Fair Exploration via Axiomatic Bargaining

Anonymous Author(s)
Affiliation
Address
email

Abstract

Motivated by the consideration of fairly sharing the cost of exploration between multiple groups in learning problems, we develop the Nash bargaining solution in the context of multi-armed bandits. Specifically, the 'grouped' bandit associated with any multi-armed bandit problem associates, with each time step, a single group from some finite set of groups. The utility gained by a given group under some learning policy is naturally viewed as the reduction in that group's regret relative to the regret that group would have incurred 'on its own'. We derive policies that yield the Nash bargaining solution relative to the set of incremental utilities possible under any policy. We show that on the one hand, the 'price of fairness' under such policies is limited, while on the other hand, regret optimal policies are arbitrarily unfair under generic conditions. Our theoretical development is complemented by a case study on contextual bandits for warfarin dosing where we are concerned with the cost of exploration across multiple races and age groups.

1 Introduction

Exploration in learning problems has an implicit cost, insomuch that exploring actions that are eventually revealed to be sub-optimal incurs regret. We study how this cost of exploration is shared in a system with multiple stakeholders. At the outset, we present two motivating examples.

Personalized Medicine and Adaptive Trials: Multi-stage, adaptive designs [1, 2, 3, 4], are widely viewed as a frontier in clinical trials. More generally, the ability to collect detailed patient level data, and real time monitoring (eg. glucose monitoring for diabetes [5, 6]) has raised the specter of learning personalized treatments. Among other formulations, such problems may be viewed as contextual bandits. For instance, for the problem of optimal warfarin dosing [7], the context at each time step corresponds to a patient's covariates, arms correspond to different dosages of warfarin, and the reward is the observed efficacy of the assigned dose. In examining such a study in retrospect, it is natural to measure the regret incurred by distinct groups of patients (eg. by race or age). What makes a profile of regret across such groups fair or unfair?

Revenue Management for Search Advertising: Ad platforms enjoy a tremendous amount of flexibility in the the choice of ads served against search queries. Specifically, this flexibility exists both in selecting a slate of advertisers to compete for a specific search, and then in picking a winner from this slate. Now a key goal for the platform is learning the affinity of any given ad for a given search. In solving such a learning problem – for which many variants have been proposed [8, 9] – we may again ask the question of who bears the cost of exploration, and whether the profile of such costs across various groups of advertisers is fair.

1.1 Bandits, Groups and Axiomatic Bargaining

Delaying a formal development to later, any bandit problem has an associated 'grouped' variant. Specifically, we are given a finite set of groups (eg. races or age groups in the warfarin example), and each group is associated with an arrival probability and a distribution over action sets. At each time step, a group and an action set is drawn from this distribution from which the learning algorithm must pick an action. Heterogeneity in groups is thus driven by differences in their respective distributions over feasible action sets. In addition to measuring overall regret, we also care about the regret incurred by specific groups, which we can view as the cost of exploration borne by that group.

In reasoning about 'fair' regret profiles we turn to the theory of axiomatic bargaining. There, a central decision maker is concerned with the incremental utility earned by each group from collaborating, relative to the utility the group would earn on its own. Here this incremental utility is precisely the reduction in regret for any given group relative to the optimal regret that group would have incurred 'on its own'. A *bargaining solution* maximizes some objective function over the set of achievable incremental utilities. The *utilitarian solution*, for instance, maximizes the sum of incremental utilities which would reduce here to the usual objective of minimizing total regret. The *Nash bargaining solution* maximizes an alternative objective, the Nash Social Welfare (SW) function. This latter solution is the unique solution to satisfy a set of axioms any 'fair' solution would reasonably satisfy. *This paper develops the Nash bargaining solution to the (grouped) bandit problem.*

1.2 Contributions

In developing the Nash bargaining solution, we focus primarily on what is arguably the simplest non-trivial grouped bandit setting. Specifically, we consider the 'grouped' K-armed bandit model, wherein each group corresponds to a subset of the K arms. We make the following contributions relative to this problem:

Regret Optimal Policies are Unfair (Theorem 3.1): We show that all regret optimal policies for the grouped K-armed bandit share a structural property that make them 'arbitrarily unfair' – in the sense that the Nash SW is $-\infty$ for these solutions – under a broad set of conditions on the problem instance.
Achievable Fairness (Theorem 3.2): We derive an instance-dependent upper bound on the Nash SW for the grouped K-armed bandit. This can be viewed as a 'fair' analogue to a regret lower bound (e.g. [10]) for the problem, since a lower bound on achievable regret (forgoing any fairness concerns) would in effect correspond to an upper bound on the utilitarian SW for the problem.
Nash Solution (Theorem 4.1): We produce a policy that achieves the Nash solution. Specifically, the Nash SW under this policy achieves the upper bound we derive on the Nash SW for all instances of the grouped K-armed bandit.
Price of Fairness for the Nash Solution (Theorem 4.2): We show that the 'price of fairness' for the Nash solution is small: the Nash solution achieves at least $O(1/\sqrt{G})$ of the reduction in regret achieved under a regret optimal solution relative to the regret incurred when groups operate separately.

Taken together, these results establish a rigorous framework for the design of bandit algorithms that yield fair outcomes across groups at a low cost to total regret. As a final contribution, we extend our framework beyond the grouped K-armed bandit and undertake an empirical study:

Linear Contextual Bandits and Warfarin Dosing: We extend our framework to grouped linear contextual bandits, yielding a candidate Nash solution there. Applied to a real-world dataset on warfarin dosing using race and age groups, we show (a) a regret optimal solution that ignores groups is dramatically unfair, and (b) the Nash solution balances out reductions in regret across groups at the cost of a small increase in total regret.

1.3 Related Literature

Two pieces of prior work have a motivation similar to our own. [11] studies a setting with multiple agents with a common bandit problem, where each agent can decide which action to take at each time. They show that 'free-riding' is possible — an agent that can access information from other agents can incur only $O(1)$ regret in several classes of problems. This is consistent with our motivation. [12] studies a very similar grouped bandit model to ours, and provides a 'counterexample' in which a group can have a negative externality on another group. This example is somewhat pathological and stems from considering an instance-specific fixed time horizon; instead, if $T \to \infty$, all externalities become non-negative (details in Appendix A.1). Our grouped bandit model is also similar to *sleeping bandits* [13], in which the set of available arms is adversarially chosen in each round. The known, fixed group structure in our model allows us to achieve tighter regret bounds than [13].

There have also been a handful of papers [14, 15, 16, 17] that study 'fairness in bandits' in a completely different context. These works enforce a fairness criterion between *arms*, which is relevant in settings where a 'pull' represents some resource that is allocated to that arm, and these pulls should be distributed between arms in a fair manner. In these models, the decision maker's objective (maximize reward) is distinct from that of a group (obtain 'pulls'), unlike our setting (and motivating examples) where the groups and decision maker are aligned in their eventual objective.

Our upper bound on Nash SW borrows classic techniques from the regret lower bound results of [10] and [18]. Our policy follows a similar pattern to recent work on regret-optimal, optimization-based policies for structured bandits [19, 20, 21, 22]. Unlike those policies, our policy has no forced exploration. Further the optimization problem defining the Nash solution can generically have multiple solutions whereas the aforementioned approaches would require this solution to be unique; our approach does not require a unique solution. Nonetheless, we believe that the framework in the aforementioned works can be fruitfully leveraged to construct Nash solutions for general grouped bandits, and we provide such a candidate solution as an extension.

Our fairness framework is inspired by the literature on fairness in welfare economics — see [23, 24]. Specifically, we study fairness in exploration through the lens of the axiomatic bargaining framework, first studied by [25], who showed that enforcing four desirable axioms induces a unique fair solution. [26] is an excellent textbook reference for this topic.

2 The Axiomatic Bargaining Framework for Bandits

Let $\theta \in \Theta$ be an unknown parameter and let $\mathcal{A}$ be the action set. For every arm $a \in \mathcal{A}$, $(Y_n(a))_{n \geq 1}$ is an i.i.d. sequence of rewards drawn from a distribution $F(\theta, a)$ parameterized by θ and a. We let $\mu(a) = \mathbb{E}[Y_1(a)]$ be the expected reward of arm a. In defining a *grouped* bandit problem, we let $\mathcal{G}$ be a set of G groups. Each group $g \in \mathcal{G}$ is associated with a probability distribution P^g over $2^{\mathcal{A}}$, and a probability of arrival p_g; $\sum_g p_g = 1$. The identity of the group arriving at time t, g_t, is chosen independently according to this latter distribution; $\mathcal{A}_t$ is then drawn according to P^{g_t}. An instance of the grouped bandit problem is specified by $\mathcal{I} = (\mathcal{A}, \mathcal{G}, p, P, F, \theta)$, where all quantities except for θ are known. At each time t, a central decision maker observes g_t and $\mathcal{A}_t$, chooses an arm $A_t \in \mathcal{A}_t$ to pull and observes the reward $Y_{N_t(A_t)+1}(A_t)$, where $N_t(a)$ is the total number of times arm a was pulled up to but not including time t. Let $A_t^* \in \operatorname{argmax}_{a \in \mathcal{A}_t} \mu(a)$ be an optimal arm at time t. Given an instance $\mathcal{I}$ and a policy π, the *total regret*, and the *group regret* for group $g \in \mathcal{G}$ are respectively

$$R_T(\pi, \mathcal{I}) = \mathbb{E}\left[\sum_{t=1}^{T} (\mu(A_t^*) - \mu(A_t))\right] \text{ and } R_T^g(\pi, \mathcal{I}) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(g_t = g)(\mu(A_t^*) - \mu(A_t))\right],$$

where the expectation is over randomness in arrivals $(g_t, \mathcal{A}_t)$, rewards $Y_n(a)$, and the policy π. Finally, so that the notion of an optimal policy for some class of instances, $\boldsymbol{\mathcal{I}}$, is well defined, we restrict attention to *consistent* policies which yield sub-polynomial regret for any instance in that class: $\Psi = \{\pi : R_T(\pi, \mathcal{I}) = o(T^b) \ \forall \mathcal{I} \in \boldsymbol{\mathcal{I}}, \forall b > 0\}$.

2.1 Background: Axiomatic Bargaining

The axiomatic bargaining problem is specified by the number of agents n, a set of feasible utility profiles $U \subseteq \mathbb{R}^n$, and a disagreement point $d \in \mathbb{R}^n$, that represents the utility profile when agents cannot come to an agreement. A solution $f(\cdot, \cdot)$ to the bargaining problem selects an agreement $u^* = f(U, d) \in U$, in which agent i receives utility u_i^*. It is assumed that there is at least one point $u \in U$ such that $u > d$, and we assume U is compact and convex.

The bargaining framework proposes a set of axioms a fair solution u^* should ideally satisfy:
(a) Pareto Optimality: There is no $u \in U$ with $u \geq u^*, u \neq u^*$.
(b) Invariance to Affine Transformations: If $U' = \{a^\top u + b : u \in U\}$ and $d' = a^\top d + b$, then $f(U', d')_i = a_i u_i^* + b_i$ for any $a \in \mathbb{R}_+^n, b \in \mathbb{R}^n$.
(c) Independence of Irrelevant Alternatives: If $V \subseteq U$ where $u^* \in V$, then $f(V, d) = u^*$.
(d) Symmetry: If U and d are symmetric, $u_i^* = u_j^* \ \forall i, j$.
Now (b) implies that $f(U, d) = f(\{u - d : u \in U\}, 0) + d$. It is therefore customary to normalize the origin to the disagreement point, i.e. assume $d = 0$, and implicitly that U has been appropriately translated. So translated, U is interpreted as a set of feasible utility *gains* relative to the disagreement

point. The seminal work of [25] showed that there is a unique bargaining solution that satisfies the above four axioms, and it is the outcome that maximizes the *Nash social welfare (SW) function* [27]:

$$W(u) = \begin{cases} \sum_{i=1}^{n} \log(u_i) & u_i > 0 \; \forall i \in [n] \\ -\infty & \text{otherwise.} \end{cases}$$

We will interchangeably refer to $u^* = \operatorname{argmax}_{u \in U} W(u)$ as the *Nash solution* or as *proportionally fair*. If $u \in U$ such that $W(u) = -\infty$, we say that u is *unfair*.

2.2 Fairness Framework for Grouped Bandits

We now consider the Nash bargaining solution in the context of the grouped bandit problem. To do so, we need to appropriately define the utility gain under any policy. We begin by formalizing the rewards to a single group under a policy where no information was shared across groups, which represents the disagreement point. Specifically, let $\mathcal{I}_g$ be the 'single-group' bandit instance obtained by considering the instance $\mathcal{I}$ restricted to arrivals of group g so that in any period t in which $g_t \neq g$, we receive no reward under any action. Let us denote by π_g^* an optimal policy for instances of type $\mathcal{I}_g$ (i.e. π_g^* is optimal in the non-grouped bandit setting) so that for any instance of type $\mathcal{I}_g$, and any other consistent policy π_g' for instances of that type,

$$\limsup_{T \to \infty} \frac{R_T(\pi_g^*, \mathcal{I}_g)}{\log T} \leq \liminf_{T \to \infty} \frac{R_T(\pi_g', \mathcal{I}_g)}{\log T}. \tag{1}$$

Now letting $\tilde{R}_T^g(\mathcal{I}) \triangleq R_T(\pi_g^*, \mathcal{I}_g)$, we define, with a slight abuse of notation, the T-period utility earned by group g under π_g^*, and any other consistent policy π for instances of type $\mathcal{I}$ respectively, as:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(g_t = g)\mu(A_t^*)\right] - \tilde{R}_T^g(\mathcal{I}) \triangleq u_T^g(\pi_g^*) \text{ and } \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(g_t = g)\mu(A_t^*)\right] - R_T^g(\pi, \mathcal{I}) \triangleq u_T^g(\pi).$$

The T-period utility gain under a policy π is then $u_T^g(\pi) - u_T^g(\pi_g^*) = \tilde{R}_T^g(\mathcal{I}) - R_T^g(\pi, \mathcal{I})$. Since our goal is to understand long-run system behavior, we define asymptotic utility gain for any group g:

$$\text{UtilGain}^g(\pi, \mathcal{I}) = \liminf_{T \to \infty} \frac{\tilde{R}_T^g(\mathcal{I}) - R_T^g(\pi, \mathcal{I})}{\log T}.$$

Equipped with this definition, we may now identify the set of incremental utilities for an instance $\mathcal{I}$, as $U(\mathcal{I}) = \{(\text{UtilGain}^g(\pi, \mathcal{I}))_{g \in \mathcal{G}} : \pi \in \Psi\}$. We can readily show that the Nash solution remains the unique solution satisfying the fairness axioms presented in Section 2.1 relative to $U(\mathcal{I})$. We finish up by finally defining the Nash solution to the grouped bandit problem. Since we find it convenient to associate a SW function with a policy (as opposed to a vector of incremental utilities), the Nash SW function for grouped bandits is equivalently defined as:

$$W(\pi, \mathcal{I}) = \begin{cases} \sum_{g \in \mathcal{G}} \log\left(\text{UtilGain}^g(\pi, \mathcal{I})\right) & \text{UtilGain}^g(\pi, \mathcal{I}) > 0 \; \forall g \in \mathcal{G} \\ -\infty & \text{otherwise.} \end{cases} \tag{2}$$

So equipped, we finish by defining the Nash solution to the grouped bandit problem.

