f_i) \cdot \frac{3 \ln T}{f^\star} \rceil$.
Initialize phase $s = 1$ and $t = 1$.
while $t \leq T$ **do**
    for *arm* $i \in A$ **do**
        Let $R_{i,s} = \emptyset$.
        for $\min(B_i, T - t)$ *iterations* **do**
            if $X_{i,t} = 1$ *and* $|R_{i,s}| < 8 \ln T \cdot 2^{2s}$ **then**
                Append $R_{i,s} \leftarrow R_{i,s} \cup \{t\}$.
            Pull $i_t = i$, and increment $t \leftarrow t + 1$.
        Calculate the mean $\mu_s(i) := -\frac{1}{\min(|R_{i,s}|, 2 \ln T \cdot 2^{2s})} \sum_{t' \in R_{i,s}} \ell_{i,t'}$ of the negative of the first $8 \ln T \cdot 2^{2s}$ observations (if more than $8 \ln T \cdot 2^{2s}$ observations are made).
        Set $\mathrm{LCB}_s(i) = \mu_s(i) - 2^{-s}$ and $\mathrm{UCB}_s(i) = \mu_s(i) + 2^{-s}$.
    For any arm $i \in A$ where $\exists j \in A$ such that $\mathrm{LCB}_s(j) > \mathrm{UCB}_s(i)$, remove i from A.
    Increment $s \leftarrow s + 1$.

the stochastic loss setting, $\ell'_{j,\phi}$ is a random variable drawn from the distribution of arm j (with mean $\bar{\ell}_j$) independently across ϕ and j.

- *Feedback realizations:* For all $j \in [K]$ and $\phi \in [T]$, let $Q_{j,\phi} \sim \text{Geom}(f_j)$ for $\phi \in [T]$ be a random variable distributed according to the geometric distribution with parameter equal to the feedback probability of arm j. (These random variables are also fully independent across values of j and ϕ.)

We are now ready to present Algorithm 10. For ease of analysis, we make the slight modification from Algorithm 7 that we convert the set $R_{i,s}$ which keeps track of time steps in the time horizon of $\text{BB}_{\text{Pull}}(\text{AAE})$ to the set $U_{i,s}$ which keeps track of time steps in the time horizon of $\textsc{Alg} = \text{AAE}$. The behavior of the algorithm remains unchanged under this change.

Algorithm 10: Simulated version of $\text{BB}_{\text{Pull}}(\text{AAE})$ (Algorithm 5 applied to AAE)

1 Maintain active set A; start with $A := [K]$.
2 Initialize phase $s = 1$, $t = 1$, and $\phi = 1$.
3 **while** $t \leq T$ **do**
4   **for** *arm* $i \in A$ **do**
5   Let $U_{i,s} = \emptyset$.
6   **while** $|U_{i,s}| \leq 8 \ln T \cdot 2^{2s}$ *and* $t \leq T$ **do**
7   Start phase s.
8   **for** $\min(Q_{i,\phi}, T - t)$ *iterations* **do**
9   Pull $i_t = i$ and let $t \leftarrow t + 1$.
10   Observe $\ell'_{i,\phi} := \ell_{i,t}$, append $U_{i,s} \cup \{\phi\}$, and let $\phi \leftarrow \phi + 1$.
11   Calculate the mean $\mu_s(i) := -\frac{1}{|U_{i,s}|} \sum_{\phi' \in U_{i,s}} \ell'_{i,\phi'}$ of the negative of all observations.
12   Set $\text{LCB}_s(i) = \mu_s(i) - 2^{-s}$ and $\text{UCB}_s(i) = \mu_s(i) + 2^{-s}$.
13   For any arm $i \in A$ where $\exists j \in A$ such that $\text{LCB}_s(j) > \text{UCB}_s(i)$, remove i from A.
14   Increment $s \leftarrow s + 1$.

Since Algorithm 10 is exactly Algorithm 5 applied to AAE, we can apply Lemma C.6 to see that the sequence of arms $\{i_t^{\text{orig}}\}_{t \in [T]}$ pulled by Algorithm 7 is distributed identically to the sequence of arms pulled by $\{i_t^{\text{sim}}\}_{t \in [T]}$ pulled by Algorithm 10. It thus suffices to analyze Algorithm 10 for the remainder of the analysis.

D.1.2 Lemmas for the analysis of $\text{BB}_{\text{Pull}}(\text{AAE})$

We now show intermediate results that build on the standard analysis of Active Arm Elimination [Even-Dar et al., 2002].

We use the following notation in these results.

1. Let S be a random variable denoting the maximum value of the variable s reached in Algorithm 10 on $\mathcal{I}$. (That is, S denotes the number of phases that Algorithm 10 *begins*.) Note that $S \leq T$ with probability 1.

2. Let E_{loss} be the "clean" event that at each phase $1 \leq s \leq S - 1$, for every arm $i \in [K]$, it holds that $\text{LCB}_s(i) \leq \bar{\ell}_i \leq \text{UCB}_s(i)$.

3. Let the random variable $L_{i,s}$ be equal to the time step t where phase s begins for arm i (i.e. the value of the variable t at line 5 when $U_{i,s}$ is initialized) if that is reached, and otherwise let $L_{i,s}$ be equal to $T + 1$.

4. For each arm i, let E_i^F be the event that at each phase $1 \leq s \leq T$, at least one of the following two conditions holds: (1) $L_{i,s} = T + 1$, or (2):

$$\sum_{\phi' \in U_s(i)} Q_{i,\phi'} \leq \frac{16 \cdot 2^{2s} \ln T}{f_i}.$$

First, we show that the clean events occur with high probability.

Lemma D.1 (Correct confidence bounds)**.** *Consider Algorithm 10 evaluated on any given instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$*. Let the event* E_{loss} *be defined as above. Then,* $\Pr[E_{loss}] \geq 1 - 2T^{-3}K$*.*

Proof. For each potential phase $1 \leq s \leq T$ and arm i, let $E^{i,s}_{\text{loss}}$ be the event that either $s \geq S$ or $\text{LCB}_s(i) \leq \bar{\ell}_i \leq \text{UCB}_s(i)$. Condition on the event that $L_{i,s} \leq T$. For ease of analysis, let us also assume that we draw additional loss values, $\ell'_{i,\phi}$ for $T \leq \phi \leq T + 8 \ln T \cdot 2^{2s}$ i.i.d. from the loss distribution of arm i.

Run the algorithm for $T + 8 \ln T \cdot 2^2 s$ time steps rather than T time steps, which ensures that line 11 for i and s is reached and the confidence bounds $\text{LCB}_s(i)$ and $\text{UCB}_s(i)$ are well-defined. We show that $\mathbb{P}[E^{i,s}_{\text{loss}} \mid L_{i,s} \leq T] \geq 1 - 2T^{-4}$:

$$\begin{aligned}
\mathbb{P}[E^{i,s}_{\text{loss}}] &\geq \mathbb{P}[E^{i,s}_{\text{loss}} \mid L_{i,s} \leq T] \cdot \mathbb{P}[L_{i,s} \leq T] + \mathbb{P}[L_{i,s} > T] \\
&\geq \mathbb{P}[\text{LCB}_s(i) \leq \bar{\ell}_i \leq \text{UCB}_s(i)] \cdot \mathbb{P}[L_{i,s} \leq T] + \mathbb{P}[L_{i,s} > T] \\
&\geq \mathbb{P}[\text{LCB}_s(i) \leq \bar{\ell}_i \leq \text{UCB}_s(i)] \\
&= \mathbb{P}\left[\left| \bar{\ell}_i - \frac{1}{|U_{i,s}|} \sum_{\phi' \in U_{i,s}} \ell'_{i,\phi} \right| \leq 2^{-s}\right].
\end{aligned}$$

Recall that we are working with stochastic losses, so $\bar{\ell}_i - \frac{1}{|U_{i,s}|} \sum_{\phi' \in U_{i,s}} \ell'_{i,\phi'}$ is distributed as an average of $|U_{i,s}| = 8 \ln T \cdot 2^{2s}$ subgaussian random variables with variance 1. Using a Chernoff bound, we have that:

$$\mathbb{P}\left[\left| \bar{\ell}_i - \frac{1}{|U_{i,s}|} \sum_{\phi' \in U_{i,s}} \ell'_{i,\phi} \right| > 2^{-s}\right] \leq 2e^{-\frac{8 \ln T \cdot 2^{2s}}{2^{2s+1}}} = 2T^{-4}.$$

Finally, we apply a union bound to bound $\Pr[E_{\text{loss}}]$. There are $S \leq T$ potential phases and K arms, so there are KT events to union bound over. We see that:

$$\Pr[E_{\text{loss}}] \geq \sum_{s=1}^{T} \sum_{i=1}^{K} \mathbb{P}[E^{i,s}_{\text{loss}}] \geq 1 - 2T^{-4} TK = 1 - 2T^{-3}K.$$

■

Lemma D.2. *Consider Algorithm 10 evaluated on any given instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ *with time horizon* T*. Suppose that the event* E_{loss} *holds. Then, the optimal arm* $i^\star = \arg\min_j \bar{\ell}_j$ *is never removed from* A*. Moreover, at every phase* $1 \leq s \leq S - 1$*, if* $i \in A$ *at the end of phase* s *(i.e. after 13 in Algorithm 10), then*

$$\bar{\ell}_i - \min_j \bar{\ell}_j \leq 4 \cdot 2^{-s}.$$

Proof. Let us condition on E_{loss}, which means that $\text{LCB}_s(i) \leq \bar{\ell}_i \leq \text{UCB}_s(i)$ for every arm i and every phase $1 \leq s \leq S - 1$. For the optimal arm $i^\star = \arg\min_j \bar{\ell}_j$ it holds for every phase s that:

$$\text{UCB}_s(i^\star) \geq -\bar{\ell}_{i^\star} = -\min_j \bar{\ell}_j = \max_j(-\bar{\ell}_j) \geq \max_j \text{LCB}_s(j),$$

so the optimal arm will never be removed from A, as desired.

If arm i is in the active arm set A at the end of phase s (i.e. after line 13), then

$$\text{UCB}_s(i) \geq \text{LCB}_s(i^\star) \geq -\bar{\ell}_{i^\star} - 2 \cdot 2^{-s}.$$

This means that

$$-\bar{\ell}_i \geq \text{LCB}_s(i) \geq \text{UCB}_s(i) - 2 \cdot 2^{-s} \geq -\bar{\ell}_{i^\star} - 4 \cdot 2^{-s} = -\min_j \bar{\ell}_j - 4 \cdot 2^{-s}.$$

Rearranging, we obtain that:

$$\bar{\ell}_i - \min_j \bar{\ell}_j \leq 4 \cdot 2^{-s}.$$

as desired. . ■

Lemma D.3. *Consider Algorithm 10 evaluated on any given instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$. For each arm i, let E_i^F be defined as above. Then,* $\Pr[E_i^F] \geq 1 - T^{-4}$.