Definition 2.1. Suppose a policy π^* satisfies $W(\pi^*, \mathcal{I}) = \sup_{\pi \in \Psi} W(\pi, \mathcal{I})$ for every instance $\mathcal{I} \in \boldsymbol{\mathcal{I}}$. Then, we say that π^* is the Nash solution for $\boldsymbol{\mathcal{I}}$ and that it is *proportionally fair*.

2.3 Grouped K-armed Bandit Model

The grouped K-armed bandit is arguably the simplest non-trivial class of grouped bandits. Let $\mathcal{A} = [K]$. Denote by $\mathcal{A}^g \subseteq \mathcal{A}$ a subset of arms corresponding to group g and by $\mathcal{G}_a$ a subset of groups corresponding to arm a. For each g, P^g places unit mass on $\mathcal{A}^g$ so that the set of arms available at time t is $\mathcal{A}_t = \mathcal{A}^{g_t}$. Assume $\theta \in (0,1)^K$, and the single period reward $Y_1(a) \sim \text{Bernoulli}(\theta(a))$. We assume that $\theta(a) \neq \theta(a')$ for all $a \neq a'$. Since the set of arms available at each time step only depends on the arriving group, we denote by $\mathsf{OPT}(g) = \max_{a \in \mathcal{A}^g} \theta(a)$ the optimal mean reward for group g. We take π_g^* to be the KL-UCB policy of [28] since KL-UCB is optimal (in the sense of (1)) for vanilla K-armed bandits. We may write the T-period regret in this model as

$$R_T(\pi, \mathcal{I}) = \sum_{g \in G} \sum_{a \in \mathcal{A}^g} \Delta^g(a) \mathbb{E}[N_T^g(a)], \tag{3}$$

where $N_T^g(a)$ is the number of times that group g has pulled arm a after T time steps, and $\Delta^g(a) = \mathsf{OPT}(g) - \theta(a)$. Lastly, we state a condition guaranteeing $U(\mathcal{I})$ contains a point $u > 0$; Proposition G.1 in Appendix G proves the following assumption is necessary and sufficient:

Assumption 2.2. *Every group g has at least one suboptimal arm that is shared with another group. That is, for every g, $\exists a \in \mathcal{A}^g$ such that $\mu(a) < \mathsf{OPT}(g)$ and $|\mathcal{G}_a| \geq 2$.*

3 Fairness-Regret Trade-off

In this section, we prove that a regret-optimal policy for a generic grouped K-armed bandit must necessarily be unfair. We then turn to deriving an upper bound on achievable Nash SW.

3.1 Unfairness of Regret Optimal Policies

We first state the main result, which states that regret optimal policies are arbitrarily unfair. In fact, we show that perversely the most 'disadvantaged' group (in a sense we make precise shortly) bears the brunt of exploration in that it sees no improvement in regret relative to if it were 'on its own'.

Theorem 3.1. *Let π be a regret optimal policy. Let $\mathcal{I}$ be an instance of the grouped K-armed bandit where $g_{\min} \triangleq \operatorname{argmin}_{g \in G} \mathsf{OPT}(g)$ is unique. Then, $W_{\mathcal{I}}(\pi) = -\infty$ and $\mathrm{UtilGain}^{g_{\min}}(\pi, \mathcal{I}) = 0$.*

Proof. We define regret optimality by proving tight lower and upper bounds on regret, and these bounds imply necessary properties of all regret optimal policies that yield the desired result.

We first lower bound the total number of pulls, $\mathbb{E}[N_T(a)]$, of a suboptimal arm. Denote by $\mathcal{A}_{\text{sub}}^g = \{a \in \mathcal{A}^g : \theta(a) < \mathsf{OPT}(g)\}$ the suboptimal actions for group g, and denote by $\mathcal{A}_{\text{sub}} = \{a \in \mathcal{A} : a \in \mathcal{A}_{\text{sub}}^g \ \forall g \in \mathcal{G}_a\}$ the set of arms that are not optimal for any group. Now since a consistent policy for the grouped K-armed bandit is automatically consistent for the vanilla K-armed bandit obtained by restricting to any of its component groups g, the standard lower bound of [10] implies that for any $a \in \mathcal{A}_{\text{sub}}^g$, $\liminf_{T\to\infty} \mathbb{E}[N_T(a)]/\log T(g) \geq J^g(a)$ where $J^g(a) \triangleq 1/\mathrm{KL}(\theta(a), \mathsf{OPT}(g))$ and $T(g)$ is the number of arrivals of group g up to and including time T. Since this must hold for any group, and since $\lim_T \log T / \log T(g) = 1$ a.s.,

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_T(a)]}{\log T} \geq J(a) \tag{4}$$

for all $a \in \mathcal{A}_{\text{sub}}$ where $J(a) = \max_{g \in \mathcal{G}_a} J^g(a)$. Now, denote by $\Gamma(a) = \operatorname{argmin}_{g \in \mathcal{G}_a} \mathsf{OPT}(g)$ the set of groups that have the smallest optimal reward out of all groups that have access to a. Then the smallest regret incurred in pulling arm a is simply $\Delta^g(a)$ for any $g \in \Gamma(a)$. With a slight abuse, we denote this quantity by $\Delta^{\Gamma(a)}(a)$. (4) immediately implies that for any consistent policy π,

$$\liminf_{T\to\infty} \frac{R_T(\pi, \mathcal{I})}{\log T} \geq \sum_{a \in \mathcal{A}_{\text{sub}}} \Delta^{\Gamma(a)}(a) J(a). \tag{5}$$

In fact, we show that the KL-UCB policy [28] (surprisingly) achieves this lower bound; the proof of this claim is somewhat involved and can be found in Appendix C. Consequently, any regret optimal policy must achieve the limit infimum in (5). In turn, this implies that a policy $\pi \in \Psi$ is regret optimal if and only if, the number of pulls of arms $a \in \mathcal{A}_{\text{sub}}$ achieve the lower bound (4), i.e.

$$\lim_{T\to\infty} \frac{\mathbb{E}[N_T(a)]}{\log T} = J(a) \quad \forall a \in \mathcal{A}_{\text{sub}} \tag{6}$$

and further that any pulls of arm a from a group $g \notin \Gamma(a)$ must be negligible, i.e.

$$\lim_{T\to\infty} \frac{\mathbb{E}[N_T^g(a)]}{\log T} = 0 \quad \forall a \in \mathcal{A}, g \notin \Gamma(a). \tag{7}$$

Now, turning our attention to $g_{\min}$, we have by assumption that $g_{\min}$ is the only group in $\Gamma(a)$ for all $a \in \mathcal{A}^{g_{\min}}$. Consequently, by (7), we must have that for any optimal policy, $\lim_{T\to\infty} \mathbb{E}[N_T^{g_{\min}}(a)]/\log T = \lim_{T\to\infty} \mathbb{E}[N_T(a)]/\log T$ for all $a \in \mathcal{A}^{g_{\min}}$. And since $J(a) = J^{g_{\min}}(a)$ for all $a \in \mathcal{A}^{g_{\min}} \cap \mathcal{A}_{\text{sub}}$, (6) then implies that the regret for group $g_{\min}$ is precisely

$$\lim_{T\to\infty} \frac{R_T^{g_{\min}}(\pi, \mathcal{I})}{\log T} = \sum_{a \in \mathcal{A}_{\text{sub}}^{g_{\min}}} \Delta^{g_{\min}}(a) J^{g_{\min}}(a).$$

But this is precisely $\lim_T \tilde{R}_T^{g_{\min}}(\mathcal{I})/\log T$. Thus, $\text{UtilGain}^{g_{\min}}(\pi, \mathcal{I}) = 0$, and $W_{\mathcal{I}}(\pi) = -\infty$. $\square$

The proof also illustrates that if $g_{\max} \triangleq \text{argmax}_{g \in G} \mathsf{OPT}(g)$ is unique, then $g_{\max}$ incurs no regret from *any* shared arm in a regret optimal policy. If all suboptimal arms for $g_{\max}$ are shared with another group, then $g_{\max}$ incurs zero (log-scaled) regret in an optimal policy. In summary, regret optimal policies are unfair, and achieve perverse outcomes with the most disadvantaged groups gaining nothing and the most advantaged groups gaining the most from sharing the burden of exploration.

3.2 Upper Bound on Nash SW

The preceding question motivates asking what is in fact possible with respect to fair outcomes. To that end, we derive an instance-dependent upper bound on the Nash SW. We may view this as a 'fair' analogue to instance-dependent lower bounds on regret.

Recall the definition of $W(\pi, \mathcal{I})$ in (2), and let $W^*(\mathcal{I}) = \sup_{\pi \in \Psi} W(\pi, \mathcal{I})$. Fix an instance $\mathcal{I}$ with unknown parameter vector θ. We first upper bound $W(\pi, \mathcal{I})$. Recall that KL-UCB is the policy π_g^* used to define $\tilde{R}_T^g(\mathcal{I})$. The fact that KL-UCB is optimal in the vanilla K-armed bandit implies:

$$\lim_{T \to \infty} \frac{\tilde{R}_T^g(\mathcal{I})}{\log T} = \sum_{a \in \mathcal{A}_{\text{sub}}^g} \Delta^g(a) J^g(a). \tag{8}$$

Next, we re-write $R_T^g(\pi, \mathcal{I})/\log T$. Given a policy π, for any action a and group g, let $q_T^g(a, \pi) \in [0, 1]$ be the *percentage* of times that group g pulls arm a, out of the total number of times arm a is pulled. That is, $\mathbb{E}[N_T^g(a)] = q_T^g(a, \pi)\mathbb{E}[N_T(a)]$, where $\sum_{g \in G} q_T^g(a, \pi) = 1$ for all a. Then,

$$\frac{R_T^g(\pi, \mathcal{I})}{\log T} = \sum_{a \in \mathcal{A}_{\text{sub}}^g} \Delta^g(a) q_T^g(a, \pi) \frac{\mathbb{E}[N_T(a)]}{\log T} \geq \sum_{a \in \mathcal{A}_{\text{sub}}^g \cap \mathcal{A}_{\text{sub}}} \Delta^g(a) q_T^g(a, \pi) \frac{\mathbb{E}[N_T(a)]}{\log T}. \tag{9}$$

Recalling $\text{UtilGain}^g(\pi, \mathcal{I}) = \liminf_{T \to \infty} \frac{\tilde{R}_T^g(\mathcal{I}) - R_T^g(\pi, \mathcal{I})}{\log T}$, combining (8), (9), and (4) yields:

$$\text{UtilGain}^g(\pi, \mathcal{I}) \leq \liminf_{T \to \infty} \sum_{a \in \mathcal{A}_{\text{sub}}^g} \Delta^g(a) \left(J^g(a) - q_T^g(a, \pi) J(a) \mathbf{1}\{a \in \mathcal{A}_{\text{sub}}\} \right).$$

Using the definition of $W(\pi, \mathcal{I})$ and taking the $\liminf$ outside of the sum gives

$$W(\pi, \mathcal{I}) \leq \liminf_{T \to \infty} \sum_{g \in \mathcal{G}} \log \left(\sum_{a \in \mathcal{A}_{\text{sub}}^g} \Delta^g(a) \left(J^g(a) - q_T^g(a, \pi) J(a) \mathbf{1}\{a \in \mathcal{A}_{\text{sub}}\} \right) \right)^+.$$

But since $\sum_{g \in \mathcal{G}} q_T^g(a, \pi) = 1$ for every T, a, it must be that the limit infimum above is achieved for some vector $(q^g(a))$ satisfying $\sum_{g \in G} q^g(a) = 1$ for all a. This immediately yields an upper bound on $W^*(\mathcal{I})$: Let $Y^*(\mathcal{I})$ be the optimal value to the program $P(\theta)$, and let q_* be an optimal solution.

$$(P(\theta)) \qquad \begin{aligned} \max_{q \geq 0} \quad & \sum_{g \in \mathcal{G}} \log \left(\sum_{a \in \mathcal{A}_{\text{sub}}^g} \Delta^g(a) \left(J^g(a) - q^g(a) J(a) \right) \right)^+ \\ \text{s.t.} \quad & \sum_{g \in \mathcal{G}} q^g(a) = 1 \quad \forall a \in \mathcal{A}_{\text{sub}} \\ & q^g(a) = 0 \quad \forall g \in G, a \notin \mathcal{A}_{\text{sub}} \cap \mathcal{A}_g. \end{aligned}$$

Then, we have shown:

Theorem 3.2. *For every instance $\mathcal{I}$ of the grouped K-armed bandit, $W^*(\mathcal{I}) \leq Y^*(\mathcal{I})$.*

4 Nash Solution for Grouped K-armed Bandits

We turn our attention in this section to constructive issues: we first develop an algorithm that achieves the Nash SW upper bound of Theorem 3.2 and thus establish that this is the Nash solution for the grouped K-armed bandit. In analogy to the unfairness of a regret optimal policy, it is then natural to ask whether the regret under this Nash solution is large relative to optimal regret; we show thankfully that this 'price of fairness' is relatively small.