Proof. For each arm i and each phase $1 \leq s \leq T$, let $E_{i,s}^F$ be the event that

$$\sum_{\phi' \in U_s(i)} Q_{i,\phi'} \leq \frac{8 \cdot 2^{2s} \ln T}{f_i}.$$

We lower bound the probability $\mathbb{P}\left[E_{i,s}^F\right]$. We analyze $\mathbb{P}\left[\sum_{\phi' \in U_s(i)} Q_{i,\phi'} \leq \frac{16 \cdot 2^{2s} \ln T}{f_i}\right]$ as follows. Let $m = \frac{16 \cdot 2^{2s} \ln T}{f_i}$. By definition, the probability that $\sum_{\phi' \in U_s(i)} Q_{i,\phi'} > m$ is equal to the probability that fewer than $8 \cdot 2^{2s} \cdot \ln T$ successes are observed after m i.i.d. Bernouilli trials with parameter f_i. This probability can be analyzed with a Chernoff bound. In particular, let $Z_j \sim \text{Bern}(f_i)$ for $1 \leq j \leq m$ be a sequence of m i.i.d. random variables. Using the multiplicative Chernoff bound, we see that:

$$\begin{aligned}
\mathbb{P}[\sum_{\phi' \in U_s(i)} Q_{i,\phi'} > \frac{16 \cdot 2^{2s} \ln T}{f_i}] &= \Pr\left[\sum_{j=1}^{m} Z_j < 8 \cdot 2^{2s} \cdot \ln T\right] \\
&\leq \Pr\left[\sum_{j=1}^{m} Z_j < m \cdot f_i \cdot 0.5\right] \\
&\leq \exp\left(-m \cdot f_i \cdot \frac{1}{8}\right) \\
&= \exp\left(-\frac{1}{f_i} \cdot 16 \cdot 2^{2s} \cdot \ln T \cdot f_i \cdot \frac{1}{8}\right) \\
&= T^{-2^{2s+1}} \\
&\leq T^{-5}.
\end{aligned}$$

This implies that $\mathbb{P}[E_{i,s}^F] \geq 1 - T^{-5}$. Union bounding over the T values of s, we obtain that $\Pr[E_i^F] \geq 1 - T^{-4}$.

■

D.1.3 Regret of $\text{BB}_{\text{Pull}}(\text{AAE})$: Proof of Theorem 4.1

Here, we prove the regret bound for $\text{BB}_{\text{Pull}}(\text{AAE})$. For convenience, we restate Theorem 4.1 below.

Theorem 4.1. *On any stochastic instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{I}\}$, $\text{BB}_{\text{Pull}}(\text{AAE})$ (presented in Algorithm 7) and $\text{BB}_{\text{Pull}}(\text{UCB})$ (presented in Algorithm 8) have regret bound of $O\left(\sqrt{T \ln(T) \sum_{i \in [K]} 1/f_i}\right)$ and an instance-dependent regret bound of $O\left(\sum_{i \in [K] | \Delta_i > 0} \frac{\ln T}{\Delta_i f_i}\right)$.*

We will prove the statement of Theorem 4.1 only for $\text{BB}_{\text{Pull}}(\text{AAE})$. In the proof of the regret bounds, we will use the following lemma.

Lemma D.4. *Let $\Delta_1, \Delta_2, \ldots, \Delta_K \geq 0$ be a sequence of nonnegative numbers. Let $N_1, \ldots, N_K \geq 0$ be a sequence of nonnegative numbers such that for some $C > 0$, it holds that $N_i \leq \frac{C \cdot \ln T}{\Delta_i^2 f_i}$ for all $1 \leq i \leq K$. Then the following two bounds hold:*

$$\sum_{1 \leq i \leq K | \Delta_i > 0} \Delta_i \cdot N_i \leq \sum_{1 \leq i \leq K | \Delta_i > 0} \frac{C \ln T}{\Delta_i f_i}$$

and

$$\sum_{1 \le i \le K | \Delta_i > 0} \Delta_i \cdot N_i \le \sqrt{CT \ln(T) \sum_{j=1}^{K} \frac{1}{f_j}}.$$

Proof. The first bound follows from:

$$\sum_{1 \le i \le K | \Delta_i > 0} \Delta_i \cdot N_i \le \sum_{1 \le i \le K | \Delta_i > 0} \Delta_i \cdot \frac{C \ln T}{\Delta_i^2 f_i} \Delta_i = \sum_{1 \le i \le K | \Delta_i > 0} \frac{C \ln T}{\Delta_i f_i}$$

For the second bound, first we rearrange the upper bound on N_i into:

$$\Delta_i \le \sqrt{\frac{C \ln T}{N_i f_i}}.$$

Now, we see that

$$\begin{aligned}
\sum_{1 \le i \le K | \Delta_i > 0} N_i \Delta_i &\le \sum_{1 \le i \le K | \Delta_i > 0} \sqrt{\frac{C N_i \ln(T)}{f_i}} \\
&= \sum_{1 \le i \le K | \Delta_i > 0} \frac{1}{f_i} \sqrt{N_i \ln(T) f_i} \\
&\le \sum_{1 \le i \le K} \frac{1}{f_i} \sqrt{N_i \ln(T) f_i} \\
&= \left(\sum_{j=1}^{K} \frac{1}{f_j} \right) \sum_{i=1}^{K} \frac{\frac{1}{f_i}}{\left(\sum_{j=1}^{K} \frac{1}{f_j} \right)} \sqrt{C N_i \ln(T) f_i} \\
&\le_{(1)} \left(\sum_{j=1}^{K} \frac{1}{f_j} \right) \sqrt{C \sum_{i=1}^{K} \frac{\frac{1}{f_i}}{\left(\sum_{j=1}^{K} \frac{1}{f_j} \right)} N_i \ln(T) f_i} \\
&= \sqrt{\sum_{j=1}^{K} \frac{1}{f_j}} \sqrt{C \sum_{i=1}^{K} N_i \ln(T)} \\
&\le_{(2)} \sqrt{\sum_{j=1}^{K} \frac{1}{f_j}} \sqrt{CT \ln(T)}
\end{aligned}$$

where (1) follows from Jensen's inequality and (2) follows from the fact that $\sum_{i=1}^{K} N_i = T$. ■

We are now ready to prove Theorem 4.1 for $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{AAE})$.

Proof of Theorem 4.1 for $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{AAE})$. By Lemma C.6, the sequence of arms $\{i_t^{\mathrm{orig}}\}_{t \in [T]}$ pulled by Algorithm 7 is distributed identically to the sequence of arms pulled by $\{i_t^{\mathrm{sim}}\}_{t \in [T]}$ pulled by Algorithm 10. Define the event E to be $E := E_{\mathrm{loss}} \cap E_F^1 \dots E_F^K$ where the events are defined as in Lemma D.1 and Lemma D.3. Union bounding, E occurs with probability at least $1 - 2T^{-3}K - KT^{-4}$. When T is sufficiently large, $\mathbb{P}[E] \ge 1 - T^{-2}$, so the event that E does not occur contributes negligibly to the regret. Let us condition on E for the remainder of the analysis.

For each arm i, let $\Delta_i = \bar{\ell}_i - \min_j \bar{\ell}_j$ be the suboptimality gap. Let N_i be the number of time steps where arm i is pulled over the course of Algorithm 10. The regret is equal to:

$$\sum_{1 \le i \le K | \Delta_i > 0} \Delta_i \cdot N_i.$$

We first show that if $\Delta_i > 0$, then arm i is pulled at most $O\left(\frac{\log T}{\Delta_i^2 f_i}\right)$ times. By Lemma D.2, arm i must be eliminated after phase $\lceil \log_2 (4/\Delta_i) \rceil$. For phases $1 \leq s \leq \lceil \log_2 (4/\Delta_i) \rceil$, recall that we have defined the random variable $L_{i,s}$ to be equal to the time step t where phase s begins (i.e. the value of the variable t at line 5 when $U_{i,s}$ is initialized) if that is reached, and otherwise let $L_{i,s}$ be equal to $T+1$. This means that arm i is pulled at most:

$$
\begin{aligned}
N_i &\leq \sum_{s=1}^{\lceil \log(4/\Delta_i) \rceil} \min\left(\sum_{\phi' \in U_s(i)} Q_{i,\phi}, T - (L_{i,s} - 1) \right) \\
&\leq \sum_{s=1}^{\lceil \log(4/\Delta_i) \rceil} \sum_{\phi' \in U_s(i)} Q_{i,\phi} \\
&\leq_{(1)} \sum_{s=1}^{\lceil \log(4/\Delta_i) \rceil} \frac{16 \cdot 2^{2s} \ln T}{f_i} \\
&\leq 16 \cdot \frac{\ln T}{f_i} \sum_{s=1}^{\lceil \log(4/\Delta_i) \rceil} 2^{2s} \\
&\leq \frac{C \cdot \ln T}{\Delta_i^2 f_i}
\end{aligned}
$$

for some universal constant $C > 0$, where (1) follows from the event E_i^F holding.

The instance-dependent and instance-independent regret bounds now both follow from Lemma D.4.

■

D.1.4 Monotonicity of BB_{Pull}(AAE): Proof of Theorem 4.2

We prove Theorem 4.2, restated below.

Theorem 4.2. *Fix a stochastic instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$. Let i be such that $\bar{\ell}_i > \min_{j \in [K]} \bar{\ell}_j$. Let $\widetilde{f}_i > f_i$, and let $\widetilde{\mathcal{I}} = \{\mathcal{A}, \widetilde{\mathcal{F}}(i), \mathcal{L}\}$. For sufficiently large T,* BB_{Pull}(AAE) *satisfies* $|\texttt{FOC}_i(\mathcal{I}) - \texttt{FOC}_i(\widetilde{\mathcal{I}})| \leq 1/T$ *and* $\texttt{APC}_i(\widetilde{\mathcal{I}}) < \texttt{APC}_i(\mathcal{I})$.

The intuition is that we can leverage the structure of AAE to refine the analysis in Theorem 3.4. In particular, in the proof of Theorem 3.4, the difference in feedback observations came from the fact that the time horizons Φ and $\tilde{\Phi}$ were different (that is, the number of calls to ALG differed for the two instances). In contrast, in the proof of Theorem 4.2, we take advantage of a key structural property of AAE: we can upper bound the number of phases until arm i is guaranteed to be eliminated. By assuming that T is sufficiently large, we can guarantee that the algorithm will reach this phase on both instances and thus the arm will be eliminated. We formalize this using a coupling argument similar to the proof of Theorem 3.4, but that leverages the structure of AAE.

Proof of Theorem 4.2 (Monotonicity for BB_{Pull}(AAE)*).* Like in Theorem 3.4, we will analyze BB_{Pull}(ALG) by comparing the behavior of Algorithm 10 on $\mathcal{I}$ and $\widetilde{\mathcal{I}}$ in three steps; the main modification is in Step 3 below, where we condition on clean events specific to AAE.

1. We construct a probability coupling between the sequence of random variables $Q_{j,\phi}$ and $\widetilde{Q}_{j,\phi}$ for $j \in [K]$ and $\phi = 1, \ldots, \infty$. This coupling ensures that $Q_{i,\phi} \geq \widetilde{Q}_{i,\phi}$ for arm i and $Q_{j,\phi} = \widetilde{Q}_{j,\phi}$ for all other arms $j \neq i$, for all ϕ.
2. We call Algorithm 5 on $\mathcal{I}$ and $\widetilde{\mathcal{I}}$ with $Q_{j,\phi}$ and $\widetilde{Q}_{j,\phi}$ for $j \in [K]$ and $\phi = 1, \ldots, \infty$, respectively. We use Lemma C.5 to argue that for any Φ^*, $(i_1^{\text{ALG}}, \ldots, i_{\Phi^*}^{\text{ALG}})$ is identically distributed to $(\tilde{i}_1^{\text{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\text{ALG}})$; then, we couple the arm pulls on each instance so $(i_1^{\text{ALG}}, \ldots, i_{\Phi^*}^{\text{ALG}}) = (\tilde{i}_1^{\text{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\text{ALG}})$.