4.1 The Nash Solution: PF-UCB

The algorithm we present here 'Proportionally Fair' UCB (or PF-UCB) works as follows: at each time step it computes the set of arms that optimize the (KL) UCB for some group. Then, when a group arrives, it asks whether any arm from this set has been 'under-explored' where the notion of under-exploration is measured relative to an estimated optimal solution to $P(\theta)$. Such an arm, if available, is pulled. Absent the availability of such an arm, a greedy selection is made.

Specifically, let $\hat{\theta}_t$ be the empirical mean estimate of θ at time t. $P(\hat{\theta}_t)$ is then our approximation to $P(\theta)$ at time t and we denote by $\hat{q}_t$ the optimal solution to this program with smallest euclidean norm. Note that finding such a solution constitutes a tractable convex optimization problem. We define the standard KL-UCB for an arm, $\mathrm{UCB}_t(a) = \max\{q : N_t(a)\mathrm{KL}(\hat{\theta}_t(a), q) \leq \log t + 3 \log\log t\}$. Finally, we denote by $A_t^{\mathrm{UCB}}(g) \in \mathrm{argmax}_{a\in\mathcal{A}^g} \mathrm{UCB}_t(a)$ the arm with the highest UCB for group g at time t, and by $\mathcal{A}_t^{\mathrm{UCB}} = \{A_t^{\mathrm{UCB}}(g) : g \in \mathcal{G}\}$ the set of arms that have the highest UCB for *some* group. PF-UCB then proceeds as follows. At time t:

1. If there is an available arm $a \in \mathcal{A}^{g_t} \cap \mathcal{A}_t^{\mathrm{UCB}}$ such that $N_t^{g_t}(a) \leq \hat{q}_t^g(a) N_t(a)$, pull a. If there are multiple arms matching this criteria, pull one of them uniformly at random.
2. Otherwise, pull the greedy arm $A_t^{\text{greedy}}(g_t) \in \mathrm{argmax}_{a\in\mathcal{A}^{g_t}} \hat{\theta}_t(a)$.

PF-UCB explores at time t by pulling an arm if it is the arm with the highest UCB for *some* group (not necessarily group g_t), *and* the current group g_t has not pulled it as many times as it should have according to the solution $\hat{q}_t$. PF-UCB constitutes a Nash solution for the grouped K-armed bandit. Specifically, we prove the following theorem in Appendix E:

Theorem 4.1. *For any instance $\mathcal{I}$ of the grouped K-armed bandit, we have for all groups g,*

$$\lim_{T\to\infty} \frac{R_T^g(\pi^{\text{PF-UCB}}, \mathcal{I})}{\log T} = \sum_{a\in\mathcal{A}^g} \Delta^g(a) q_*^g(a) J(a).$$

It is worth noting that relative to the existing optimization-based algorithms for structured bandits (e.g. [19, 20, 21, 22]), PF-UCB does no forced sampling. In addition, we make no requirement that the solution to the optimization problem $P(\theta)$ is unique as these existing policies require. In fact, optimal solutions to $P(\theta)$ are not unique, and the choice of a solution that has smallest euclidean norm is carefully shown to provide the necessary 'stability' while being computationally tractable. That said, the next section shows how we can fruitfully leverage an existing algorithm from [22] to construct a candidate Nash solution for a setting beyond the grouped K-armed bandit.

4.2 Price of Fairness

Whereas PF-UCB is proportionally fair, what price do we pay with respect to regret? To answer this question we compute in this section an upper bound on the 'price of fairness'. Specifically, define

$$\text{SYSTEM}(\mathcal{I}) = \textstyle\sum_{g\in\mathcal{G}} \mathrm{UtilGain}^g(\pi^{\text{KL-UCB}}, \mathcal{I}) \text{ and } \text{FAIR}(\mathcal{I}) = \textstyle\sum_{g\in\mathcal{G}} \mathrm{UtilGain}^g(\pi^{\text{PF-UCB}}, \mathcal{I}).$$

$\mathrm{UtilGain}^g(\pi^{\text{KL-UCB}}, \mathcal{I})$ is the reduction in group g's regret under a *regret optimal* policy in the grouped setting relative to the optimal regret it would have endured on its own; $\text{SYSTEM}(\mathcal{I})$ aggregates this reduction in regret across all groups. Similarly, $\mathrm{UtilGain}^g(\pi^{\text{PF-UCB}}, \mathcal{I})$ is the reduction in group g's regret under a *proportionally fair* policy, and $\text{FAIR}(\mathcal{I})$ aggregates this across groups. The price of fairness (PoF) asks what fraction of the optimal reduction in regret is lost to fairness:

$$\mathrm{PoF}(\mathcal{I}) \triangleq \frac{\text{SYSTEM}(\mathcal{I}) - \text{FAIR}(\mathcal{I})}{\text{SYSTEM}(\mathcal{I})}.$$

Of course, $\mathrm{PoF}(\mathcal{I})$ is a quantity between 0 and 1, where smaller values are preferable.

Now for an instance $\mathcal{I}$, let $s^g(\mathcal{I}) = \sup_{\pi\in\Psi^+(\mathcal{I})} \mathrm{UtilGain}^g(\pi, \mathcal{I})$ be the maximum achievable utility gain (or equivalent, the largest reduction in regret possible) for group g, where $\Psi^+(\mathcal{I}) = \{\pi \in \Psi : \mathrm{UtilGain}^g(\pi, \mathcal{I}) \geq 0 \ \forall g \in \mathcal{G}\}$. Then, $R(\mathcal{I}) = \min_{g\in\mathcal{G}} s^g(\mathcal{I}) / \max_{g\in\mathcal{G}} s^g(\mathcal{I})$ is a measure of the inherent asymmetry of the instance $\mathcal{I}$ with respect to utility gain across groups. We show:

Theorem 4.2. *For an instance $\mathcal{I}$ of the grouped K-armed bandit,* $\mathrm{PoF}(\mathcal{I}) \leq 1 - R(\mathcal{I}) \frac{2\sqrt{G}-1}{G}$.

The proof relies on an analysis of the price of fairness for general convex allocation problems in [29] and may be found in Appendix F. The key takeaway from this result is that, treating the inherent asymmetry $R(\mathcal{I})$ as a constant, the price of fairness grows *sub-linearly* in the number of groups G. It is unclear we can expect this with other fairness solution concepts: for instance, we would expect the price of fairness under a max-min solution to grown linearly with the number of groups [29]. Further, whereas the bound above depends on the topology of the instance only through $R(\mathcal{I})$, a topology specific analysis may well yield stronger results. For instance:

Proposition 4.3. *Let $\mathcal{I}$ be an instance such that for every arm $a \in \mathcal{A}$, either $\mathcal{G}_a = \mathcal{G}$ or $|\mathcal{G}_a| = 1$. Then* $\mathrm{PoF}(\mathcal{I}) \leq \frac{1}{2}$.

This result shows that for a specific class of topologies, the price of fairness is a constant independent of any parameters including the number of groups or the mean rewards. In Section 6 we study the price of fairness computationally in the context of random families of instances.

5 Extension to Grouped Contextual Linear Bandits

In this section, we introduce the grouped linear contextual bandit model and propose a candidate Nash solution by extending the regret optimal policy of [22] (without theory). We apply this model and the policies in Section 6 for an empirical case study.

Grouped Linear Contextual Bandit Model: Let $\theta \in \mathbb{R}^d$ and $\mathcal{A} \subseteq \mathbb{R}^d$. The reward for pulling arm a for the n'th time is $Y_n(a) = \langle a, \theta \rangle + \varepsilon_{a,n}$, where $\varepsilon_{a,n}$ is distributed i.i.d. $N(0,1)$. Let $\mathcal{M} \subseteq \mathbb{R}^d$ be the set of contexts, where $|\mathcal{M}| = M < \infty$, and each $m \in \mathcal{M}$ is associated with an action set $\mathcal{A}(m) \subseteq \mathcal{A}$. Each group $g \in \mathcal{G}$ has a probability of arrival, p^g, and a distribution P^g over contexts $[M]$. At each time t, a group g_t is drawn independently from $(p^g)_g$, then a random context $m_t \sim P^{g_t}$ is drawn. The action set at time t is $\mathcal{A}_t = \mathcal{A}(m_t)$. Let $\mathcal{M}^g$ be the contexts in the support of P^g. Let $\mathsf{OPT}(m) = \max_{a \in \mathcal{A}(m)} \langle a, \theta \rangle$ and $\Delta(m,a) = \mathsf{OPT}(m) - \langle a, \theta \rangle$.

Regret Optimal Policy: [22] provides an instance-dependent lower bound for linear contextual bandits as the optimal value of the following optimization problem:

$$
(L(\theta)) \qquad
\begin{aligned}
Y(\mathcal{M}) = \min_{Q \geq 0} \quad & \textstyle\sum_{m \in \mathcal{M}} \sum_{a \in \mathcal{A}(m)} Q(m,a) \Delta(m,a) \\
\text{s.t.} \quad & Q(a) = \textstyle\sum_{m : a \in \mathcal{A}(m)} Q(m,a) \quad \forall a \in \mathcal{A} \\
& (Q(a))_{a \in \mathcal{A}} \in \mathcal{Q},
\end{aligned}
$$

where $\mathcal{Q}$ is the following polytope ensuring the consistency of the policy:

$$
\mathcal{Q} = \left\{ (Q(a))_{a \in \mathcal{A}} : ||a||^2_{H_Q^{-1}} \leq \Delta(m,a)^2/2 \; \forall m \in [M], a \in \mathcal{A}(m), H_Q = \textstyle\sum_{a \in \mathcal{A}} Q(a) a a^\top \right\}.
$$

The variable $Q(m,a)$ represents how often context m pulls arm a. [22] provides a policy (OAM) whose regret matches this lower bound. At a high level, like PF-UCB, OAM solves $L(\hat{\theta}_t)$ at each time step and 'follows' the solution; but it does not make use of a UCB and rather uses forced exploration. There are many details in the OAM policy and the full description can be found in Appendix A.2.

Candidate Nash Solution: We propose a policy which runs exactly OAM, except that the optimization problem solved at every time step is changed to the following:

$$
(L^{\text{fair}}(\theta)) \qquad
\begin{aligned}
\max_{Q \geq 0} \quad & \textstyle\sum_{g \in \mathcal{G}} \log \Big(Y(\mathcal{M}^g) - \sum_{m \in \mathcal{M}^g} \sum_{a \in \mathcal{A}(m)} Q^g(m,a) \Delta(m,a) \Big)^+ \\
\text{s.t.} \quad & Q(a) = \textstyle\sum_{g \in \mathcal{G}} \sum_{m \in \mathcal{M}^g : a \in \mathcal{A}(m)} Q^g(m,a) \quad \forall a \in \mathcal{A} \\
& (Q(a))_{a \in \mathcal{A}} \in \mathcal{Q}.
\end{aligned}
$$

Compared to $(L(\theta))$, the objective is modified to maximize the Nash SW, and the new variable $Q^g(m,a)$ represents how often group g with context m should pull arm a.

We do not have a theoretical guarantee that this extension of OAM is indeed the Nash solution. This is not implied by [22] since there is an added group structure on the bandit model and OAM requires that the optimization problem has a unique solution, which $(L^{\text{fair}}(\theta))$ does not. Proving such a guarantee is a natural direction for future work.

6 Experiments

We consider two sets of experiments. The first seeks to understand the PoF in synthetic instances to shed further light on the impact of topology. The second is a real-world case study that returns to the Warfarin dosing example discussed in motivating the paper where we seek to understand unfairness under a regret optimal policy and the extent to which the Nash solution can mitigate this problem.

Synthetic Grouped K-Armed Bandits: We consider two generative models that differ in how the bipartite graph matching groups to available arms is generated. In 'i.i.d.', each edge appears independently with probability 0.5, and $K = 10$ is fixed. The mean reward of each arm is i.i.d. $\mathrm{U}(0,1)$. In 'Skewed', $K = G + 1$, and a group $g \in \{1, \ldots, G-1\}$ has access to arms $\{g, G\}$, while the last group $g = G$ has access to all arms. The rewards of arms $1, \ldots, G-1$ are equal, and $\mu(1) < \mu(G) < \mu(G+1)$ are generated randomly by sorting three i.i.d. $\mathrm{U}(0,1)$ random variables.

Table 1 shows that the PoF is very small in the 'i.i.d.' setting, and contrary to Theorem 4.2 the PoF actually decreases as G gets large. This suggests an interesting conjecture for future research: the PoF may actually grow negligible in large random bandit instances. The 'Skewed' structure is motivated by our PoF analysis where we see that the PoF increase – albeit slowly – with G.

Table 1: The median and 95th percentile of the PoF for synthetic instances of the grouped K-armed bandit over 500 runs of each method.

	i.i.d.				Skewed			
G	3	5	10	50	3	5	10	50
Median	0.073	0.054	0.040	0.015	0.327	0.407	0.454	0.521
95th percentile	0.289	0.177	0.142	0.063	0.632	0.764	0.845	0.924

Table 2: Asymptotic disagreement point, regret, and utility gains for each group under the regret optimal and fair policies, where groups are either based on race or age. The numbers are derived from the optimal solution to ($L(\theta)$) and ($L^{\text{fair}}(\theta)$) for the regret optimal and fair policies respectively, for the grouped linear contextual bandit instance based on the warfarin dataset. As regret scales logarithmically as $T \to \infty$, these numbers represent the coefficient of $\log T$ term.