3. By this step, random variables $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, i_ϕ^{ALG} and $\tilde{i}_\phi^{\text{ALG}}$ are coupled as described above. Let E be the event that $E_{\text{loss}} \cap E_1^F \cap \cdots \cap E_{i-1}^F \ldots E_{i+1}^F \cap \cdots \cap E_K^F$ holds (these events are defined in Appendix D.1.2). We condition on the event E and analyze `FOC`. We then analyze `APC`.

Step 1: Coupling realizations of feedback observations. We couple the distributions over the feedback observations in the same way as in the proof of Theorem 3.4. Note that for $\tilde{f}_i > f_i$, the distribution of $\widetilde{Q}_{i,\phi}$ is stochastically dominated by $Q_{i,\phi}$. Therefore, there is a joint probability distribution over $(Q_{j,\phi}, \widetilde{Q}_{j,\phi})$ such that for all ϕ, with probability 1 the following holds: $Q_{i,\phi} \geq \widetilde{Q}_{i,\phi}$ and for all $j \neq i$, $Q_{j,\phi} = \widetilde{Q}_{j,\phi}$. This also gives us a coupling, that is a joint distribution, over $\left(\{Q_{j,\phi}\}_{j\in[K],\phi\in\{1,\ldots,\infty\}}, \{\widetilde{Q}_{j,\phi}\}_{j\in[K],\phi\in\{1,\ldots,\infty\}}\right)$ that meets the aforementioned property.

Step 2: Coupling arms pulled by Alg across instances $\mathcal{I}$ and $\widetilde{\mathcal{I}}$. We couple the arms in the same way as in the proof of Theorem 3.4. We condition on the sequences $Q_{j,\phi}$ and $\widetilde{Q}_{j,\phi}$ for $j \in [K]$ and $\phi = 1, \ldots, \infty$, respectively, as coupled in in Step 1, and we apply Lemma C.5. As before, the preconditions of this lemma are met for any Φ^*, so we have that $(i_1^{\text{ALG}}, \ldots, i_{\Phi^*}^{\text{ALG}})$ and $(\tilde{i}_1^{\text{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\text{ALG}})$ are identically distributed. This allows us to consider a joint probability distribution over $(i_1^{\text{ALG}}, \ldots, i_{\Phi^*}^{\text{ALG}}, \tilde{i}_1^{\text{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\text{ALG}})$ such that $i_\phi^{\text{ALG}} = \tilde{i}_\phi^{\text{ALG}}$ for all $\phi \in [\Phi^*]$.

Step 3: Condition on E. We use the coupling thus far with the property that $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, i_ϕ^{ALG}, $\tilde{i}_\phi^{\text{ALG}}$ are all fixed and for all $\phi = 1, \cdots, \infty$ satisfy $i_\phi^{\text{ALG}} = \tilde{i}_\phi^{\text{ALG}}$, $Q_{i,\phi} \geq \widetilde{Q}_{i,\phi}$, and $Q_{j,\phi} = \widetilde{Q}_{j,\phi}$ for $j \neq i$. Moreover, we condition on the event $E = E_{\text{loss}} \cap E_1^F \cap \cdots \cap \cdots \cap E_K^F$ holds on $\mathcal{I}$ (these events are defined in Appendix D.1.2).

Notation for Analyzing `FOC`. To formalize these claims, we introduce the following additional notation. Let $\texttt{FOC}_i^{[T]}(\mathcal{I})$ be the number of times feedback is observed on arm i in timesteps in R on $\widetilde{\mathcal{I}}$, and let $\texttt{FOC}_i^{[T]}(\widetilde{\mathcal{I}})$ be the number of times feedback is observed on arm i in timesteps on $\tilde{\mathcal{I}}$. Since we have conditioned on $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, i_ϕ^{ALG}, $\tilde{i}_\phi^{\text{ALG}}$, we see that at this point $\texttt{FOC}_i^{[T]}(\mathcal{I})$ and $\texttt{FOC}_i^{[T]}(\widetilde{\mathcal{I}})$ are both deterministic.

Analyzing `FOC`. We first show that arm i will be eliminated on both instances before the end of the time horizon. By Lemma D.2, from phase $s \geq s' := 3 - \log(\Delta_i)$ onwards, the arm i is guaranteed to not be pulled. Using events E_j^F, we see that phase $s' = 3 - \log(\Delta_i)$ must be reached on $\mathcal{I}$ within the following number of time steps:

$$
\begin{aligned}
\sum_{s=1}^{s'-1} \sum_{i' \in [K]} \sum_{\phi' \in U_s(i')} Q_{i',\phi} &\leq \sum_{s=1}^{s'-1} \sum_{i' \in [K]} \frac{16 \cdot 2^{2s} \ln T}{f_{i'}} \\
&= (16 \ln T) \left(\sum_{i' \in [K]} \frac{1}{f_{i'}} \right) \sum_{s=1}^{s'-1} 4^s \\
&= \frac{16 \cdot (4^{s'} - 1) \ln T}{3} \left(\sum_{i' \in [K]} \frac{1}{f_{i'}} \right) \\
&= \frac{16 \cdot 4^{3-\log(\Delta_i)} \ln T}{3} \left(\sum_{i' \in [K]} \frac{1}{f_{i'}} \right),
\end{aligned}
$$

which grows logarithmically in T. Thus, for sufficiently large T, $\frac{16 \cdot 4^{3-\log(\Delta_i)} \ln T}{3} \left(\sum_{i' \in [K]} \frac{1}{f_{i'}} \right) \leq T$, which means that phase s' will be reached on $\mathcal{I}$ within a time horizon of T. For $\tilde{\mathcal{I}}$, we use the fact that $\tilde{Q}_{j,\phi} \leq Q_{j,\phi}$ in our coupling and moreover $i_\phi^{\text{ALG}} = \tilde{i}_\phi^{\text{ALG}}$ for all $\phi \in [\Phi^*]$, so phase s' will be reached on $\tilde{\mathcal{I}}$ as well within a time horizon of T.

We are now ready to analyze FOC. We observe that:

$$\begin{aligned}
\texttt{FOC}_i^{[T]}(\mathcal{I}) &= \sum_{\phi=1}^{\Phi} \mathbb{1}[i_\phi = i] \cdot \mathbb{1}\left[\sum_{\phi'=1}^{\phi} Q_{i_{\phi'},\phi'} \leq T\right] \\
&= \sum_{s'=1}^{3-\log(\Delta_i)} \sum_{\phi \text{ in phase } s} \mathbb{1}[i_\phi = i] \\
&= \sum_{\phi=1}^{\tilde{\Phi}} \mathbb{1}[i_\phi = i] \cdot \mathbb{1}\left[\sum_{\phi'=1}^{\phi} \tilde{Q}_{i_{\phi'},\phi'} \leq T\right] \\
&= \texttt{FOC}_i^{[T]}(\tilde{\mathcal{I}}).
\end{aligned}$$

Applying the law of total expectation, taking an expectation over the coupled random variables $Q_{j,\phi}$, $\widetilde{Q}_{j,\phi}$, i_ϕ^{ALG}, $\tilde{i}_\phi^{\text{ALG}}$, we see that, conditioned on E,

$$\texttt{FOC}_i(\tilde{\mathcal{I}}) - \texttt{FOC}_i(\mathcal{I}) = \mathbb{E}\left[\texttt{FOC}_i^{[T]}(\widetilde{\mathcal{I}}) - \texttt{FOC}_i^{[T]}(\mathcal{I})\right] \geq 0.$$

The event that E does not hold contributes negligibly (i.e., at most $1/T$) to both $\texttt{FOC}(\mathcal{I})$ and $\texttt{FOC}(\tilde{\mathcal{I}})$. Taking expectations and including the possibility of $1/T$ error from the event E not holding, we obtain that:

$$|\texttt{FOC}_i(\mathcal{I}) - \texttt{FOC}_i(\tilde{\mathcal{I}})| \leq 1/T,$$

as desired.

Analyzing APC. For APC, the above result implies that

$$\texttt{FOC}_i(\widetilde{\mathcal{I}}) < \texttt{FOC}_i(\mathcal{I}) + 1/T.$$

Applying Lemma 2.1, we have that:

$$\texttt{APC}_i(\widetilde{\mathcal{I}}) < \texttt{APC}_i(\mathcal{I})\frac{f_i}{\widetilde{f}_i} + \frac{1}{T\widetilde{f}_i}.$$

Recall that $\widetilde{f}_i > f_i$, so the RHS above is less than $\texttt{APC}_i(\mathcal{I})$ as long as $T > \frac{1}{\texttt{APC}_i(\widetilde{\mathcal{I}})(\widetilde{f}_i - f_i)}$. By the definition of Algorithm 5, every arm must be pulled at least once. Thus, for sufficiently large T, we see that

$$\texttt{APC}_i(\widetilde{\mathcal{I}}) < \texttt{APC}_i(\mathcal{I})$$

as desired.

■

D.2 Analysis of BB_{Pull}(UCB): Proof of Theorem 4.1

Here, we prove the regret bound of Theorem 4.1 for BB_{Pull}(UCB). For convenience, we restate Theorem 4.1 below.

Theorem 4.1. *On any stochastic instance* $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{I}\}$, BB_{Pull}(AAE) *(presented in Algorithm 7) and* BB_{Pull}(UCB) *(presented in Algorithm 8) have regret bound of* $O\left(\sqrt{T\ln(T)\sum_{i\in[K]} 1/f_i}\right)$ *and an instance-dependent regret bound of* $O\left(\sum_{i\in[K]|\Delta_i>0} \frac{\ln T}{\Delta_i f_i}\right)$.

We consider the simulated version of BB_{Pull}(UCB) given by Algorithm 5 applied to UCB. For convenience, we explicitly state this algorithm below (Algorithm 11).

Let us define the same random variables as those used in Algorithm 5, restated for convenience. (Recall that ϕ indexes losses for the time horizon of ALG, Φ is the total number of times ALG is called by BB_{Pull}(ALG), and $\Phi \leq T$ because ALG can be called at most T times.)

- *Losses:* For each round $\phi \in [\Phi]$ of $\textsc{Alg} = \text{UCB}$ and each arm $j \in [K]$, let $\ell'_{j,\phi} := \ell_{j,t}$ be the loss for arm j at a time step t that corresponds to the last time step in block ϕ of $\text{BB}_{\text{Pull}}(\text{UCB})$. Since we are in the stochastic loss setting, $\ell'_{j,\phi}$ is a random variable drawn from the distribution of arm j (with mean $\bar{\ell}_j$) independently across ϕ and j.
- *Feedback realizations:* For all $j \in [K]$ and $\phi \in [T]$, let $Q_{j,\phi} \sim \text{Geom}(f_j)$ for $\phi \in [T]$ be a random variable distributed according to the geometric distribution with parameter equal to the feedback probability of arm j. (These random variables are also fully independent across values of j and ϕ.)

We are now ready to present Algorithm 11. For ease of analysis, we define the lower confidence bounds LCB within the algorithm, even though the algorithm does not ever use these quantities.