		Race				Age		
		A	B	C	Total	A	B	Total
Regret	Disagreement point	25.6	74.8	78.6	179.1	164.7	78.0	242.8
	Regret optimal	1.9	5.6	71.1	78.6	151.6	23.2	174.8
	Fair	0.0	25.4	54.0	79.4	149.3	29.3	178.7
Utility Gain	Regret optimal	23.7	69.2	7.6	100.4	13.1	54.9	68.0
	Fair	25.6	49.4	24.6	99.6	15.4	48.7	64.1

Warfarin Dosing Case Study: Warfarin is a common blood thinner whose optimal dose varies widely across patients. We use a publicly available dataset [30] to evaluate the effect of using a proportionally fair policy on learning the optimal personalized dose of warfarin. A detailed description of the experimental setup is deferred to Appendix A.3. The dataset contains covariates and the optimal dose of warfarin for 5700 patients. Both the age and race of patients are available and we use these to define groups. We use a linear contextual bandit setup with five features and an intercept; three actions (dose levels) are available to any arriving patient.

The results in Table 2 shows that for both groups based on race and age, the fair solution effectively 'balances out' the utility gains across groups with a small increase in regret. For race, we see that the disagreement point for groups B and C are very similar, but the regret optimal solution ends up benefitting B substantially more than C. The fair solution is able to 'even out' the utility gain between C to B for a small increase in regret. For age, the impact of fairness is smaller than with race which is potentially since there is less opportunity to learn across age groups than across race.

References

[1] Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al. The battle trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1):44–53, 2011.

[2] Donald A Berry. Adaptive clinical trials in oncology. *Nature reviews Clinical oncology*, 9(4):199, 2012.

[3] Donald A Berry. The brave new world of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research. *Molecular oncology*, 9(5):951–959, 2015.

[4] Hope S Rugo, Olufunmilayo I Olopade, Angela DeMichele, Christina Yau, Laura J van't Veer, Meredith B Buxton, Michael Hogarth, Nola M Hylton, Melissa Paoloni, Jane Perlmutter, et al. Adaptive randomization of veliparib–carboplatin treatment in breast cancer. *New England Journal of Medicine*, 375(1):23–34, 2016.

[5] Richard M Bergenstal, Mary Johnson, Rebecca Passi, Anuj Bhargava, Natalie Young, Davida F Kruger, Eran Bashan, Stanley G Bisgaier, Deanna J Marriott Isaman, and Israel Hodish. Automated insulin dosing guidance to optimise insulin management in patients with type 2 diabetes: a multicentre, randomised controlled trial. *The Lancet*, 393(10176):1138–1148, 2019.

[6] Revital Nimri, Tadej Battelino, Lori M Laffel, Robert H Slover, Desmond Schatz, Stuart A Weinzimer, Klemen Dovc, Thomas Danne, and Moshe Phillip. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. *Nature medicine*, 26(9):1380–1384, 2020.

[7] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

[8] Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *ICML*, 2010.

[9] Deepak Agarwal, Bo Long, Jonathan Traupman, Doris Xin, and Liang Zhang. Laser: A scalable response prediction platform for online advertising. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 173–182, 2014.

[10] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[11] Christopher Jung, Sampath Kannan, and Neil Lutz. Quantifying the burden of exploration and the unfairness of free riding. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1892–1904. SIAM, 2020.

[12] Manish Raghavan, Aleksandrs Slivkins, Jennifer Vaughan Wortman, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. In *Conference on Learning Theory*, pages 1724–1738. PMLR, 2018.

[13] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2):245–272, 2010.

[14] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *arXiv preprint arXiv:1605.07139*, 2016.

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

[16] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *arXiv preprint arXiv:1802.06936*, 2018.

[17] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. Achieving fairness in the stochastic multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5379–5386, 2020.

[18] Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

[19] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

[20] Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. *arXiv preprint arXiv:1711.00400*, 2017.

[21] Bart Van Parys and Negin Golrezaei. Optimal learning for structured bandits. *Available at SSRN 3651397*, 2020.

[22] Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.

[23] H Peyton Young. *Equity: in theory and practice*. Princeton University Press, 1995.

[24] Amartya Sen and James Eric Foster. *On Economic Inequality*. Oxford university press, 1997.

[25] John F Nash. The bargaining problem. *Econometrica: Journal of the econometric society*, pages 155–162, 1950.

[26] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

[27] Mamoru Kaneko and Kenjiro Nakamura. The nash social welfare function. *Econometrica: Journal of the Econometric Society*, pages 423–435, 1979.

[28] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

[29] Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. The price of fairness. *Operations research*, 59(1):17–31, 2011.

[30] Michelle Whirl-Carrillo, Ellen M McDonagh, JM Hebert, Li Gong, K Sangkuhl, CF Thorn, Russ B Altman, and Teri E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012.

[31] International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

[32] Fumihiko Takeuchi, Ralph McGinnis, Stephane Bourgeois, Chris Barnes, Niclas Eriksson, Nicole Soranzo, Pamela Whittaker, Venkatesh Ranganath, Vasudev Kumanduri, William McLaren, et al. A genome-wide association study confirms vkorc1, cyp2c9, and cyp4f2 as principal genetic determinants of warfarin dose. *PLoS Genet*, 5(3):e1000433, 2009.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Total social welfare can decrease, and the extent of this is evaluated in Section 4.2 and is one focus of the experiments in Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]

(b) Did you include complete proofs of all theoretical results? [Yes] All proofs are in supplemental material.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report both median and 95 percentile.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] [31]
   (b) Did you mention the license of the assets? [Yes] In Appendix A.3.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code for experiments in supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] In Appendix A.3.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] In Appendix A.3.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Deferred Descriptions

A.1 Negative Externality Example from [12]

[12] provide an example of an instance where there exists a sub-population that is better off when UCB is run on that sub-population alone, compared to running UCB on the entire population. The example they provide depends on the total time horizon T. We claim that this does not occur when you fix an instance and consider asymptotic log-scaled regret, $\lim_{T\to\infty} \frac{R_T}{\log T}$.

Fix any time T_0, and consider the two-armed instance according to $T = T_0$ from Definition 1 of [12]. The population consists of three buckets that depend on their starting location: A, B, and C. The sub-population consisting of B and C is dubbed the "minority", while A is the "majority". Note that only B has access to both arms and hence it is the only bucket that can ever incur regret. Group B pulls the arm that has a higher UCB, defined as $\hat{\theta}_t(a) + \sqrt{\frac{\alpha \log T_0}{N_t(a)}}$ for some tuning parameter $\alpha > 0$.

We first summarize informally how the negative externality arises. Because arms 1 and 2 are so close together, even after $O(T_0)$ time steps, which arm has a higher UCB is not dominated by the difference between their empirical means, but it is dominated the second term of the UCB: $\sqrt{\frac{\alpha \log T_0}{N_t(a)}}$, which is just a function of the number of pulls $N_t(a)$. That is, group B essentially ends up pulling the arm that has fewer pulls. Therefore, when only the minority exists, since C only pulls arm 2, arm 1 ends up having a higher UCB, and hence B ends up always pulling arm 1. However, if the majority group exists, arm 1 always has more pulls than arm 2 since there are more people from A then C. Then, B ends up essentially always pulling arm 2. If arm 2 is the arm that has a lower true reward than arm 1, then regret is higher when the majority group exists — therefore, the existence of the majority can have a "negative externality" on the minority.

However, if we fix this instance and let $T \to \infty$, then no matter which arms is better, from Theorem C.1, the total log-scaled regret is 0 from running KL-UCB. Moreover, when the majority does not exist, then the minority incurs non-zero log-scaled regret when $\theta_1 < \theta_2$. Therefore, the presence of the majority can only help the minority. Now, as explained in [12], it is true that the presence of the majority can negatively affect the minority in the early time steps (i.e. $t < T_0$). In the asymptotic regime, such a negative externality corresponds to adding $o(\log T)$ regret, which is deemed insignificant in our setting.

A.2 Optimal Allocation Matching (OAM) Policy

We describe the OAM algorithm from [22].

Preliminaries: Let $G_t = \sum_{s=1}^{t-1} A_s A_s^\top$ and let $\hat{\theta}_t = G_t^{-1} \sum_{s=1}^{t-1} A_s Y_s$ be the least squares estimate of θ at time t. Let $\hat{\Delta}_t^m(a) = \max_{a' \in \mathcal{A}(m)} \langle a' - a, \hat{\theta}_t \rangle$ be the corresponding estimate of $\Delta^m(a)$. Let $\hat{\Delta}_t^{\min} = \min_{m\in[M]} \min_{a\in\mathcal{A}(m), \hat{\Delta}_t(m,a)>0} \hat{\Delta}_t(m,a)$ be the smallest nonzero instantaneous regret. Let

$$f_{T,\delta} = 2\left(1 + \frac{1}{\log T}\right) \log\left(\frac{1}{\delta}\right) + cd\log(d\log T),$$

where c is an absolute constant. Let $f_T = f_{T,1/T}$.

Define the following optimization problem that takes $\tilde{\Delta}(m,a)$ as input:

$$(K) \qquad \begin{aligned} &\min \sum_{m\in\mathcal{M}} \sum_{a\in\mathcal{A}(m)} Q(m,a)\tilde{\Delta}(m,a) \\ &\text{s.t.} \quad ||a||^2_{H_T^{-1}} \leq \frac{\tilde{\Delta}(m,a)^2}{f_T} \quad \forall m \in \mathcal{M}, a \in \mathcal{A}(m) \\ &\qquad Q(m,a) \geq 0 \quad \forall m \in \mathcal{M}, a \in \mathcal{A}, \end{aligned}$$

where $H_T = \sum_{m\in\mathcal{M}} \sum_{a\in\mathcal{A}(m)} Q(m,a)aa^\top$ is invertible. Let $(\hat{Q}_t(m,a))_{m\in\mathcal{M},a\in\mathcal{A}}$ be the solution to (K) using $\tilde{\Delta} = \hat{\Delta}_t$.

Algorithm: We are now ready to state the algorithm. At each time step t, observe context m_t and do the following. First, check whether

$$||a||^2_{G_t^{-1}} \leq \frac{\hat{\Delta}_t(m,a)^2}{f_T} \quad \forall a \in \mathcal{A}(m_t). \tag{10}$$

If (10) is satisfied, we exploit; otherwise, we explore.

Exploit: Pull the greedy arm: $\mathrm{argmax}_{a\in\mathcal{A}(m_t)}\langle a, \hat{\theta}_t\rangle$.

Explore: Let $s(t)$ be the total number of exploration rounds so far. Solve the empirical optimization problem (K) to get solution $\hat{Q}_t(m,a)$.

1. Check whether $N_t^{m_t}(a) \geq \min(\hat{Q}_t(m_t,a), f_T/(\hat{\Delta}_t^{\min})^2)$ holds for all available arms $a \in \mathcal{A}(m_t)$. If so, pull the UCB arm $A_t = \mathrm{argmax}_{a\in\mathcal{A}(m_t)}\langle a, \hat{\theta}_t\rangle + \sqrt{f_{T,1/s(t)^2}}||a||_{G_t^{-1}}$.
2. Check whether there exists an available arm $a \in \mathcal{A}(m_t)$ such that $N_t(a) \leq \varepsilon_t s(t)$, where $\varepsilon_t = 1/\log\log t$. If there is, then pull $A_t = \mathrm{argmin}_{a\in c\mathcal{A}^{m_t}} N_t(a)$.
3. If the above two criteria are not true, then pull $A_t = \mathrm{argmin}_{a\in\mathcal{A}^{m_t}} \frac{N_t(a)}{\min(\hat{Q}_t(m_t,a), f_T/(\hat{\Delta}_t^{\min})^2)}$.

A.3 Warfarin Experiment Details

We use a publicly available dataset for warfarin dosing that was collected by the Pharmacogenomics Knowledge Base (PharmGKB [30]), which is under a Creative Commons license[1]. The dataset contains 5700 patients who were treated with warfarin from 21 research groups over 9 countries. Consent for all patients was obtained previously from each center, and no personally identifiable information was used. The dataset contains the optimal dose of warfarin for each patient, which was found by doctors through trial and error. It also includes many other covariates for each patient including demographics, clinical features, and genetic information.

Groups: We group the patients either by race or age. There were three distinct races in the dataset, which we label as A, B ,and C. For age, we split the patients into two age groups, where the threshold age was 70.

Contexts: The OAM and PF-OAM policies assume a finite number of possible feature vectors, and the optimization problem $(L(\theta))$ scales with this number. Therefore, for tractability, we only use five features for the contexts of the patients, where we discretize each feature into two bins. We use the five features that are most correlated with the optimal warfarin dosage, and we use the empirical median of each feature to discretize them. The five features that we use are: age, weight, whether the patient was taking another drug (amiodarone), and two binary features capturing whether the patient has a particular genetic variant of genes Cyp2C9 and VKORC1, two genes that are known to affect warfarin dosage [32]. Out of $2^5 = 32$ different possible feature vectors, there were 21 that were present in the data.

Rewards: We bin the optimal dosage levels into three arms as was done in [7]: Low (under 3 mg/day), Medium (3-7 mg/day), and High (over 7 mg/day). To ensure that the model is correctly specified, for each arm, we train a linear regression model using the entire dataset from the five contexts to the binary reward on whether the optimal dosage for that patient belongs in that bin. Let $\theta_a \in \mathbb{R}^6$ be the learned linear regression parameter for each arm ($d = 6$ to include the intercept).[2] To model this as grouped linear contextual bandits as described in Section 5, we let $d = 18$ and let $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^d$. When a patient with covariates $X \in \mathbb{R}^6$ arrives, the actions available are $\{(X, \mathbf{0}, \mathbf{0}), (\mathbf{0}, X, \mathbf{0}), (\mathbf{0}, \mathbf{0}, X)\}$, and their expected reward from arm a is $\langle X, \theta_a\rangle$ for $a \in \{1, 2, 3\}$.