Algorithm 11: Simulated version of $\text{BB}_{\text{Pull}}(\text{UCB})$

1 Initialize number of pulls $n_i = 0$ for all $i \in [K]$.
2 Initialize empirical mean $\mu(i) = 0$ for all $i \in [K]$.
3 Initialize $t = 1$ and $\phi = 1$.
4 **while** $t \leq T$ **do**
5   Initialize $i_\phi = 0$.
6   **if** $n_i = 0$ *for any arm* $i \in [K]$ **then**
7   Let i_ϕ be the arm with the smallest index such that $n_{i_\phi} = 0$.
8   **else**
9   For every arm $i \in [K]$, compute $\text{UCB}(i) = \mu(i) + \sqrt{\frac{6 \ln T}{n_i}}$ and $\text{LCB}(i) = \mu(i) - \sqrt{\frac{6 \ln T}{n_i}}$.
10   Let $i_\phi = \text{argmax}_{j \in [K]} \text{UCB}(j)$.
11   **for** $\min(Q_{i_\phi,\phi}, T - t)$ *iterations* **do**
12   Pull $i_\phi = i$ and let $t \leftarrow t + 1$.
13   Observe $\ell'_{i_\phi,\phi} := \ell_{i_\phi,t}$.
14   Update the empirical mean $\mu(i) \leftarrow \frac{n_{i_\phi} \cdot \mu(i)}{n_{i_\phi}+1} - \frac{\ell'_{i_\phi,\phi}}{n_{i_\phi}+1}$.
15   Increment $n_{i_\phi} \leftarrow n_{i_\phi} + 1$.
16   Increment $\phi \leftarrow \phi + 1$.

Since Algorithm 11 is exactly Algorithm 5 applied to AAE, we can apply Lemma C.6 to see that the sequence of arms $\{i_t^{\text{orig}}\}_{t \in [T]}$ pulled by Algorithm 8 is distributed identically to the sequence of arms pulled by $\{i_t^{\text{sim}}\}_{t \in [T]}$ pulled by Algorithm 11.

D.2.1 Lemmas for the analysis of $\text{BB}_{\text{Pull}}(\text{UCB})$

We now show the following intermediate results that build on the standard analysis of UCB [Auer et al., 2002a]. First, we see immediately that for $1 \leq \phi \leq K$, the *if* statement on line 6 of Algorithm 11 will be met, so $i_{\phi=1} = 1$, $i_{\phi=2} = 2, \ldots, i_{\phi=K} = K$. We will handle the regret from these rounds ($\phi = 1 \ldots K$) separately.

We define the following two clean events.

1. First, recall that Φ is the maximum value of ϕ realized by Algorithm 11. Let $E_{\text{UCB,loss}}$ be the "clean" event that at each round $K + 1 \leq \phi \leq \Phi$, for every arm $i \in [K]$, it holds that $\text{LCB}(i) \leq \bar{\ell}_i \leq \text{UCB}(i)$ at line 9 for round ϕ.
2. Let the random variable L_ϕ be equal to the time step t where round ϕ begins (i.e. the value of the variable t at line 5 when i_ϕ is initialized.) if that is reached, and otherwise let L_ϕ be equal to $T + 1$. For each arm i and any value $M_i \geq 0$ let $E_{i,M_i}^{F,\text{UCB}}$ be the event that
$$\sum_{\phi=1}^{\Phi} \min(Q_{i,\phi}, T - (L_\phi - 1)) \cdot \mathbb{1}[i_\phi = i] \leq \frac{6 \cdot M_i}{f_i}.$$

Lemma D.5. *Consider Algorithm 10 evaluated on any given instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$. Let the event $E_{UCB,\ loss}$ be the "clean" event defined above. Then,* $\Pr[E_{UCB,loss}] \geq 1 - 2T^{-3}K$.

Proof. Consider arm $i \in [K]$ and potential number of arm pulls $1 \leq n \leq T$. Let $E^{i,n}_{\text{UCB, loss}}$ be the event that either $n_i = n$ is not reached by the algorithm or $\text{LCB}(i) \leq \bar{\ell}_i \leq \text{UCB}(i)$ at line 9 when $n_i = n$. Let $\tilde{\mu}_n(i)$ be the empirical mean of n i.i.d. samples from the loss distribution for arm i. Following the standard analysis of UCB confidence sets, we see:

$$\mathbb{P}[E^{i,n}_{\text{UCB, loss}}] \geq \mathbb{P}\left[|\tilde{\mu}_n(i) - \bar{\ell}_i| \leq \sqrt{\frac{6 \ln T}{n}}\right] \geq 1 - 2e^{\frac{6n \ln T}{2n}} = 1 - 2T^{-3}.$$

We union bound over $1 \leq n \leq T$ and $i \in [K]$ to obtain $\Pr[E_{\text{UCB,loss}}] \geq 1 - 2T^{-3}K$. ∎

Lemma D.6. *Consider Algorithm 10 evaluated on any given instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$, and consider $i \in \mathcal{A}$. Let M_i be such that $\mathbb{P}[\sum_{\phi=1}^{\Phi} \mathbb{1}[i_\phi = i] \geq M_i] \leq 2T^{-3}K$ and $M_i \geq 6 \ln T$. Let $E^{F,UCB}_{i,M_i}$ be defined as above. Then it holds that $\mathbb{P}[E^{F,UCB}_{i,M_i}] \geq 1 - T^{-4} - 2T^{-3}K$.*

Proof. For the sake of this proof, let's assume that we realize $2T$ random variables $Q_{i,\phi}$ for $1 \leq \phi \leq 2T$ instead of T random variables.

For $1 \leq n \leq T$ such that $\sum_{\phi=1}^{\Phi} \mathbb{1}[i_\phi = i] \geq n$, let $\Phi_{n,i}$ be equal to minimum value $\phi' \geq 1$ such that $\sum_{\phi=1}^{\phi'} 1[i_\phi = i] = n$ (that is, the time step ϕ at which that arm i is pulled by $\textsc{Alg} = \text{UCB}$ for the nth time). For $1 \leq n \leq T$ such that $\sum_{\phi=1}^{\Phi} \mathbb{1}[i_\phi = i] < n$ (i.e. the arm is pulled by $\textsc{Alg} = \text{UCB}$ less than n times), for technical convenience, let $\Phi_{n,i} = T + n \leq 2T$. Observe that:

$$\sum_{\phi=1}^{\Phi} \min(Q_{i,\phi}, T - (L_\phi - 1)) \cdot \mathbb{1}[i_\phi = i] \leq \sum_{\phi=1}^{\Phi} Q_{i,\phi} \cdot \mathbb{1}[i_\phi = i] = \sum_{n \geq 1 \text{ s.t. } \sum_{\phi=1}^{\Phi} \mathbb{1}[i_\phi = i] \geq n} Q_{i,\Phi_{n,i}}.$$

Moreover, by the definition of M_i, we see that

$$\mathbb{P}\left[\sum_{n \geq 1 \text{ s.t. } \sum_{\phi=1}^{\Phi} \mathbb{1}[i_\phi = i] \geq n} Q_{i,\Phi_{n,i}} \leq \sum_{n=1}^{M_i} Q_{i,\Phi_{n,i}}\right] \geq 1 - T^{-2}.$$

We thus focus on bounding

$$\mathbb{P}\left[\sum_{n=1}^{M_i} Q_{i,\Phi_{n,i}} > \frac{6M_i}{f_i}\right].$$

It is easy to see that $Y := \sum_{n=1}^{N_i} Q_{i,\Phi_{n,i}}$ is distributed as the number of Bernoulli trials with parameter f_i needed to observe M_i successes. We can analyze the probability $\mathbb{P}[Y > \frac{6M_i}{f_i}]$ as follows. By definition, this is equal to the probability that fewer than M_i successes are observed after $\frac{6M_i}{f_i}$ Bernoulli trials with parameter f_i. If we let Z_j denote i.i.d. Bernoullis with parameter f_i, this probability can be analyzed by a multiplicative Chernoff bound:

$$\mathbb{P}\left[\sum_{n=1}^{M_i} Q_{i,\Phi_{n,i}} > \frac{6M_i}{f_i}\right] = \mathbb{P}[Y > \frac{6M_i}{f_i}] = \mathbb{P}\left[\sum_{j=1}^{6M_i/f_i} Z_j \leq M_i\right] \leq T^{-4},$$

where we use that $M_i \geq 6 \ln T$.

Union bounding, we obtain that $\mathbb{P}[E^{F,\text{UCB}}_{i,m_i}] \geq 1 - T^{-4} - T^{-2}$. ∎

Lemma D.7. *Consider Algorithm 10 evaluated on any given instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$. If the event $E_{UCB,\ loss}$ holds, then $\sum_{\phi=1}^{\Phi} \mathbb{1}[i_\phi = i] \leq \frac{6 \ln T}{\Delta_i^2}$.*

Proof. This follows from the standard analysis of UCB. Let us condition on $E_{\text{UCB, loss}}$, and let i^* be the arm with optimal mean loss. If $i_\phi = i$ and $\sum_{\phi'=1}^{\phi-1} \mathbb{1}[i_{\phi'} = i] = n$, then it must hold that $-\bar{\ell}_i + \sqrt{\frac{6 \ln T}{n}} = \text{UCB}(i) \geq \text{UCB}(i^*) \geq -\bar{\ell}_{i^*}$. Solving for n, we obtain that

$$n \leq \frac{6 \ln T}{\Delta_i^2}.$$

■

Now, we are ready to prove Theorem 4.1.

Proof of Theorem 4.1 for $\text{BB}_{\text{Pull}}(\text{UCB})$*.* By Lemma C.6, the sequence of arms $\{i_t^{\text{orig}}\}_{t \in [T]}$ pulled by Algorithm 7 is distributed identically to the sequence of arms pulled by $\{i_t^{\text{sim}}\}_{t \in [T]}$ pulled by Algorithm 10. Let $M_i = \frac{6 \ln T}{\Delta_i^2}$ for $i \in [K]$ and let the event E be defined to be $E_{\text{UCB, loss}} \cap E_{1, M_1}^{\text{F, UCB}} \cap \ldots E_{1, M_K}^{\text{F, UCB}}$. We apply Lemma D.5, Lemma D.6, and Lemma D.7 to see that E occurs with probability at least $1 - 2T^{-3}K - 2T^{-3}K^2 - KT^{-4}$. When T is sufficiently large, $\mathbb{P}[E] \geq 1 - T^{-2}$, so the event that E does not occur contributes negligibly to the regret. Let us condition on E for the remainder of the analysis.

For each arm i, let $\Delta_i = \bar{\ell}_i - \min_j \bar{\ell}_j$ be the suboptimality gap. Let M_i be the number of time steps where arm i is pulled over the course of the algorithm. The regret is equal to:

$$\sum_{1 \leq i \leq K | \Delta_i > 0} \Delta_i \cdot M_i.$$

We first observe by event $E_{i, M_i}^{\text{F, UCB}}$ that if $\Delta_i > 0$, then arm i is pulled at most $\frac{36 \ln T}{\Delta_i^2 f_i}$ times. The instance-independent and instance-dependent regret bounds now follow from Lemma D.4.

■

D.3 Analysis of BB_{DA}(AAE)

We prove the monotonicity properties of $\text{BB}_{\text{DA}}(\text{AAE})$.As before, we also construct a simulated version of Algorithm 9. We formalize this simulated version in Algorithm 12.

Let us define the following random variables. (Recall that ϕ indexes losses for the time horizon of Alg, Φ is the total number of times Alg is called by $\text{BB}_{\text{Pull}}(\text{Alg})$, and $\Phi \leq T$ because Alg can be called at most T times.)

- *Losses:* For each round $\phi \in [T]$ of Alg = AAE and each arm $j \in [K]$, let $\ell'_{j,\phi}$ denote a stochastic loss sampled from the distribution of arm j (with mean $\bar{\ell}_j$). These random variables are fully independent across values of j and ϕ. Note that unlike in Algorithm 10 or 11, it is not guaranteed that $\ell_{j,\phi}^{\text{AAE}}$ observed by AAE is $\ell'_{j,\phi}$, because with a fixed block size, there will always be some likelihood that no feedback is observed.
- *Feedback probabilities:* Let $U_{j,\phi} \sim \text{Bern}(1 - (1 - f_j)^{B_j})$ for $j \in [K]$ and $\phi \in [T]$ denote the indicator variable for whether feedback will be observed in block ϕ, where $B_j = \lceil \frac{3 \ln T}{\min_i f_i} (1 + f_j) \rceil$. (These random variables are also fully independent across values of j and ϕ.)

Again, note that Algorithm 12 is a direct application of Algorithm 6 to AAE (lines 7-11 in Algorithm 12 reflect Algorithm 6, while the rest are for Alg = AAE). This allows us use Lemma C.8 directly to argue that the arms selected by Algorithm 12 are distributed identically to those selected by Algorithm 9.

For convenience, we restate Theorem 4.3 below.

Theorem 4.3. *Fix a stochastic instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$. Let i be such that $\bar{\ell}_i > \min_{j \in [K]} \bar{\ell}_j$. Let $\widetilde{f}_i > f_i$, and let $\widetilde{\mathcal{I}} = \{\mathcal{A}, \mathcal{F}(\widetilde{i}), \mathcal{L}\}$. For any $f^\star \leq \min_i f_i$ and sufficiently large T,* $\text{BB}_{\text{DA}}(\text{AAE}, f^\star)$ *satisfies* $\texttt{APC}_i(\widetilde{\mathcal{I}}) > \texttt{APC}_i(\mathcal{I})$ *and* $\texttt{FOC}_i(\widetilde{\mathcal{I}}) > \texttt{FOC}_i(\mathcal{I})$.

Algorithm 12: Simulated version of $\mathrm{BB}_{\mathrm{DA}}(\mathrm{AAE})$

For arm $i \in [K]$, set $B_i = \lceil (1+f_i) \cdot \frac{3 \ln T}{f^\star} \rceil$.
Initialize $t = 1$, $\phi = 1$ and phase $s = 1$. Maintain active set A; start with $A := [K]$.
while $t \leq T$ **do**
  Start phase s.
  for *arm* $j \in A$ **do**
    for $2^{2s+1} \cdot \ln T$ *iterations* **do**
      for $\min(B_j, T - t)$ *iterations* **do**
        Pull $i_t = j$ and let $t \leftarrow t + 1$.
        if $t = T$ **then return**.
      if $U_{j,\phi} = 1$ **then** observe $\ell^{\mathrm{AAE}}_{j,\phi} := \ell'_{j,\phi}$ and let $\phi \leftarrow \phi + 1$.
      else observe $\ell^{\mathrm{AAE}}_{j,\phi} := 1$ and let $\phi \leftarrow \phi + 1$.
    Let $\psi_s(j) := \{\phi - 8 \cdot 2^{2s} \cdot \ln T, \ldots, \phi\}$ be the set of ϕ timesteps in which arm j was pulled for phase s.
    Compute empirical mean $\mu_s(j) = \frac{1}{8 \cdot 2^{2s} \cdot \ln T} \sum_{\phi \in \psi_s(j)} \ell^{\mathrm{AAE}}_{j,\phi}$.
    Set $\mathrm{LCB}_s(i) = \mu_s(i) - 2^{-s}$ and $\mathrm{UCB}_s(i) = \mu_s(i) + 2^{-s}$.
  For any arm $i \in A$ where $\exists j \in A$ such that $\mathrm{LCB}_s(j) > \mathrm{UCB}_s(i)$, remove i from A.
  Increment $s \leftarrow s + 1$.

The proof of Theorem 4.3 follows from adjusting the ideas in the proof of Theorem 4.2 and Theorem 3.6. The high-level intuition is that for a sufficiently large T, we must reach a phase in both instances where i is eliminated, if i is a suboptimal arm. If ALG takes the same number of phases s^* to eliminate i in both instances (which a similar coupling argument as above will ensure), then by definition of the block sizes in each algorithm, $\mathtt{APC}_i(\widetilde{\mathcal{I}}) = s^* \cdot \frac{3 \ln T}{f^\star} \cdot (1 + \widetilde{f}_i)$ and $\mathtt{APC}_i(\mathcal{I}) = s^* \cdot \frac{3 \ln T}{f^\star} \cdot (1 + f_i)$.

In these results, we use the following notation. (Items 1 and 2 are analogous notation to in the analysis of $\mathrm{BB}_{\mathrm{Pull}}(\mathrm{AAE})$ from Appendix D.1.2, restated below for convenience.)

1. Let S be a random variable denoting the maximum value of the variable s reached in Algorithm 12 on $\widetilde{\mathcal{I}}$. (That is, S denotes the number of phases that Algorithm 12 *begins*.) Note that $S \leq T$ with probability 1.

2. Let E_{loss} be the "clean" event that at each phase $1 \leq s \leq S - 1$, for every arm $i \in [K]$, it holds that $\mathrm{LCB}_s(i) \leq \bar{\ell}_i \leq \mathrm{UCB}_s(i)$.

3. Let $\widetilde{A}$ denote the active set on $\widetilde{\mathcal{I}}$, and A denote the active set on $\mathcal{I}$.

4. Let E be the "clean" event that $U_{j,\phi} = \widetilde{U}_{j,\phi} = 1$ for all $j \in [K], \phi \in [T]$.

Again, we begin by arguing that "clean events" occur with high probability.

Lemma D.8. *Consider Algorithm 12 evaluated on any given instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$. Condition on the event E defined above, i.e. that $U_{j,\phi} = \widetilde{U}_{j,\phi} = 1$ for all $j \in [K], \phi \in [T]$. Let the event E_{loss} be the defined as above. Then, $\Pr[E_{loss}] \geq 1 - 2T^{-3}K$.*

Lemma D.9. *Consider Algorithm 12 evaluated on any given instance $\mathcal{I} = \{\mathcal{A}, \mathcal{F}, \mathcal{L}\}$ with time horizon T. Suppose that the events E_{loss} and E both hold. Then, the optimal arm $i^\star = \arg\min_j \bar{\ell}_j$ is never removed from A. Moreover, at every phase $1 \leq s \leq S - 1$, if $i \in A$ at the end of phase s (i.e. after 13 in Algorithm 10), then*

$$\bar{\ell}_i - \min_j \bar{\ell}_j \leq 4 \cdot 2^{-s}.$$

The proof of Lemma D.8 is identical to the proof of Lemma D.1, and the proof of Lemma D.9 is identical to the proof of Lemma D.2, because the analysis is specific to AAE, rather than the black-box transformations; we accordingly omit them here.

Proof of Theorem 4.3 (Monotonicity of $\mathrm{BB}_{\mathrm{DA}}(\mathrm{AAE})$*).* As before, we condition on a series of clean events, construct a coupling, then analyze the phase at which arm i must be eliminated.

Step 1: Condition on feedback observations. This step is identical to Step 1 in the proof of Theorem 3.6. Let E be the event that $U_{j,\phi} = \widetilde{U}_{j,\phi} = 1$ for all $j \in [K], \phi \in [T]$. By Lemma C.4, $\Pr[E] \geq 1 - 1/T^2$. Then, for any $\phi > T$, we let $U_{j,\phi}$ and $\widetilde{U}_{j,\phi}$ take on arbitrary values in $\{0, 1\}$. We condition on E for the following steps.

Step 2: Couple arms pulled by Alg across instances $\mathcal{I}$ and $\widetilde{\mathcal{I}}$. We couple arms in the same way as in the proof of Theorem 3.2. We can apply Lemma C.7, letting $\Phi^* = T$, so that $(i_1^{\mathrm{ALG}}, \ldots, i_{\Phi^*}^{\mathrm{ALG}})$ and $(\tilde{i}_1^{\mathrm{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\mathrm{ALG}})$ are identically distributed. This allows us to consider a joint probability distribution over $(i_1^{\mathrm{ALG}}, \ldots, i_{\Phi^*}^{\mathrm{ALG}}, \tilde{i}_1^{\mathrm{ALG}}, \ldots, \tilde{i}_{\Phi^*}^{\mathrm{ALG}})$ such that $i_\phi^{\mathrm{ALG}} = \tilde{i}_\phi^{\mathrm{ALG}}$ for all $\phi \in [\Phi^*]$.

Step 3: Condition on E_{loss}. This step is similar to Step 2 in the proof of Theorem 4.2. We will condition on E_{loss}, i.e. that confidence bounds are correct: $\mathrm{LCB}_s(i) \leq \bar{\ell}_i \leq \mathrm{UCB}_s(i)$, for every phase s and every arm $i \in [K]$. By Lemma D.8, we have that $\Pr[E_{\mathrm{loss}}] \geq 1 - 2T^3K$.

Step 4: Run Algorithm 12 and analyze `APC`. This step is similar to Step 3 in the proof of Theorem 4.2. However the interpretation of the number of rounds ϕ in which any arm is selected by AAE is different. While in Algorithm 10, the number of rounds ϕ was equivalent to the number of feedback observations for that arm, the number of rounds ϕ specifies the *number of blocks* in which that arm is pulled by Algorithm 12. For each arm i, within each block where the arm i is selected, the arm will be pulled exactly $\lceil (1+f_i) \cdot \frac{3\ln T}{f^\star} \rceil$ times on $\mathcal{I}$, and exactly $\lceil (1+\widetilde{f}_i) \cdot \frac{3\ln T}{f^\star} \rceil$ times on $\widetilde{\mathcal{I}}$, by the definition of Algorithm 12.