Algorithms: We assume a patient is drawn i.i.d. from the dataset at each time step, and we compute the asymptotic group regret of the OAM policy ('Regret optimal') and the fair extension ('Fair') as described in Section 5:

[1] `https://creativecommons.org/licenses/by-sa/4.0/`

[2] The linear regression step is done solely to remove model misspecification. The purpose of this study is not to show that the linear contextual bandit is a good fit for this dataset — this was already demonstrated in [7]. rather, the purpose is to demonstrate how incorporating fairness changes the outcome from a policy that does not take fairness into account on a bandit instance that approximates a real-world setting. rather, the purpose is to

- *Regret optimal:* Using the true values θ, we solve ($L(\theta)$) and obtain solution $(Q(m,a))_{m\in[M],a\in\mathcal{A}}$. Then, the total (log-scaled) regret incurred by context m is $\sum_{a\in\mathcal{A}} \Delta(m,a)Q(m,a)$. Since we assume the group arrivals are i.i.d., for each context, we allocate the regret to groups in proportion to the group's frequency. That is, for each m, let $(w^g(m))_{g\in\mathcal{G}}$, $\sum_{g\in\mathcal{G}} w^g(m) = 1$ be the empirical distribution of groups among patients with context m. Then, the total regret assigned to group g is $\sum_{m\in[M]} w^g(m) \sum_{a\in\mathcal{A}} \Delta(m,a)Q(m,a)$.
- *Fair:* Using the true values Δ, we solve ($L^{\text{fair}}(\theta)$) and obtain solution $(Q^g(m,a))_{g\in\mathcal{G},m\in[M],a\in\mathcal{A}}$. The total regret assigned to group g is $\sum_{m\in[M]} \sum_{a\in\mathcal{A}} \Delta(m,a)Q^g(m,a)$.

All experiments were run on a Macbook Pro with a 2.5 GHz Intel Core i7 processor.

B Proof Preliminaries

B.1 Notation

For all of the subsequent proofs, we assume that an instance $\mathcal{I}$ is *fixed*. We often use big-O notation, which is with respect to $T \to \infty$, unless otherwise specified. The big-O hides constants that may depend on any other parameter other than T, including the instance $\mathcal{I}$. In general, when we introduce a *constant*, it may depend on any other parameters other than T . We are usually not concerned with the values of the constants as we are concerned with asymptotic results (though we do concern ourselves with constants in front of the leading term, usually $\log T$). We sometimes re-use letters like c for constants but they do not refer to the same value.

The UCB of an arm is defined as:

$$\text{UCB}_t(a) = \max\{q : N_t(a)\mathbf{KL}(\hat{\theta}_t(a), q) \le \log t + 3\log\log t\}. \tag{11}$$

Let $\text{Pull}_t(a)$ be the indicator for arm a being pulled at time t, and let $\text{Pull}_t^g(a)$ be the indicator for when arm a is pulled by group g. We define the class of log-consistent policies:

Definition B.1. A policy π for the grouped bandit problem is *log-consistent* for if for any instance $(\theta, G, (p_g)_{g\in G}, (\mathcal{A}_g)_{g\in G})$, for any group g,

$$\mathbb{E}\left[\sum_{a\in\mathcal{A}_{\text{sub}}(g)} N_T^g(a)\right] = O(\log T). \tag{12}$$

That is, the expected number of times that group g pulled a suboptimal arm by time t is logarithmic in the number of arrivals of g.

B.2 Commonly Used Lemmas

We state a few lemmas that are used several times for both Theorem C.1 and Theorem 4.1. These lemmas do not depend on the policy that is used. The first result shows that the number of times that an arm's UCB is smaller than its true mean is small.

Lemma B.2. *Let $\Lambda_t = \{\text{UCB}_t(a) \ge \theta(a)\ \forall a \in \mathcal{A}\}$ be the event that UCB for every arm is valid at time t.*

$$\sum_{t=1}^{T} \Pr(\bar{\Lambda}_t) = O(\log\log T).$$

Proof. For a fix arm a, $\sum_{t=1}^T \Pr(\text{UCB}_t(a) < \theta(a)) = O(\log\log T)$ follows from Theorem 10 of [28], plugging in $\delta = \log t + 3\log\log t$ as is done in the proof of Theorem 2 of [28]. The result follows from a union bound over all actions $a \in \mathcal{A}$. □

The second lemma states a relationship between the radius of the UCB of an arm and the number of pulls of the arm.

Lemma B.3. *Let $0 < \alpha < \beta < 1$. There exists a constant $c > 0$ such that if $\hat{\theta}_t(a) \le \alpha$ and $\text{UCB}_t(a) \ge \beta$, then $N_t(a) < c\log t$.*

Proof. Suppose $\hat{\theta}_t(a) \le \alpha$ and $\mathrm{UCB}_t(a) \ge \beta$. Then, $\mathrm{KL}(\hat{\theta}_t(a), \mathrm{UCB}_t(a)) \ge \mathrm{KL}(\alpha, \beta)$. Let $c = \frac{4}{\mathrm{KL}(\alpha,\beta)}$. By definition of the UCB (11), $N_t(a) \le \frac{\log t + 3\log\log t}{\mathrm{KL}(\hat{\theta}_t(a), \mathrm{UCB}_t(a))} \le c \log t$. □

This result essentially states that if the radius of the UCB of an arm is larger than a constant, then the number of pulls of the arm is at most $O(\log t)$; this result follows simply from the definition of the UCB (11). The next result states that if an arm a is pulled, then its empirical mean will be close to its true mean.

Lemma B.4. *For any group g and arm $a \in \mathcal{A}^g$, if $L < \theta(a) < U$,*

$$\sum_{t=1}^{T} \Pr(\mathrm{Pull}_t(a), \hat{\theta}_t(a) \notin [L, U]) = O(1).$$

where big-O hides constants that may depend on the instance and L, U.

Proof. Let $\hat{\theta}^n(a)$ be the empirical mean after n pulls of arm a. Let $E_{t,n}$ be the event that the number of times arm 1 has been pulled before time t is exactly n.

$$\begin{aligned}
&\sum_{t=1}^{T} \Pr(\mathrm{Pull}_t(a), \hat{\theta}_t(a) \notin [L, U]) \\
&= \sum_{t=1}^{T} \sum_{n=1}^{T} \Pr(\mathrm{Pull}_t(a), \hat{\theta}^n(a) \notin [L, U], E_{t,n}) \\
&= \sum_{n=1}^{T} \sum_{t=1}^{T} \Pr(\hat{\theta}^n(a) \notin [L, U] \mid \mathrm{Pull}_t(a), E_{t,n}) \Pr(\mathrm{Pull}_t(a), E_{t,n})
\end{aligned}$$

If $F_{t,n} = \{\mathrm{Pull}_t(a), E_{t,n}\}$, then for any n, the events $F_{1,n}, \dots, F_{T,n}$ are disjoint. Then, by the law of total probability, $\Pr(\hat{\theta}^n(a) \notin [L, U]) \ge \sum_{t=1}^{T} \Pr(\hat{\theta}^n \notin [L, U] | F_{t,n}) \Pr(F_{t,n})$. Therefore,

$$\sum_{t=1}^{T} \Pr(\mathrm{Pull}_t(a), \hat{\theta}_t(a) \notin [L, U]) \le \sum_{n=1}^{T} \Pr(\hat{\theta}^n(a) \notin [L, U]) \le \sum_{n=1}^{T} \exp(-\alpha n).$$

for some $\alpha > 0$ since the rewards of arm a are Bernoulli. Therefore, $\sum_{t=1}^{T} \Pr(\mathrm{Pull}_t(a), \hat{\theta}_t(a) \notin [L, U]) = O(1)$. □

C Proof that KL-UCB is Regret Optimal

In this section, we prove that the KL-UCB policy is regret-optimal. At each time step, $\pi^{\text{KL-UCB}}$ chooses the arm with the highest UCB, defined as (11), out of all arms available.

Theorem C.1. *For all instances $\mathcal{I}$ of the grouped K-armed bandit,*

$$\liminf_{T \to \infty} \frac{R_T(\pi^{\text{KL-UCB}}, \mathcal{I})}{\log T} \le \sum_{a \in \mathcal{A}_{sub}} \Delta^{\Gamma(a)}(a) J(a). \tag{13}$$

The first step of the proof is to show that the *number of pulls* of a suboptimal arm is optimal:

Proposition C.2. *Let $a \in \mathcal{A}_{sub}$ be a suboptimal arm. KL-UCB satisfies*

$$\limsup_{T \to \infty} \frac{\mathbb{E}[N_T(a)]}{\log T} \le J(a).$$

This result can be shown using the existing analysis of KL-UCB from [28]. The next step is to analyze how these pulls are distributed across groups. In particular, we need to show that a group never pulls a suboptimal arm a if $g \notin \Gamma(a)$. This is the result of the next theorem:

Proposition C.3. *Let $a \in \mathcal{A}$. Let $g \in G_a$, $g \notin \Gamma(a)$ be a group that has access to the arm but is not the group that has the smallest optimal out of G_a. Then, KL-UCB satisfies*

$$\mathbb{E}[N_T^g(a)] = O(\log\log T),$$

where the big-O hides constants that depend on the instance.

This result implies that for any arm a, the regret incurred by group $g \notin \Gamma(a)$ pulling the arm is $o(\log T)$, and is equal to 0 when scaled by $\log T$. Theorem C.1 then follows from combining Proposition C.2 and Proposition C.3.

In this section, we prove Proposition C.3. Let $a \in \mathcal{A}$ and let $A \in \Gamma(a)$ be a group that has access to that arm with the smallest OPT. Let group $B \notin \Gamma(a)$ be another group that has access to arm a. Let θ^A, θ^B be the optimal arms for group A and B respectively. We use θ^A, θ^B to refer to both the arm and the arm means. Our goal is to show $\mathbb{E}\left[N_T^B(a)\right] = O(\log\log T)$.

$$\begin{aligned}\mathbb{E}\left[N_T^B(a)\right] &= \sum_{t=1}^{T} \Pr(\text{Pull}_t^B(a)) \\ &= \sum_{t=1}^{T} \Pr(\text{Pull}_t^B(a), \text{UCB}_t(\theta^B) \geq \theta^B) + \sum_{t=1}^{T} \Pr(\text{Pull}_t^B(a), \text{UCB}_t(\theta^B) < \theta^B).\end{aligned}$$

The second sum can be bounded by Lemma B.2, since $\sum_{t=1}^{T} \Pr(\text{Pull}_t^B(a), \text{UCB}_t(\theta^B) < \theta^B) \leq \sum_{t=1}^{T} \Pr(\bar{\Lambda}_t) = O(\log\log T)$. Therefore, our goal is to show

$$\sum_{t=1}^{T} \Pr(\text{Pull}_t^B(a), \text{UCB}_t(\theta^B) \geq \theta^B) = O(\log\log T). \tag{14}$$

We state a slightly more general result that implies (14).

Lemma C.4. *Suppose we run any log-consistent policy π. Let $r > 0$ be fixed. For any $a \in \mathcal{A}$,*

$$\sum_{t=1}^{T} \Pr(\text{Pull}_t(a), \text{UCB}_t(a) \geq \mathsf{OPT}(\Gamma(a)) + r) = O(\log\log T),$$

where the constant in the big-O may depend on the instance and r.

The rest of this section proves Lemma C.4.

C.1 Probabilistic Lower Bound of $N_t(a)$ for Grouped Bandit

One of the main tools used in the proof of Lemma C.4 is a high probability lower bound on the number of pulls of a suboptimal arm. Let $W_t(g)$ be the number of arrivals of group g by time t. Let $R_t^g = \{W_t(g) \geq \frac{p_g t}{2}\}$ be the event that the number of arrivals of group g is at least half of the expected value. We condition on the event R_t^g to ensure that a group has arrived a sufficient number of times.

Proposition C.5. *Let g be a group, and let $a \in \mathcal{A}_{\text{sub}}^g$ be a suboptimal arm for group g. Fix $\varepsilon \in (0, 1)$. Suppose we run a log-consistent policy as defined in Definition B.1. Then,*

$$\Pr\left(N_t(a) \leq \frac{(1-\varepsilon)\log t}{KL(\theta(a), \mathsf{OPT}(g))} \;\middle|\; R_t^g\right) = O\left(\frac{1}{\log t}\right),$$

where the big-O notation is with respect to $t \to \infty$.

The proof of this result can be found in Appendix D.3. For an arm $a \notin \mathcal{A}_{\text{sub}}$, we have the following stronger result:

Proposition C.6. *Let a be an arm that is optimal for some group g. Suppose we run a log-consistent policy. Then, for any $b > 0$,*

$$\Pr\left(N_t(a) \leq b\log t \;\middle|\; R_t^g\right) = O\left(\frac{1}{\log t}\right),$$

where the big-O notation is with respect to $t \to \infty$ and hide constants that depend on both b and the instance.

C.2 Proof of Lemma C.4

Outline: Let $A \in \Gamma(a)$ be a group that has the smallest optimal out of all arms with access to a. The main idea of this lemma is that group A does not "allow" the UCB of arm a to grow as large as $\mathsf{OPT}(A) + r$, as group A would pull arm a once the UCB is above $\mathsf{OPT}(A)$. Proposition C.5 implies that $\mathrm{UCB}_t(a)$ is not larger than $\mathsf{OPT}(A)$ with high probability. If this occurs at time t, since the radius of the UCB grows slowly (logarithmically), the earliest time that the UCB can grow to $\mathsf{OPT}(A) + r$ is t^γ, for some $\gamma > 1$. We divide the time steps into epochs, where if epoch k starts at time s_k, it ends at s_k^γ. This exponential structure gives us $O(\log \log T)$ epochs in total, and we show that the expected number of times that $\mathrm{UCB}_t(a) > \mathsf{OPT}(A) + r$ during one epoch is $O(1)$.