We first claim that arm i will be eliminated before the end of the time horizon is reached on both instances. Because we have conditioned on E_{loss}, we can apply Lemma D.9 to argue that from phase $s \geq s' := 3 - \log(\Delta_i)$ onwards, arm i is guaranteed not to be pulled. Furthermore, to our coupling, in each phase s, arm i is in the active set A for Algorithm 12 on $\mathcal{I}$ if and only if it is also in the active set $\widetilde{A}$ for Algorithm 12 on $\widetilde{\mathcal{I}}$. To show that arm i is eliminated, we next count the number of times that an arm is pulled in a given phase s. For any phase s on $\mathcal{I}$, the total number of (t-indexed) rounds *within* that phase is at most $2 \cdot K \cdot \frac{3\ln T}{f^\star} \cdot 2^{2s+1} \ln T = 2^{2s+2} \cdot \frac{3K(\ln T)^2}{f^\star}$. (To see this, note that A contains at most K arms, each of which have a block size of at most multiplied by the maximum block size per arm of $2 \cdot \frac{3\ln T}{f^\star}$, multiplied by $2^{2s+3}\ln T$ pulls per arm per block within phase s.) Let t_s be the total number of rounds elapsed by the end of phase s. Because each previous phase takes $1/4$ as many t-indexed rounds as the current phase, we can see that for any s,

$$t_s \leq \frac{4}{3} \cdot 2^{2s+4} \cdot \frac{3K(\ln T)^2}{f^\star} = 2^{2s} \cdot \frac{K(\ln T)^2}{f^\star}.$$

Then, for $s' = 3 - \log(\Delta_i)$, we have

$$t_{s'} \leq 2^{2(3-\log(\Delta_i))} \cdot \frac{K(\ln T)^2}{f^\star} = \frac{64}{\Delta_i^2} \cdot \frac{K(\ln T)^2}{f^\star}.$$

We will have $t_{s'} \leq T$ as long as $\Delta_i \geq \frac{8\sqrt{K}\ln T}{\sqrt{Tf^\star}}$ (which holds for any $f^\star$ and Δ_i as $T \to \infty$). Note that due to our coupling, this analysis holds for $\widetilde{\mathcal{I}}$ as well. Altogether, this proves the number of blocks in which arm i will be eliminated before the end of the time horizon on both instances.

We are now ready to analyze `APC` on $\mathcal{I}$ and $\widetilde{\mathcal{I}}$. To formalize the rest of our analysis, we introduce the following additional notation. Let $\mathtt{APC}_i^{[T]}(\mathcal{I})$ be the number of times arm i is pulled in timesteps on $\mathcal{I}$, and let $\mathtt{APC}_i^{[T]}(\widetilde{\mathcal{I}})$ be the number of times arm i is pulled in timesteps on $\widetilde{\mathcal{I}}$. Since we have conditioned on $U_{j,\phi}$, $\widetilde{U}_{j,\phi}$, i_ϕ^{ALG}, $\tilde{i}_\phi^{\mathrm{ALG}}$, we see that at this point $\mathtt{APC}_i^{[T]}(\mathcal{I})$ and $\mathtt{APC}_i^{[T]}(\widetilde{\mathcal{I}})$ are both deterministic.

We show that $\mathtt{APC}_i^{[T]}(\widetilde{\mathcal{I}}) > \mathtt{APC}_i^{[T]}(\mathcal{I})$. We observe that the number of rounds ϕ in which arm i is pulled on $\mathcal{I}$ is equal to the number of rounds ϕ in which arm i is pulled on $\widetilde{\mathcal{I}}$ (this is using the fact that arm i will be eliminated before the end of the time horizon on both instances and using the property of the coupling that

$\tilde{i}^{\textsc{Alg}}_\phi = i^{\textsc{Alg}}_\phi$ for all rounds ϕ). Using the equality in the number of blocks in which arm i is pulled on each instance, we see that:

$$\frac{\texttt{APC}_i^{[T]}(\widetilde{\mathcal{I}})}{\texttt{APC}_i^{[T]}(\mathcal{I})} = \frac{\tilde{B}_i}{B_i} = \frac{\lceil (1+\tilde{f}_i) \cdot \frac{3\ln T}{f^*} \rceil}{\lceil (1+f_i) \cdot \frac{3\ln T}{f^*} \rceil},$$

which is strictly greater than 1 as long as T is sufficiently large. This implies that $\texttt{APC}_i^{[T]}(\widetilde{\mathcal{I}}) > \texttt{APC}_i^{[T]}(\mathcal{I})$ as desired.

We can apply the law of total expectation over the sequences $U_{j,\phi}$, $\widetilde{U}_{j,\phi}$, $i^{\textsc{Alg}}_\phi$, $\tilde{i}^{\textsc{Alg}}_\phi$. Let $\texttt{APC}_i(\mathcal{I} \mid E, E_{\text{loss}})$ notate the metric $\texttt{APC}_i$ on instance $\mathcal{I}$ conditioned on the clean events E and E_{loss}. We see that:

$$\texttt{APC}_i(\widetilde{\mathcal{I}} \mid E, E_{\text{loss}}) - \texttt{APC}_i(\mathcal{I} \mid E, E_{\text{loss}}) = \mathbb{E}\left[\texttt{APC}_i^{[T]}(\widetilde{\mathcal{I}}) - \texttt{APC}_i^{[T]}(\mathcal{I}) \mid E, E_{\text{loss}}\right] > 0.$$

This means that:

$$\texttt{APC}_i(\widetilde{\mathcal{I}} \mid E, E_{\text{loss}}) > \texttt{APC}_i(\mathcal{I} \mid E, E_{\text{loss}}).$$

Step 4: Handle the APC contributions of the conditioning steps. Finally, we handle the possibility that the events E and E_{loss} do not hold, i.e., that we do not see feedback in every block on each instance, and the possibility that our confidence bounds are not good. Because we first conditioned on E and then conditioned on E_{loss}, we will remove the conditioning in the reverse order:

$$\left|\texttt{APC}_i(\widetilde{\mathcal{I}}|E) - \frac{1+\widetilde{f}_i}{1+f_i} \cdot \texttt{APC}_i(\mathcal{I}|E)\right| \leq \frac{1}{T^2}$$

$$\implies \left|\texttt{APC}_i(\widetilde{\mathcal{I}}) - \frac{1+\widetilde{f}_i}{1+f_i} \cdot \texttt{APC}_i(\mathcal{I})\right| \leq \frac{1}{T}$$

Combining the above result with Lemma 2.1 implies that FOC must be strictly increasing in f_i. ■

E Supplemental Materials for Section 4.3

E.1 Linear Regret of Standard EXP3

We first illustrate how the standard EXP3 algorithm may achieve linear regret in the probabilistic feedback setting.

Proposition E.1 (Regret of Standard EXP3)**.** *Standard EXP3 obtains regret $\Omega(T)$ when arms have $f_i \neq 1, \forall i \in [K]$.*

Proof. We work with utilities here instead of losses because the intuition is clearer. To obtain the result for losses, one can use the standard transformation that loss $= 1-$ utilities.

Consider two instances:

Instance 1. Let there be two arms. Arm 1 has reward distribution u_1 with expectation $\mathbb{E}[u_1] = 1$ and $f_1 = 1/4$. Arm 2 has reward distribution u_2 with expectation $\mathbb{E}[u_2] = 1/2$ and $f_2 = 1$.

Instance 2. Let there be two arms. Arm $1'$ has reward distribution such that with probability $3/4$, $u_{1'} = 0$, and with probability $1/4$, $u_{1'} \sim u_1$ (that is, sample the deterministic value 0 with probability $3/4$, and sample the reward distribution u_1 with probability $1/4$). Arm 2 has the reward distribution $u_{2'}$. Let $f_{1'} = f_{2'} = 1$.

Fix an infinite tape of independent draws from u_1 (call it p_{u_1}) and fix an infinite tape of draws from u_2 (call it p_{u_2}). Fix an infinite tape of draws from a Bernoulli distribution with rate $1/4$ (call it p_{f_1}) and for all $\pi \in [0,1]$, fix an infinite tape $p_{\mathcal{B}(\pi)}$ of random draws from a Bernoulli distribution with rate π.

We will define the trajectory of EXP3 run on either instance according to these sequences, and show that fixing this sequence of draws, EXP3 must pull the same arms and maintain the same values of $w_{i,t}$ and

$\pi_{i,t}$ across all rounds for corresponding arms across the instances. We call the algorithm running on the respective instances World and World'.

Base case: Let $t = 0$. Then, both algorithms have initialized the weights to 1, so trivially, the weights are the same. Moreover, by these weights, $\pi_1 = \pi_{1'} = \pi_2 = \pi_{2'} = 1/2$. Then, we use the first bit in the tape $p_{\mathcal{B}(1/2)}$ to determine which arm to pull in $t = 1$. If this value is 0, then pull arm 1 in World and World'; else, pull arm 2 in both Worlds.

Inductive case: Suppose the algorithms have pulled the exact same arms up to time $t-1$, and have maintained the same weights and probabilities so that $w_{1,t} = w_{1',t}$ and $w_{2,t} = w_{2',t}$ and $\pi_{1,t} = \pi_{1',t}$ and $\pi_{2,t} = \pi_{2',t}$. Now, use the next unused bit in the tape $p_{\mathcal{B}(\pi_{1,t})}$ to determine which arm to pull. If this value is 0, pull arm 1 in both Instance 1 and Instance 2; else, pull arm 2 in both Instances. This realizes the correct probabilities in both instances. Now, if arm 2 is pulled, let the reward be the next available draw from the tape p_{u_2}; this realizes the correct reward distribution in both instances. If arm 1 is pulled, first take the next unused bit in p_{f_1}. If it is 0, in both worlds set the observed utility to 0, making the estimator of the utility $\hat{u}1, t = \hat{u}1', t = 0$. If the bit is 1, then draw the observed utility by taking the next unused bit from p_{u_1}, and use it to compute the utility estimator in both worlds, so that again, $\hat{u}1, t = \hat{u}1', t$. This realizes the correct distribution of the estimator in both worlds.

Because the estimators are equal, and by the induction hypothesis $w_{1,t} = w_{1',t}$ and $w_{2,t} = w_{2',t}$ and $\pi_{1,t} = \pi_{1',t}$ and $\pi_{2,t} = \pi_{2',t}$, we have that $w_{1,t+1} = w_{1',t+1}$ and $w_{2,t+1} = w_{2',t+1}$ and $\pi_{1,t+1} = \pi_{1',t+1}$ and $\pi_{2,t+1} = \pi_{2',t+1}$.

EXP3 is guaranteed to get sublinear regret when the f_i's are uniformly 1; thus, it must get sublinear regret in instance 2, and thus must pull arm 1 a subconstant number of times. It directly follows that EXP3 must then also pull arm 1 a sublinear (in T) number of times in instance 2, meaning that it pulls arm 2 (the suboptimal arm in instance 2) a linear (in T) number of times, thus incurring linear regret in instance 1. ■

E.2 Monotonicity of 3-Phase EXP3

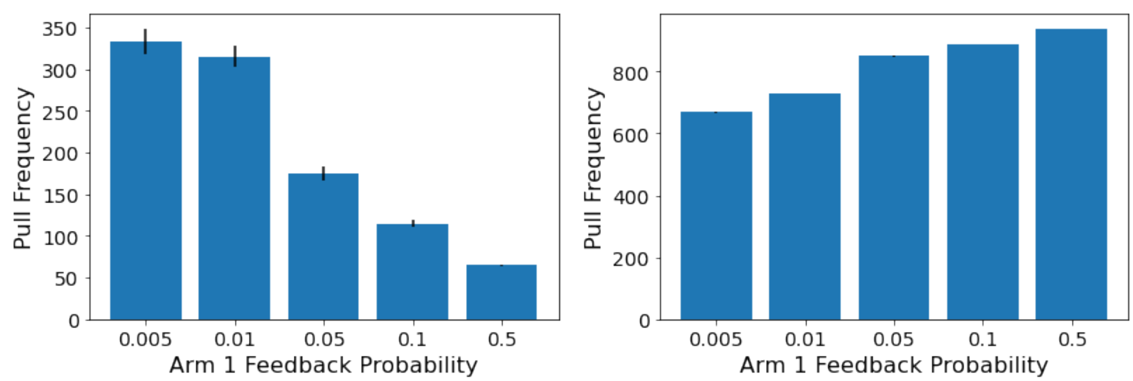

Figure 3: Analysis of APC for a simplified version of 3-Phase EXP3 (Algorithm 4) in two instances where $K = 2$ and $T = 1000$. In Instance 1 (left), Arm 1 has constant loss 0.9 and Arm 2 has constant loss 0.1; In Instance 2 (right), Arm 1 has constant loss 0.1 and Arm 2 has constant loss 0.9. APC is strictly negative monotonic in Instance 1 and strictly positive in Instance 2. These differing directions of monotonicity suggest that Algorithm 4 does not exhibit clean monotonicity guarantees.

E.3 Regret of 3-Phase EXP3 (Algorithm 4)

We first discuss the regret bound provided in 4.4 and its implications in the context of related work; we prove this result in the remainder of the section.

E.3.1 Lemmas for Proof of Theorem 4.4

We first provide several useful lemmas for formalizing the proof of Theorem 4.4.

We start by proving that the estimates built on Phase 2 of the algorithm are close to the true f_i's. As is customary, we prove our results for the pseudo-regret, which coincides with the expected regret for the case of an oblivious adversary.

Lemma E.2. *For all $i \in [K]$, the estimate P_i^E obtained in Phase 2 of Algorithm 4 satisfies $\mathbb{E}[P_i^E] = 1/f_i$ and $\mathbb{E}[(P_i^E)^2] \leq 2/f_i^2$.*

Proof. To see that $\mathbb{E}[P_i^E] = 1/f_i$, note that P_i^E is distributed as a geometric distribution with parameter f_i. To see that $\mathbb{E}[(P_i^E)^2] \leq 2/f_i^2$, note that

$$\mathbb{E}\left[\left(P_i^E\right)^2\right] = \mathbb{E}\left[P_i^E\right]^2 + \mathtt{Var}\left(P_i^E\right) \leq \frac{1}{f_i^2} + \frac{1}{f_i^2} = \frac{2}{f_i^2}.$$

■

Lemma E.3. *The estimates P_i^{LR} obtained in Phase 1 of Algorithm 4 satisfy the tail bound:*

$$\Pr\left[\forall i \in [K], \frac{1}{2f_i} \leq P_i^{LR} \leq \frac{2}{f_i}\right] \geq 1 - \frac{2}{T}.$$

Proof. Since we can union over all $i \in [K]$, it suffices to show that the following tail bound for each $i \in [K]$:

$$\Pr\left[P_i^{LR} > \frac{2}{f_i}\right] \leq \frac{1}{TK} \text{ and } \Pr\left[P_i^{LR} < \frac{1}{2f_i}\right] \leq \frac{1}{TK}.$$

First, we show the upper tail bound. Since $N \cdot P_i^{LR}$ is a random variable counting the number of trials until N observations are made, we can rewrite $\Pr[P_i^{LR} > \frac{2}{f_i}]$ as a tail bound on a binomial random variable. More specifically, note that $\Pr[P_i^{LR} > \frac{2}{f_i}]$ is equal to the probability that less than N observations appear after $\frac{2N}{f_i}$ trials which is equal to $\Pr[Y < N]$, where $Y \sim \text{Bin}(2N/f_i, f_i)$. We can now apply a multiplicative Chernoff bound to obtain a bound on $\Pr[Y_u < N]$. Let $Z_1, \ldots, Z_{2N/f_i}$ be a sequence of Bernoulli random variables with probability f_i, then we see that:

$$\begin{aligned}\Pr\left[P_i^{LR} > \frac{2}{f_i}\right] &= \Pr[Y_u < N] = \Pr\left[\sum_{j=1}^{2N/f_i} Z_j < N\right] \\ &= \Pr\left[\sum_{j=1}^{2N/f_i} Z_j < 0.5 \cdot \mathbb{E}\left[\sum_{j=1}^{2N/f_i} Z_j\right]\right] \leq e^{-\frac{2N}{8}} = e^{-N/4},\end{aligned}$$

by applying a multiplicative Chernoff bound. We thus obtain a tail bound of at most $1/(TK)$ with our setting of $N = 8\log(TK)$.

Next, let's show the lower tail bound. As before, since $N \cdot P_i^{LR}$ is a random variable counting the number of trials until N observations are made, we can rewrite $\Pr[P_i^{LR} < \frac{1}{2f_i}]$ as a tail bound on a binomial random variable. More specifically, note that $\Pr[P_i^{LR} < \frac{1}{2f_i}]$ is equal to the probability that at least N observations appear after $\frac{N}{2f_i} - 1$ trials which is equal to $\Pr[Y_l \geq N]$, where $Y' \sim \text{Bin}(0.5N/f_i - 1, f_i)$. We can now apply a multiplicative Chernoff bound to obtain a bound on $\Pr[Y' \geq N]$. Let $Z_1, \ldots, Z_{0.5N/f_i - 1}$ be a sequence of Bernoulli random variables with probability f_i, then we see that:

$$\begin{aligned}\Pr\left[P_i^{LR} < \frac{0.5}{f_i}\right] &= \Pr[Y_l \geq N] = \Pr\left[\sum_{j=1}^{0.5N/f_i - 1} Z_j \geq N\right] \\ &= \Pr\left[\sum_{j=1}^{0.5N/f_i - 1} Z_j > 2 \cdot \mathbb{E}\left[\sum_{j=1}^{0.5N/f_i - 1} Z_j\right]\right] \leq e^{-\frac{(0.5N - f_i)}{3}} = e^{-N/8},\end{aligned}$$

by applying a multiplicative Chernoff bound. Our setting of $N = 8\log(TK)$ thus ensures a tail bound of at most $1/(TK)$.

■

Next, we analyze the regret of Phase 3 conditional on the event that the estimates are close to the f_i's.

Lemma E.4. *Conditional on the estimates $\{P_i^{LR}\}_{i\in[K]}$ being close to $\{1/f_i\}_{i\in[K]}$ as in Lemma E.3, the regret incurred in Phase 3 is*

$$\sqrt{\frac{2T\log K}{\sum_{i\in[K]}\frac{1}{f_i}}}$$

The proof of Lemma E.4 builds on the standard analysis of the loss estimator that EXP3 maintains (e.g. Hazan et al. [2016]), which we state and reprove here for completeness.

Lemma E.5 (EXP3 bound on estimated rewards)**.** *Let $i^\star = \arg\min_{i\in[K]}\sum_{t\in[T]}\ell_{i,t}$ be the optimal arm in hindsight. Then, for the loss estimator $\widehat{\ell}_{i,t}$ that EXP3 maintains it holds that:*

$$-\sum_{t\in[T]}\widehat{\ell}_{i^\star,t} \le -\sum_{t\in[T]}\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}_{i,t} + \eta\sum_{t\in[T]}\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}^2_{i,t} + \frac{\log K}{\eta}$$

Proof of Lemma E.5. Let $W_t = \sum_{i\in[K]} w_{i,t}$ be the sum of weights of all arms for round t. This serves as our potential function. Our goal is to upper and lower bound quantity W_T. For the lower bound:

$$\begin{aligned}
W_T &= \sum_{i\in[K]} w_{i,T} && \text{(by definition)}\\
&\ge w_{i^\star,T} && (w_{i,t}\ge 0, \forall i,t)\\
&= \exp\left(-\eta\left(\sum_{t\in[T]}\widehat{\ell}_{i^\star,t}\right)\right) && (4)
\end{aligned}$$

For the upper bound:

$$\begin{aligned}
W_T &= W_{T-1}\sum_{i\in[K]}\pi_{i,t}\cdot\exp\left(-\eta\widehat{\ell}_{i,t}\right) && \text{(by definition of the update step)}\\
&\le W_{T-1}\sum_{i\in[K]}\pi_{i,t}\cdot\left(1-\eta\widehat{\ell}_{i,t}+\eta^2\widehat{\ell}^2_{i,t}\right) && (e^{-x}\le 1-x+x^2 \text{ for } x\ge 0)\\
&= W_{T-1}\left(1-\eta\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}_{i,t}+\eta^2\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}^2_{i,t}\right) && \left(\textstyle\sum_{i\in[K]}\pi_{i,t}=1\right)\\
&\le W_{T-1}\exp\left(-\eta\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}_{i,t}+\eta^2\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}^2_{i,t}\right) && (1+x\le e^x \text{ for all } x)\\
&= W_0\exp\left(-\eta\sum_{t\in[T]}\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}_{i,t}+\eta^2\sum_{t\in[T]}\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}^2_{i,t}\right) && \text{(telescoping for } W_t\text{, for } t\in[T-1])\\
&= K\exp\left(-\eta\sum_{t\in[T]}\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}_{i,t}+\eta^2\sum_{t\in[T]}\sum_{i\in[K]}\pi_{i,t}\cdot\widehat{\ell}^2_{i,t}\right) && (5)
\end{aligned}$$

where the last inequality comes from the fact that $W_0 = K$. Combining Equations (4) and (5), taking the log on both sides, then dividing both sides by η we get the result. ■

Now we prove Lemma E.4.

Proof of Lemma E.4. We first analyze the first and the second moments of the estimator $\widehat{u}_{i,t}$. Let H_{Ph1} encompass the randomness of Phase 1; H_{Ph2} encompass the randomness of Phase 2; H_{t-1} encompass the randomness of the algorithm in Phase 3 up to time $t-1$; and H_{Alg} encompass the randomness of the algorithm at time t.

For the first moment, we have:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\ell}_{i,t} \mid H_{Ph1}\right] &= \mathbb{E}_{H_{Ph2}}\left[\mathbb{E}_{H_{t-1}}\left[\mathbb{E}_{H_{Alg}}\left[\widehat{\ell}_{i,t}|H_{t-1}, H_{Ph2}, H_{Ph1}\right]|H_{Ph2}, H_{Ph1}\right] \mid H_{Ph1}\right] \\
&=_{(A)} \mathbb{E}_{H_{Ph2}}\left[\ell_{i,t}\cdot f_i \cdot P_i^E|H_{Ph1}\right] \\
&= \ell_{i,t}\cdot f_i\cdot \mathbb{E}_{H_{Ph2}}\left[P_i^E|H_{Ph1}\right] \\
&=_{(B)} \ell_{i,t}, \qquad (6)
\end{aligned}$$

where (A) follows from the fact that $\mathbb{E}_{H_{t-1}}\left[\mathbb{E}_{H_{Alg}}\left[\widehat{\ell}_{i,t}|H_{t-1}, H_{Ph2}, H_{Ph1}\right]|H_{Ph2}, H_{Ph1}\right] = \ell_{i,t} f_i P_i^E$ and (B) follows from Lemma E.2.