Proof: We denote by θ_a the true mean reward of arm a and by $\hat{\theta}_t$ the empirical mean reward of a at the start of time t. Let $U = \mathsf{OPT}(\Gamma(a)) + r$. Let $A \in \Gamma(a)$, and let $\theta^A = \mathsf{OPT}(A)$. If $a \notin \mathcal{A}_{\text{sub}}$, then let $\theta^A = \mathsf{OPT}(A) + r/2$. Let $b > 0$ such that $\frac{\mathrm{KL}(\theta_a, U)}{\mathrm{KL}(\theta_a, \theta^A)} = 1 + b$. Define $\theta_u \in [\theta_a, \theta^A]$ such that $\frac{\mathrm{KL}(\theta_u, U)}{\mathrm{KL}(\theta_a, \theta^A)} = 1 + \frac{b}{2}$. We have $\theta_a < \theta_u < \theta^A < U$. Define $\gamma \triangleq 1 + \frac{b}{4}$. Let $\varepsilon > 0$ such that $\frac{1-\varepsilon}{1+\varepsilon} \cdot \frac{\mathrm{KL}(\theta_u, U)}{\mathrm{KL}(\theta_a, \theta^A)} = \gamma$.

By Lemma B.4, $\sum_{t=1}^T \Pr(\mathrm{Pull}_t(a), \hat{\theta}_t(a) > \theta_u) = O(1)$. Therefore, we can assume $\hat{\theta}_t(a) \le \theta_u$. Denote the event of interest by $E_t = \{\mathrm{Pull}_t(a), \mathrm{UCB}_t(a) \ge \theta^A + r, \hat{\theta}_t(a) \le \theta_u\}$. Our goal is to show $\sum_{t=1}^T \Pr(E_t) = O(\log \log T)$.

Divide the time interval T into $K = O(\log \log T)$ epochs. Let epoch k start at $s_k \triangleq \left\lceil 2^{\gamma^k} \right\rceil$ for $k \ge 0$. Let $\mathcal{T}_k = \{s_k, s_k + 1, \ldots, s_{k+1} - 1\}$ be the time steps in epoch k. This epoch structure satisfies the following properties:

1. The total number of epochs is $O(\log \log T)$.
2. $\frac{\log s_{k+1}}{\log s_k} = \gamma$ for all $k \ge 0$.

We will treat each epoch separately. Fix an epoch k. Our goal is to bound $\mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t)\right]$. Lemma B.3 implies that there exists a constant $c > 0$ such that if E_t occurs, it must be that $N_t(a) < c \log t$. Hence,

$$\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t) \le c \log s_{k+1}.$$

Define the event $G_t = \left\{N_t(a) \ge (1 - \varepsilon) \frac{\log t}{\mathrm{KL}(\mu, \theta^A)}\right\}$. The following claim says that if G_{s_k} is true, then E_t never happens during that epoch.

Claim C.7. *Suppose G_{s_k} is true. Let t_0 be such that if $t \ge t_0$, $\log \log t \le \varepsilon \log t$. Then, if $s_k \ge t_0$, $\sum_{t=s_k}^{s_{k+1}} \mathbf{1}(E_t) = 0$.*

This result follows from the fact that the event G_{s_k} implies that the radius of the UCB is "small" at time s_k, and the epoch is defined so that the radius will not grow large enough that E_t can occur during epoch k. Therefore, we have the following:

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t)\right] = \mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t) \middle| \bar{G}_{s_k}\right] \Pr\left(\bar{G}_{s_k}\right) \le c \log s_{k+1} \Pr\left(\bar{G}_{s_k}\right).$$

We can bound $\Pr\left(\bar{G}_{s_k}\right)$ using the probabilistic lower bound of Proposition C.5.

Claim C.8. $\Pr\left(\bar{G}_{s_k}\right) \le O\left(\frac{1}{\log s_k}\right)$.

Then, property 2 of the epoch structure implies $\mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t)\right] = O(1)$. Since the number of epochs is $O(\log \log T)$,

$$\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}(E_t)\right] \le \sum_{k=1}^K \mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t)\right] = O(\log \log T),$$

as desired.

C.3 Proof of Claims

Proof of Claim C.7. Let $t = s_k > t_0$ and let $t' \geq t$ such that $E_{t'}$ is true. By definition of KL-UCB,

$$N_{t'}(a) \leq \frac{\log t' + 3 \log t'}{\mathrm{KL}(\hat{\theta}_{t'}, \mathrm{UCB}_{t'}(\theta))}.$$

Since $E_{t'}$ implies $\mathrm{UCB}_{t'}(a) > \theta^B$ and $\hat{\theta}_{t'} \leq \theta_u$, we have $N_{t'}(a) \leq \frac{\log t' + 3 \log t'}{\mathrm{KL}(\theta_u, \theta^B)}$. Since G_{s_k} is true, $N_{t'}(a) \geq (1-\varepsilon) \frac{\log s_k}{\mathrm{KL}(\theta_a, \theta^A)}$. Therefore, it must be that

$$\begin{aligned}
&(1-\varepsilon)\frac{\log s_k}{\mathrm{KL}(\theta_a, \theta^A)} \leq \frac{\log t' + 3 \log \log t'}{\mathrm{KL}(\theta_u, \theta^B)} \leq \frac{(1+\varepsilon)\log t'}{\mathrm{KL}(\theta_u, \theta^B)} \\
&\Rightarrow \frac{1-\varepsilon}{1+\varepsilon} \cdot \frac{\mathrm{KL}(\theta_u, \theta^B)}{\mathrm{KL}(\theta_a, \theta^A)} \log s_k \leq \log t' \\
&\Rightarrow t' \geq s_k^{\gamma}.
\end{aligned}$$

This implies that t' is not in epoch k. □

Proof of Claim C.8. For group $g = A$, Proposition C.5 (or Proposition C.6 if $a \notin \mathcal{A}_{\mathrm{sub}}$) states that

$$\Pr\left(\bar{G}_{s_k} \mid R^g_{s_k}\right) = O\left(\frac{1}{\log s_k}\right).$$

(We show in Appendix D.1 that KL-UCB is log-consistent.)

Now we need to bound $\Pr(\bar{R}^g_{s_k}) = \Pr\left(M_{s_k}(A) \leq \frac{p_A s_k}{2}\right)$. Note that $M_s(A) = \sum_{t=1}^{s} Z_i^A$, where $Z_t^A \overset{\text{iid}}{\sim} \mathrm{Bern}(p_A)$. By Hoeffding's inequality,

$$\Pr\left(M_{s_k}(A) \leq \frac{p_A s_k}{2}\right) < \exp\left(-\frac{1}{2} p_A^2 s_k\right).$$

Combining, we have

$$\Pr(\bar{G}_k) \leq \Pr(\bar{R}_k) + \Pr(\bar{G}_k \mid R_k) \leq O\left(\frac{1}{\log s_k}\right).$$

□

D Deferred Proofs for Theorem C.1

For any $\varepsilon > 0$, let

$$K^g_{\varepsilon}(x) = \left\lceil \frac{1+\varepsilon}{\mathrm{KL}(\theta_a, \mathsf{OPT}(g))} (\log x + 3 \log \log x) \right\rceil.$$

To show both Proposition C.2 and the fact that KL-UCB is log-consistent, we make use of the following lemma.

Lemma D.1. *Let $a \in \mathcal{A}$. Let $g \in G_a$ be a group in which a is suboptimal. For any $\varepsilon > 0$,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(\mathrm{Pull}^g_t(a), N_t(a) \geq K^g_{\varepsilon}(T))\right] = O(\log \log T). \tag{15}$$

Proof. Let $\varepsilon > 0$. Recall that A^*_g is the optimal arm for group g, and $\mathsf{OPT}(g)$ is the mean reward of A^*_g.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(\mathrm{Pull}^g_t(a), N_t(a) \geq K^g_{\varepsilon}(T))\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(\mathrm{Pull}^g_t(a), N_t(a) \geq K^g_{\varepsilon}(T), \mathrm{UCB}_t(A^*_g) \geq \mathsf{OPT}(g))\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(\mathrm{Pull}^g_t(a), \mathrm{UCB}_t(A^*_g) < \mathsf{OPT}(g))\right]
\end{aligned}$$

The second term is $O(\log\log T)$ from Lemma B.2. We will show that the first term is $O(1)$. Let $\hat{\theta}_s(a)$ be the empirical mean of a after s pulls. Consider the event $\{A_t = a, g_t = g, N_t(a) = s, \mathrm{UCB}_t(A_g^*) \geq \mathsf{OPT}(g)\}$, where $s \geq K_n$. Suppose this is true at time t. Then, it must be that $\mathrm{UCB}_t(a) \geq \mathsf{OPT}(g)$. For this to happen, by definition of KL-UCB, it must be that

$$s\mathsf{KL}(\hat{\theta}_s(a), \mathsf{OPT}(g)) \leq \log t + 3\log\log t. \tag{16}$$

Since $s \geq K_\varepsilon^g(T)$ and $t \leq T$, we must have

$$\mathrm{KL}(\hat{\theta}_s(a), \mathsf{OPT}(g)) \leq \frac{\log T + 3\log\log T}{K_\varepsilon^g(T)} = \frac{\mathrm{KL}(\theta_a, \mathsf{OPT}(g))}{1+\varepsilon}. \tag{17}$$

Let $r > \theta_a$ such that $\mathrm{KL}(r, \mathsf{OPT}(g)) = \frac{\mathsf{KL}(\theta_a, \mathsf{OPT}(g))}{1+\varepsilon}$. Then, for (17) to occur, it must be that $\hat{\theta}_s(a) \geq r$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(\mathrm{Pull}_t^g(a), N_t(a) \geq K_\varepsilon^g(n), \mathrm{UCB}_t(A_g^*) \geq \mathsf{OPT}(g))\right] \\
=&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=K_n}^{\infty} \mathbf{1}(\mathrm{Pull}_t^g(a), N_t(a) = s, \mathrm{UCB}_t(A_g^*) \geq \mathsf{OPT}(g))\right] \\
\leq&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=K_n}^{\infty} \mathbf{1}(\mathrm{Pull}_t^g(a), N_t(a) = s, \hat{\theta}_s(a) \geq r)\right] \\
=&\mathbb{E}\left[\sum_{s=K_n}^{\infty} \mathbf{1}(\hat{\theta}_s(a) \geq r)\sum_{t=1}^{T} \mathbf{1}(\mathrm{Pull}_t^g(a), N_t(a) = s)\right] \\
\leq& \sum_{s=K_n}^{\infty} \Pr(\hat{\theta}_s(a) \geq r).
\end{aligned}$$

Since $r > \mu(a)$, there exists a constant $C_3 > 0$ that depends on ε and r such that $\Pr(\mu_s(a) \geq r) \leq \exp(-sC_3)$. Therefore, $\sum_{s=K_n}^{\infty} \Pr(\hat{\theta}_s(a) \geq r) = O(1)$ and we are done.

□

D.1 Proof that KL-UCB is log-consistent

This basically follows from Lemma D.1. Let $\varepsilon = 1/2$. Fix a group g, and let a be a suboptimal arm for g.

$$\begin{aligned}
\mathbb{E}[N_T^g(a)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(\mathrm{Pull}_t^g(a))\right] \\
&\leq K_\varepsilon^g(T) + \mathbb{E}\left[\sum_{t=1}^{t_{g(n)}} \mathbf{1}(\mathrm{Pull}_t^g(a), N_t(a) \geq K_\varepsilon^g(T))\right] \\
&= K_\varepsilon^g(T) + \log\log(T).
\end{aligned}$$

We are done since $K_\varepsilon^g(T) = O(\log T)$.

D.2 Proof of Proposition C.2

Let $a \in \mathcal{A}_{\text{sub}}$ be a suboptimal arm. Let $\varepsilon > 0$. Let

$$K_T = \max_{g \in G_a} K_\varepsilon^g(T).$$

Clearly, the maximum is attained in the group g with the smallest $\mathsf{OPT}(g)$, so.

$$K_T = \left\lceil \frac{1+\varepsilon}{\mathsf{KL}(\theta_a, \mathsf{OPT}(\Gamma(a)))} (\log T + 3\log\log T) \right\rceil.$$

$$
\begin{aligned}
\mathbb{E}[N_T(a)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(A_t = a)\right] \\
&\le K_T + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(A_t = a, N_t(a) \ge K_T)\right] \\
&\le K_T + \sum_{g \in G_a} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(\mathrm{Pull}_t^g(a), N_t(a) \ge K_T)\right] \\
&\le K_T + \sum_{g \in G_a} O(\log\log T).
\end{aligned}
$$

where the last inequality follows from Eq. (15) of Lemma D.1. Since this holds for any $\varepsilon > 0$, the desired result holds.

D.3 Proof of Proposition C.5 and Proposition C.6

Let g be a group, and let j be a suboptimal arm for group g; i.e. $\theta_j < \mathsf{OPT}(g)$. Fix $\varepsilon > 0$. We assume that the event $R_t^g = \{W_t(g) \ge \frac{p_g t}{2}\}$ holds. Fix $\delta > 0$ such that $\frac{1-\delta}{1+\delta} = 1 - \varepsilon$. Let $a = \delta/2$. We construct another instance γ where arm j is replace with λ so that arm j is the optimal arm for g in the same manner as the Lai-Robbins proof. Specifically, $\lambda > \theta_j$ such that

$$
\mathsf{KL}(\theta_j, \lambda) = (1+\delta)\mathsf{KL}(\theta_j, \mathsf{OPT}(g)).
$$

Our goal is to bound the probability of event $\left\{N_t(j) \le \frac{(1-\delta)\log t}{\mathrm{KL}(\theta_j,\lambda)}\right\}$, which we split into two events:

$$
\begin{aligned}
C_t &= \left\{N_t(j) \le \frac{(1-\delta)\log t}{\mathrm{KL}(\theta_j,\lambda)}, L_{N_t(j)} \le (1-a)\log t\right\}, \\
E_t &= \left\{N_t(j) \le \frac{(1-\delta)\log t}{\mathrm{KL}(\theta_j,\lambda)}, L_{N_t(j)} > (1-a)\log t\right\},
\end{aligned}
$$

where $L_m = \sum_{i=1}^{m} \log\left(\frac{f(Y_i;\theta_j)}{f(Y_i;\lambda)}\right)$.