For the second moment, we have:

$$\begin{aligned}
\mathbb{E}\left[\pi_{i,t}\widehat{\ell}_{i,t}^2 \mid H_{Ph1}\right] &= \mathbb{E}_{H_{Ph2}}\left[\mathbb{E}_{H_{t-1}}\left[\mathbb{E}_{H_{Alg}}\left[\pi_{i,t}\widehat{\ell}_{i,t}^2|H_{t-1}, H_{Ph2}, H_{Ph1}\right] \mid H_{Ph2}, H_{Ph1}\right]\Big|H_{Ph1}\right] \\
&= \mathbb{E}_{H_{Ph2}}\left[\mathbb{E}_{H_{t-1}}\left[\pi_{i,t} f_i \cdot \frac{\ell_{i,t}^2}{\pi_{i,t}}\cdot (P_i^E)^2 \mid H_{Ph2}, H_{Ph1}\right]\Big|H_{Ph1}\right] \\
&= \mathbb{E}_{H_{Ph2}}\left[f_i\cdot \ell_{i,t}^2\cdot (P_i^E)^2\Big|H_{Ph1}\right] \\
&= f_i\cdot \ell_{i,t}^2\cdot \mathbb{E}_{H_{Ph2}}\left[(P_i^E)^2\Big|H_{Ph1}\right] \\
&= f_i\cdot \ell_{i,t}^2\cdot \mathbb{E}_{H_{Ph2}}\left[(P_i^E)^2\right] \\
&\le \frac{2\ell_{i,t}^2}{f_i}, \qquad (7)
\end{aligned}$$

where the last inequality is due to the fact that $\mathbb{E}[(P_i^E)^2] \le 2/f_i^2$ (see Lemma E.2).

Taking expectations on both sides of Lemma E.5 and substituting Equation (6) and Equation (7) we get:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t\in[T]} \ell_{i^\star,t} - \sum_{t\in[T]} \ell_{i_t,t} \Big| H_{Ph1}\right] &\leq \mathbb{E}\left[\eta \sum_{t\in[T]}\sum_{i\in[K]} \frac{2\ell_{i,t}^2}{f_i} + \frac{\log K}{\eta}\Big| H_{Ph1}\right] \\
&\leq_{(A)} \mathbb{E}\left[2\eta T \sum_{i\in[K]} \frac{1}{f_i} + \frac{\log K}{\eta}\Big| H_{Ph1}\right] \\
&= \mathbb{E}\left[2\sqrt{\frac{\log K}{T\sum_{i\in[K]} P_i^{LR}}}\, T \sum_{i\in[K]} \frac{1}{f_i} + \frac{\log K}{\sqrt{\frac{\log K}{T\sum_{i\in[K]} P_i^{LR}}}}\Big| H_{Ph1}\right] \\
&\leq_{(B)} 2\sqrt{\frac{\log K}{T\sum_{i\in[K]} 0.5(1/f_i)}}\, T \sum_{i\in[K]} \frac{1}{f_i} + \sqrt{T\left(\sum_{i\in[K]} \frac{2}{f_i}\right)\log(K)} \\
&\leq 2\sqrt{2T\left(\sum_{i\in[K]} \frac{1}{f_i}\right)\log(K)} + \sqrt{2T\left(\sum_{i\in[K]} \frac{1}{f_i}\right)\log(K)} \\
&= 4\sqrt{2}\sqrt{T\log(K)\sum_{i\in[K]} \frac{1}{f_i}},
\end{aligned}
$$

where (A) follows from the fact that $\ell_{i,t} \leq 1$ and (B) follows from the fact that we conditioned on Lemma E.3. ■

E.3.2 Proof of Theorem 4.4

We are now ready to prove Theorem 4.4.

Proof of Theorem 4.4. The regret Algorithm 4 can be decomposed to the regret of the three phases of the algorithm:

$$
\begin{aligned}
R(T) &= R_{\texttt{Phase1}}(T) + R_{\texttt{Phase2}}(T) + R_{\texttt{Phase3}}(T) \\
&\leq \mathbb{E}\left[(N+1)\sum_{i\in[K]} P_i\right] + R_{\texttt{Phase3}}(T) && \text{(expected number of rounds to obtain feedback)} \\
&= \sum_{i\in[K]} \frac{8\log(TK)}{f_i} + R_{\texttt{Phase3}}(T).
\end{aligned}
$$

We next analyze term $R_{\texttt{Phase3}}(T)$. From Lemma E.3, with probability at least $1-\delta$ the estimates $1/P_i^{LR}$ are close to f_i. But with probability at most δ, the estimates are far away and the regret that we pick up in these rounds is at most 1. Putting everything together, we have:

$$
\begin{aligned}
R_{\texttt{Phase3}}(T) &\leq 4\sqrt{2}(1-\delta)\sqrt{T\log(K)\sum_{i\in[K]} \frac{1}{f_i}} + \delta T \\
&\leq 4\sqrt{2}\sqrt{T\log(K)\sum_{i\in[K]} \frac{1}{f_i}} + 1 && \text{(Lemma E.4)}
\end{aligned}
$$

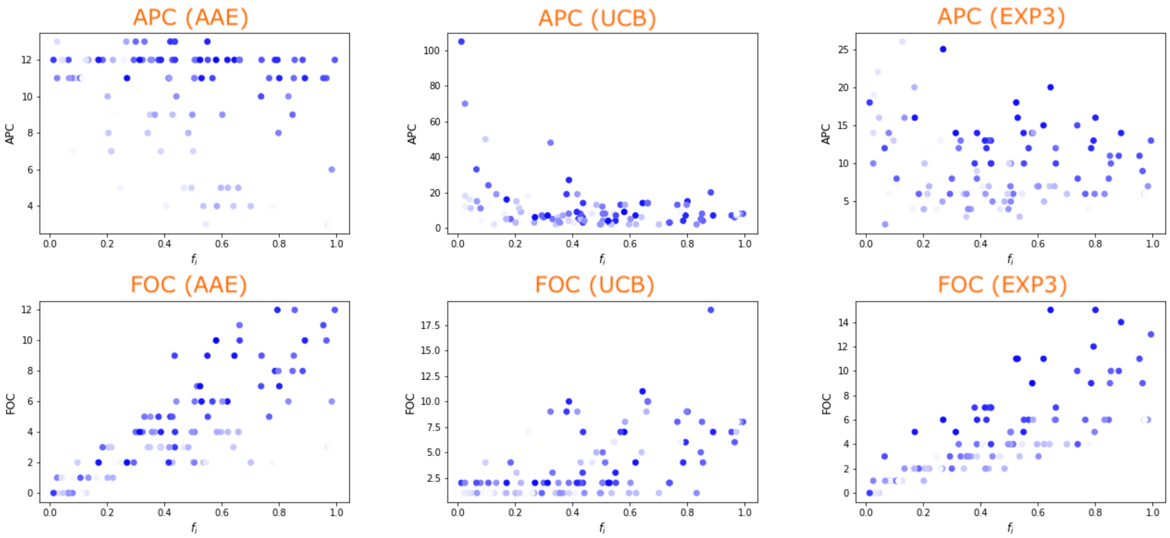

Figure 4: Correlations induced between f_i and $\texttt{APC}_i$ (top row) as well as $\texttt{FOC}_i$ (bottom row) by $\text{BB}_{\text{Pull}}(\text{AAE})$ (left column), $\text{BB}_{\text{Pull}}(\text{UCB})$ (middle), and 3-Phase EXP3 (right). There are $K = 100$ arms and $T = 1000$ rounds. The darkness of a point indicates the corresponding arm's average utility; darker is higher.

This proves a regret bound of:

$$\mathcal{O}\left(\sqrt{T\log(K)\sum_{i\in[K]}\frac{1}{f_i}} + \sum_{i\in[K]}\frac{\log(T)}{f_i} + \sum_{i\in[K]}\frac{\log(K)}{f_i}\right).$$

In general $T \geq K$, so in order to derive the bound in the theorem statement, we need to argue that $\sum_{i\in[K]}\log T/f_i$ is order smaller than $\sqrt{T\sum_{i\in[K]}1/f_i}$. Note that this is the case when $T \geq \sqrt{\sum_{i\in[K]}1/f_i}$, which is true for large enough time horizons. ■

F Beyond Monotonicity: An Empirical Study of Correlations

To better understand and control the downstream impacts of the relationship between feedback and APC/FOC, our theoretical analysis focuses on the *monotonicity* properties of these relationships, not just correlation. Monotonic dependence shifts the state of the entire system, rather than only in certain pockets of the content landscape, and so it is a stronger property to study. However, the weaker notion of *correlations* may also be of concern; here, we initiate a numerical exploration of *correlation* induced by bandit algorithms between f_i and APC/FOC.

Fig. 4 shows the correlation between either measure and the f_i's in a *single* instance. In this example, by inspection appears that $\texttt{APC}_i$ is weakly negatively correlated with f_i across algorithms, and $\texttt{FOC}_i$ is somewhat more strongly positively correlated with f_i across algorithms. Furthermore, these trends hold consistently across randomly generated instances. We give experimental details below.

Algorithms. On each instance, we run three of the algorithms from Section 4: $\text{BB}_{\text{Pull}}(\text{UCB})$ (Algorithm 8), $\text{BB}_{\text{Pull}}(\text{AAE})$ (Algorithm 7), and 3-Phase EXP3 (Algorithm 4) with the simplification that the algorithm is given the f_i's as inputs (rather than estimating them in the first two phases).

Instance generation methods. All of our instances have $K = 100$ arms and $T = 1000$ rounds. We first uniformly randomly generate the means of arms' utility / loss distributions (utilities for $\text{BB}_{\text{Pull}}(\text{UCB})$ and $\text{BB}_{\text{Pull}}(\text{AAE})$, losses for 3-Phase EXP3). These means range from 0 to 1 for $\text{BB}_{\text{Pull}}(\text{UCB})$, 0 to 5 for $BB_{Pull}(AAE)$ [17], and -1 to 0 for 3-Phase EXP3. Then, for each arm, we sample realized rewards for each time step in $[T]$ from a Gaussian distribution centered at that arm's mean with standard deviation 0.1 for $BB_{Pull}(UCB)$ and 3-Phase EXP3 and 0.5 for $BB_{Pull}(AAE)$ (commensurate with the scaled up mean). Negative utilities / positive losses are truncated to 0. Finally, we uniformly draw each arm's f_i from the interval $[0, 1]$.

Results. For illustrative purposes, the scatter plots below show the correlation between either measure and the f_i's over arms in a *single* random instance. In this single example, by inspection it looks as if $\texttt{APC}_i$ is weakly negatively correlated with f_i across algorithms, and $\texttt{FOC}_i$ is somewhat more strongly positively correlated with f_i across algorithms.

We show these trends to hold consistently across randomly generated instances: we randomly generate 100 instances by the same procedure as above, and evaluate the Pearson correlation coefficient[18] between $\texttt{APC}_i$ and f_i, $\texttt{FOC}_i$ and f_i for each of the three algorithms on every instance. In Table 2 below, we report the average Pearson correlation coefficients across these instances, which are consistent with the inspected trends in our scatter plots.

	APC_i and f_i			FOC_i and f_i		
	mean	min	max	mean	min	max
$BB_{Pull}(UCB)$	-0.33	-0.51	-0.16	0.43	0.22	0.59
3-Phase EXP3	-0.23	-0.41	0.03	0.72	0.49	0.86
$BB_{Pull}(AAE)$	-0.33	-0.53	-0.11	0.74	0.54	0.88

Table 2: Correlations between f_i and $\texttt{APC}/\texttt{FOC}$ observed in 100 randomly generated instances.

[17]The larger magnitude for AAE ensures that arms are actually eliminated in the time horizon T we have chosen to be standard across algorithms.

[18]The Pearson correlation coefficient is the ratio of two variables' covariance and the product of their standard deviations. It ranges from -1 to 1, where positive (resp. negative) values indicate positive (resp. negative) correlation, and magnitude indicates the strength of correlation.