Assumption (12), there exists a constant c such that if t is large enough that $\Pr(R_t^g) \ge 1/2$,

$$
\mathbb{E}_\gamma\left[\sum_{a \in \mathcal{A}_{\text{sub}}} N_t^g(a) \,\middle|\, R_t^g\right] \le c\log t.
$$

Since j is the unique optimal arm under γ,

$$
\mathbb{E}_\gamma\left[W_t(g) - N_t^g(j) \,\middle|\, R_t^g\right] \le c\log t.
$$

Using Markov's inequality and using the fact that $W_t(g) \ge \frac{p_g t}{2}$, we get

$$
\begin{aligned}
\Pr_\gamma\left(N_t^g(j) \le \frac{(1-\delta)\log t}{\mathrm{KL}(\theta_j,\lambda)} \,\middle|\, R_t^g\right) &= \Pr_\gamma\left(W_t(g) - N_t^g(j) \ge W_t(g) - \frac{(1-\delta)\log t}{\mathrm{KL}(\theta_j,\lambda)} \,\middle|\, R_t^g\right) \\
&\le \Pr_\gamma\left(W_t(g) - N_t^g(j) \ge \frac{p_g t}{2} - \frac{(1-\delta)\log t}{\mathrm{KL}(\theta_j,\lambda)} \,\middle|\, R_t^g\right) \\
&\le \frac{\mathbb{E}\left[W_t(g) - N_t^g(j) \mid R_t^g\right]}{\frac{p_g t}{2} - \frac{(1-\delta)\log t}{\mathrm{KL}(\theta_j,\lambda)}} \\
&= O\left(\frac{\log t}{t}\right).
\end{aligned}
$$

Bounding $\Pr(C_t \mid R_t^g)$**:** Following through with the same steps as the original proof, we can replace (2.7) with

$$\Pr_\theta(C_t \mid R_t^g) \le t^{1-a} \Pr_\gamma(C_t \mid R_t^g) \le t^{1-a} O\left(\frac{\log t}{t}\right) = O\left(\frac{\log t}{t^a}\right).$$

Bounding $\Pr(E_t \mid R_t^g)$**:** Next, we need to show a probabilistic result in lieu of (2.8) of [10]. Let $m = \frac{(1-\delta)\log t}{\mathrm{KL}(\theta_j,\lambda)}$ and let $\alpha > 0$ such that $(1+\alpha) = \frac{1-a}{1-\delta}$. We need to upper bound

$$\begin{aligned}\Pr_\theta\left(\max_{j \le m} L_j > (1-a)\log t\right) &= \Pr_\theta\left(\max_{j \le m} L_j > (1+\alpha)\mathrm{KL}(\theta_j,\lambda)m\right) \\ &\le \Pr_\theta\left(\max_{j \le m}\{L_j - j\mathrm{KL}(\theta_j,\lambda)\} > \alpha\mathrm{KL}(\theta_j,\lambda)m\right).\end{aligned}$$

Let $Z_i = \log\left(\frac{f(Y_i;\theta_j)}{f(Y_i;\lambda)}\right) - \mathrm{KL}(\theta_j,\lambda)$. We have $\mathbb{E}[Z_i] = 0$. Let $\mathrm{Var}(Z_i) = \sigma^2$. Then, by Kolmogorov's inequality, we have

$$\begin{aligned}\Pr_\theta\left(\max_{j \le m}\sum_{i=1}^{j} Z_i > \alpha\mathrm{KL}(\theta_j,\lambda)m\right) &\le \frac{1}{\alpha^2\mathrm{KL}(\theta_j,\lambda)^2 m^2}\mathrm{Var}\left(\sum_{i=1}^{m} Z_i\right) \\ &= \frac{\sigma^2}{\alpha^2\mathrm{KL}(\theta_j,\lambda)^2 m} \\ &= O\left(\frac{1}{\log t}\right),\end{aligned}$$

since $m = \Theta(\log t)$.

Combine: Combining, we have

$$\begin{aligned}\Pr_\theta\left(N_t(j) \le \frac{(1-\delta)\log n}{\mathrm{KL}(\theta_j,\lambda)} \;\middle|\; R_t^g\right) &= \Pr_\theta(C_n \mid R_t^g) + \Pr_\theta(E_n \mid R_t^g) \\ &= O\left(\frac{\log t}{t^a}\right) + O\left(\frac{1}{\log t}\right).\end{aligned}$$

Since $\mathrm{KL}(\theta_j,\lambda) \le (1+\delta)\mathrm{KL}(\theta_j,\mathsf{OPT}(g))$ and $\frac{1-\delta}{1+\delta} = 1-\varepsilon$, we have

$$\Pr_\theta\left(N_t(j) \le \frac{(1-\varepsilon)\log t}{\mathrm{KL}(\theta_j,\mathsf{OPT}(g))} \;\middle|\; R_t^g\right) \le O\left(\frac{1}{\log t}\right)$$

as desired.

Proof of Proposition C.6. The proof of this result follows the same steps as Proposition C.5. Let $\varepsilon = 1/2$ and let $\theta^* > \theta_j$ so that $\frac{1-\varepsilon}{\mathrm{KL}(\theta_j,\theta^*)} = b$. In the proof of Proposition C.5, replace $\mathsf{OPT}(g)$ with θ^*. Then, the same proof goes through and we get $\Pr\left(N_t(j) \le b\log n \mid R_t^g\right) = O\left(\frac{1}{\log t}\right)$. □

E Proof of Theorem 4.1

To prove Theorem 4.1, our goal is to show that the total number of pulls of a suboptimal arm a is $J(a)\log T$, and those pulls are distributed amongst groups according to $q_*^g(a)$. The policy PF-UCB assigns arms in a way that the distribution of groups that have pulled arm a converges to $\hat{q}_t^g(a)$. Hence, our goal is to show that $\hat{q}_t^g(a)$ is usually "close" to $q_*^g(a)$.

Let $\delta_0 = \min_{a \ne a'} \frac{|\theta(a)-\theta(a')|}{4}$. For $\delta \in (0,\delta_0)$ let $H_t(\delta) = \{\hat{\theta}_t(a) \in [\theta(a)-\delta, \theta(a)+\delta]\ \forall a \in \mathcal{A}\}$ be the event that all arms are within their "δ-boundaries". Since $\delta < \delta_0$, this implies that the ranking of the arms do not change if $H_t(\delta)$ is true (i.e. $\theta(a) < \theta(a') \Rightarrow \hat{\theta}_t(a) < \hat{\theta}(a')$). We first state a result pertaining to the program $(P(\theta))$, which states that if $H_t(\delta)$ is true, the approximate solution $\hat{q}_t$ is also close to the true solution q_*.

Proposition E.1. *For any $\varepsilon > 0$, there exists $\delta > 0$ such that if $H_t(\delta)$, then $\hat{q}_t^g(a) \in [q_*^g(a) - \varepsilon, q_*^g(a) + \varepsilon]$ for all $a \in \mathcal{A}$ and $g \in \mathcal{G}$.*

The proof of Proposition E.1 can be found in Appendix G.4. This result implies that when we have good empirical estimates of θ (i.e. $H_t(\delta)$ is true), the policy of 'following' the solution $\hat{q}_t^g(a)$ will give us the desired 'split' of pulls between groups. Therefore, our goal is to show that suboptimal arms are pulled only when $H_t(\delta)$ is true.

For $a \in \mathcal{A}_{\text{sub}}^g$, there are two reasons why $\text{Pull}_t^g(a)$ would occur: (i) $a = A_t^{\text{UCB}}(g')$ for some group g', or (ii) $a = A_t^{\text{greedy}}(g)$. We show that the regret from (ii) is negligible:

Proposition E.2. *Let g be a group, and let $a \in \mathcal{A}_{\text{sub}}^g$ be a suboptimal arm for g.*

$$\sum_{t=1}^{T} \Pr(\text{Pull}_t^g(a), A_t^{\textit{greedy}}(g) = a) = O(\log\log T).$$

Therefore, all of the regret stems from pulls of type (i), when an arm has the highest UCB. The next result says that essentially all pulls occur when $H_t(\delta)$ is true:

Proposition E.3. *Let $\delta > 0$. For any group g and action $a \in \mathcal{A}_{\text{sub}}^g$,*

$$\sum_{t=1}^{T} \Pr(\text{Pull}_t^g(a), A_t^{\textit{greedy}}(g) \neq a, \bar{H}_t(\delta)) = O(\log\log T).$$

Lastly, we show that the total number of times an arm $a \in \mathcal{A}_{\text{sub}}$ is pulled matches the lower bound:

Proposition E.4. *Let $a \in \mathcal{A}_{\text{sub}}$.*

$$\lim_{T\to\infty} \frac{\mathbb{E}[N_T(a)]}{\log T} = J(a).$$

We now prove Theorem 4.1 using Propositions E.2-E.4.

Proof of Theorem 4.1. Fix a group g and an arm $a \in \mathcal{A}_{\text{sub}}^g$. Let $\varepsilon > 0$. Let $\delta \in (0, \delta_0)$ according to Proposition E.1. Let $H_t = H_t(\delta)$.

$$\begin{aligned}
\mathbb{E}[N_T^g(a)] &= \sum_{t=1}^{T} \Pr(\text{Pull}_t^g(a)) \\
&= \sum_{t=1}^{T} (\Pr(\text{Pull}_t^g(a), A_t^{\text{greedy}}(g) \neq a, H_t) \\
&\quad + \Pr(\text{Pull}_t^g(a), A_t^{\text{greedy}}(g) = a) + \Pr(\text{Pull}_t^g(a), A_t^{\text{greedy}}(g) \neq a, \bar{H}_t)) \\
&\leq \sum_{t=1}^{T} \Pr(\text{Pull}_t^g(a), a \in \mathcal{A}_t^{\text{UCB}}, H_t) + O(\log\log T). \qquad (18)
\end{aligned}$$

where the last step follows from Proposition E.3 and Proposition E.2.

First, assume that $a \notin \mathcal{A}_{\text{sub}}$. That is, there exists a group g' such that a is optimal for g'. We claim that $\Pr(\text{Pull}_t^g(a) \mid a \in \mathcal{A}_t^{\text{UCB}}, H_t) = 0$. Notice that when H_t is true, a is not the greedy arm for g, and moreover, $a \notin \hat{\mathcal{A}}_{\text{sub}}$. Therefore, a is not involved in the optimization problem ($P(\theta)$), and a is not the greedy arm for g, so g would not pull a when H_t is true. Therefore, $\text{Pull}_t^g(a) = 0$ when H_t is true. This implies that if $a \notin \mathcal{A}_{\text{sub}}$,

$$\lim_{T\to\infty} \frac{\mathbb{E}[N_T^g(a)]}{\log T} = 0. \qquad (19)$$

Next, assume $a \in \mathcal{A}_{\text{sub}}$. By definition of the algorithm, if $\{\text{Pull}_t^g(a), a \in \mathcal{A}_t^{\text{UCB}}\}$ occurs, then $N_t^g(a) \leq \hat{q}_t^g(a) N_t(a)$. If $H_t(\delta)$, then $\hat{q}_t^g(a) \leq q_t^g(a) + \varepsilon$. Therefore, $\sum_{t=1}^{T} \mathbf{1}(\text{Pull}_t^g(a), a \in$

$\mathcal{A}_t^{\mathrm{UCB}}, H_t(\delta)) \le (q_t^g(a) + \varepsilon) N_T(a)$. Then, using (18), we can write

$$\begin{aligned}
\limsup_{T\to\infty} \frac{\mathbb{E}[N_T^g(a)]}{\log T} &= \limsup_{T\to\infty} \frac{\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}(\mathrm{Pull}_t^g(a), a \in \mathcal{A}_t^{\mathrm{UCB}}, H_t(\delta))\right] + O(\log\log T)}{\log T} \\
&\le \limsup_{T\to\infty} \frac{(q^g(a)+\varepsilon)\mathbb{E}[N_T(a)]}{\log T} \\
&\le (q^g(a)+\varepsilon) J(a),
\end{aligned}$$

where the last inequality follows from Proposition E.4. Since this holds for all $\varepsilon > 0$,

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_T^g(a)]}{\log T} \le q^g(a) J(a). \tag{20}$$

Recall that Proposition E.4 states

$$\lim_{T\to\infty} \frac{\mathbb{E}[N_T(a)]}{\log T} = J(a). \tag{21}$$

This implies that (20) must be an equality all g. If this weren't the case, then $\limsup_{T\to\infty} \frac{\mathbb{E}[N_T(a)]}{\log T}$ would be strictly less than $J(a)$, which would be a contradiction.

Moreover, we claim that (20) and (21) implies $\lim_{T\to\infty} \frac{\mathbb{E}[N_T^g(a)]}{\log T} = q^g(a)J(a)$ for all g. By contradiction, suppose there exists a $g' \in \mathcal{G}$ such that $\liminf_{T\to\infty} \frac{\mathbb{E}[N_T^{g'}(a)]}{\log T} = q^{g'}(a)J(a) - \alpha$ for some $\alpha > 0$. Then, (21) implies that $\limsup_{T\to\infty} \sum_{g\neq g'} \frac{\mathbb{E}[N_T^{g'}(a)]}{\log T} \ge (1 - q^{g'}(a))J(a) + \alpha$, which is a contradiction. Therefore, for every g,

$$\lim_{T\to\infty} \frac{\mathbb{E}[N_T^g(a)]}{\log T} = q^g(a) J(a).$$

Combining with (19) yields the desired result:

$$\lim_{T\to\infty} \frac{\mathbb{E}[\mathrm{Regret}_T^g(a)]}{\log T} = \lim_{T\to\infty} \frac{\sum_{a\in\mathcal{A}} \Delta^g(a)\mathbb{E}[N_T^g(a)]}{\log T} = \lim_{T\to\infty} \sum_{a\in\mathcal{A}_{\mathrm{sub}}} \Delta^g(a) q^g(a) J(a).$$

□

E.1 Proof of Propositions E.2-E.4

Proof of Proposition E.2. Let $g \in \mathcal{G}$ and let $a \in \mathcal{A}_{\mathrm{sub}}^g$. We bound $\sum_{t=1}^T \Pr(\mathrm{Pull}_t^g(a), a = A_t^{\mathrm{greedy}}(g))$. We can assume that the events $\hat\theta_t(a) \in [\theta(a)-\delta, \theta(a)+\delta]$ and Λ_t occur using Lemma B.4, and Lemma B.2 respectively. Since a is the greedy arm, it must be that $\hat\theta_t(a') \le \theta(a) + \delta$ for all $a' \in \mathcal{A}^g$.

Define the event

$$R_t = \{A_t^{\mathrm{greedy}}(g) = a, \Lambda_t, \hat\theta_t(a) \le \theta(a) + \delta, \hat\theta_t(a') \le \theta(a) + \delta \ \forall a' \in \mathcal{A}^g\}.$$

Our goal is to bound $\sum_{t=1}^T \Pr(R_t)$.

For R_t to occur, $\hat\theta_t(a') \le \theta(a) + \delta$ (since a is the greedy arm) and $\mathrm{UCB}_t(a') \ge \mathsf{OPT}(g)$ (since Λ_t) for all $a' \in \mathcal{A}_{\mathrm{opt}}^g$. By Lemma B.3 there exists a constant $c > 0$ such that if $N_t(a') > c\log t$ for some $a' \in \mathcal{A}_{\mathrm{opt}}^g$, R_t cannot happen. Moreover, for every $a' \in \mathcal{A}_{\mathrm{opt}}^g$, $\Pr(N_t(a') < c\log t) < O\left(\frac{1}{\log t}\right)$ from Proposition C.6.

Divide the time period into epochs, where epoch k starts at time $s_k = 2^{2^k}$. Let $\mathcal{T}_k$ be the time steps in epoch k. Let $G_k = \{N_{s_k}(a) > 3c\log s_k \ \forall a \in \mathcal{A}_{\mathrm{opt}}^g\}$ be the event that all optimal arms were pulled at least $3c\log s_k$ times by the start of epoch k. If G_k occurs, since $s_k = \sqrt{s_{k+1}}$,

$N_{s_{k+1}}(a) > \frac{3}{2} r \log s_{k+1} > r \log s_{k+1}$, and hence R_t can never happen during epoch k. Moreover, $\Pr(\bar{G}_k) = O\left(\frac{1}{\log s_k}\right)$ for any k.

Suppose we are in a "bad epoch", where G_k does not occur. We claim that R_t can't occur more than $O(\log s_{k+1})$ times during epoch k. For R_t to occur, the arm j with the highest UCB satisfies $\mathrm{UCB}_t(j) \geq \mathsf{OPT}(g)$ and $\hat{\theta}_t(j) \leq \theta(a) + \delta$.

Claim E.5. *For any action $j \in \mathcal{A}^g$, $\sum_{t=1}^{s} \Pr(A_t^{UCB}(g) = j, \mathrm{UCB}_t(j) \geq \mathsf{OPT}(g), \hat{\theta}_t(j) \leq \theta(a) + \delta \mid \bar{G}_k) = O(\log s)$.*

Using Claim E.5 and taking a union bound over all actions j implies $\sum_{t \in \mathcal{T}_k} \Pr(R_t \mid \bar{G}_k) = \sum_{t \in \mathcal{T}_k} \sum_{j \in \mathcal{A}^g} \Pr(R_t, A_t^{\mathrm{UCB}}(g) = j \mid \bar{G}_k) = O(\log s_{k+1})$. Since $\Pr(\bar{G}_k) = O\left(\frac{1}{\log s_k}\right)$, $\sum_{t \in \mathcal{T}_k} \Pr(R_t) = O(1)$. Since there are $O(\log \log T)$ epochs, $\sum_{t=1}^{T} \Pr(R_t) = O(\log \log T)$.

$\square$

Proof of Proposition E.3. Let $H_t = H_t(\delta)$. Fix a group g and an arm $a \in \mathcal{A}_{\mathrm{sub}}^g$. For g to pull a when $A_t^{\mathrm{greedy}}(g) \neq a$, it must be that $a \in \mathcal{A}_t^{\mathrm{UCB}}$.

First, assume $a \notin \mathcal{A}_{\mathrm{sub}}$. Then, there exist groups $G \subseteq \mathcal{G}$ in which a is optimal. If a is the greedy arm for some $g' \in G$, then $a \notin \hat{\mathcal{A}}_{\mathrm{sub}}$, implying a is not considered in the optimization problem $(\hat{P}_t)$. In this case, group g would never pull arm a. Therefore, it must be that a is not the greedy arm for all groups in G. We show the following lemma, which proves the proposition for an arm $a \notin \mathcal{A}_{\mathrm{sub}}$.

Lemma E.6. *Let $a \notin \mathcal{A}_{\mathrm{sub}}$, and let G be the set of groups in which a is optimal. Then,*

$$\sum_{t=1}^{T} \Pr(\mathrm{Pull}_t(a), A_t^{greedy}(g) \neq a \ \forall g \in G, a \in \mathcal{A}_t^{\mathrm{UCB}}) = O(\log \log T).$$

Now assume $a \in \mathcal{A}_{\mathrm{sub}}$. We assume that the events Λ_t and $\hat{\theta}_t(a) \in [\theta(a) - \delta, \theta(a) + \delta]$ hold using Lemma B.2 and Lemma B.4. Since $a \in \mathcal{A}_t^{\mathrm{UCB}}$ and Λ_t, it must be that $\mathrm{UCB}_t(a) \geq \mathsf{OPT}(\Gamma(a))$. Let $E_t = \{\mathrm{Pull}_t^g(a), \Lambda_t, \hat{\theta}_t(a) \in [\theta(a) - \delta, \theta(a) + \delta], \mathrm{UCB}_t(a) \geq \mathsf{OPT}(\Gamma(a))\}$ Our goal is to show

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(E_t, \bar{H}_t)\right] = O(\log \log T).$$

Divide the time interval into epochs, where epoch k starts at time $s_k = 2^{2^k}$. Let $K = O(\log \log T)$ be the total number of epochs. Let $\mathcal{T}_k$ be the time steps in epoch k.

Let $H_k = \cap_{t \in \mathcal{T}_k} H_t$. Clearly, if H_k is true, then by definition, $\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t, \bar{H}_t) = 0$. Therefore, we can write

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(E_t, \bar{H}_t)\right] = \sum_{k=1}^{K} \mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t, \bar{H}_t)\right] = \sum_{k=1}^{K} \left(\mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t, \bar{H}_t) \,\middle|\, \bar{H}_k\right] \Pr(\bar{H}_k)\right)$$

We bound the expectation and the probability separately.

1) Bounding $\mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t, \bar{H}_t) \,\middle|\, \bar{H}_k\right]$: If E_t occurs at some time step t, $\mathrm{UCB}_t(a) \geq \mathsf{OPT}(\Gamma(a))$ and $\hat{\theta}_t(a) \leq \theta(a) + \delta$. By Lemma B.3 it must be that $N_t(a) = O(\log t)$. Clearly, $N_s(a) \geq \sum_{t=1}^{s} \mathbf{1}(E_t)$, implying that $\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t) = O(\log s_{k+1})$. Therefore, $\sum_{t \in \mathcal{T}_k} \mathbf{1}(E_t, \bar{H}_t) \leq \sum_{t=1}^{s_{k+1}} \mathbf{1}(E_t) = O(\log s_{k+1})$

2) Bounding $\Pr(\bar{H}_k)$: For $a \in \mathcal{A}_{\mathrm{sub}}$ let $c_a = \frac{0.9}{\mathsf{KL}(\theta(a), \mathsf{OPT}(\Gamma(a)))}$. For $a \notin \mathcal{A}_{\mathrm{sub}}$, let $c_a = 1$. Let $F_k = \{\hat{\theta}_{s_k}(a) \in [\theta(a) - \delta/2, \theta(a) + \delta/2], N_{s_k}(a) \geq c_a \log s_k \ \forall a \in \mathcal{A}\}$ be the event that at time s_k, all arms a have been pulled $c_a \log s_k$ times and all arms are within an "inner" boundary (half as small as the boundary defined for H_t). We bound $\Pr(\bar{H}_k)$ by conditioning on the event F_k. Firstly, we bound $\Pr(\bar{F}_k)$ using the probabalistic lower bound of Proposition C.5-C.6:

Lemma E.7. *For any k,* $\Pr(\bar{F}_k) = O\left(\frac{1}{\log s_k}\right)$.

Next, we show that if F_k is true, then H_k occurs with probability at least $1 - O\left(\frac{1}{\log s_k}\right)$.

Lemma E.8. *For any action a,* $\Pr\left(\hat{\theta}_t(a) \notin [\theta(a) - \delta, \theta(a) + \delta] \text{ for some } t \in \mathcal{T}_k \mid F_k\right) \leq O\left(\frac{1}{\log s_k}\right)$.

Therefore,

$$\Pr(\bar{H}_k) \leq \Pr(\bar{F}_k) + \Pr(\bar{H}_k \mid F_k) = O\left(\frac{1}{\log s_k}\right).$$

3) Combine: Combining, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(E_t, \bar{H}_t)\right] &\leq \sum_{k=1}^{K} \left(O(\log s_{k+1}) O\left(\frac{1}{\log s_k}\right)\right) \\
&\leq \sum_{k=1}^{K} O(1) \\
&= O(\log \log T),
\end{aligned}$$

where the last inequality follows due to the fact that $\frac{\log s_{k+1}}{\log s_k} = 2$ for any k. □

Proof of Proposition E.4. Let $a \in \mathcal{A}_{\text{sub}}$. We need to show $\limsup_{T\to\infty} \frac{\mathbb{E}[N_T(a)]}{\log T} \leq J(a)$, as the lower bound is implied by (4). By Proposition E.2, the number of times a is pulled when a is the greedy arm for some group g is $O(\log \log T)$. Therefore,

$$\mathbb{E}[N_T(a)] = \sum_{t=1}^{T} \Pr(\text{Pull}_t(a), a \in \mathcal{A}_t^{\text{UCB}}, H_t(\delta)) + O(\log \log T).$$

The rest of the proof relies on the same argument as Proposition C.2. The main idea is that after $J(a) \log T + o(\log T)$ pulls of a, the UCB of a will not be larger than $\mathsf{OPT}(\Gamma(a))$, and therefore $a \notin \mathcal{A}_t^{\text{UCB}}$. □

E.2 Deferred Proofs

Proof of Claim E.5. Recall that $G_k = \{N_{s_k}(a) > 3c \log s_k \ \forall a \in \mathcal{A}_{\text{opt}}^g\}$. We will show $\sum_{t=1}^{T} \Pr(A_t^{\text{UCB}} = j, \text{UCB}_t(j) \geq \mathsf{OPT}(g), \hat{\theta}_t(j) \leq \theta(a) + \delta \mid \bar{G}_k) = O(\log \log T)$. From Lemma B.3, there exists a constant c' such that if $N_t(j) > c' \log T$ then, $\{\text{UCB}_t(j) \geq \mathsf{OPT}(g), \hat{\theta}_t(j) \leq \theta(a) + \delta\}$ cannot occur.

$$\begin{aligned}
&\sum_{t \in \mathcal{T}_k} \Pr(A_t^{\text{UCB}}(g) = j, \text{UCB}_t(j) \geq \mathsf{OPT}(g), \hat{\theta}_t(j) \leq \theta(a) + \delta \mid \bar{G}_k) \\
&= \sum_{n=1}^{c' \log T} \sum_{t \in \mathcal{T}_k} \Pr(A_t^{\text{UCB}}(g) = j, \text{UCB}_t(j) \geq \mathsf{OPT}(g), \hat{\theta}_t(j) \leq \theta(a) + \delta, N_t(a) = n \mid \bar{G}_k) \\
&\leq \sum_{n=1}^{c' \log T} \sum_{t \in \mathcal{T}_k} \Pr(A_t^{\text{UCB}}(g) = j, N_t(a) = n \mid \bar{G}_k). \qquad (22)
\end{aligned}$$

Our goal is to show that $\sum_{t \in \mathcal{T}_k} \Pr(A_t^{\text{UCB}}(g) = j, N_t(a) = n \mid \bar{G}_k) = O(1)$ for any n. Fix n, and write

$$\sum_{t \in \mathcal{T}_k} \Pr(A_t^{\text{UCB}}(g) = j, N_t(j) = n \mid \bar{G}_k) = \mathbb{E}\left[\sum_{t \in \mathcal{T}_k} \mathbf{1}(A_t^{\text{UCB}}(g) = j, N_t(j) = n) \,\middle|\, \bar{G}_k\right]$$

Let $L_t = \mathbf{1}(A_t^{\text{UCB}}(g) = j, N_t(j) = n)$ be the indicator for the event of interest. Our goal is to count the number of times L_t occurs. Let $Y_m = \{\exists\, t : \sum_{s=1}^{t} L_s = m\}$ be the event that L_s occurs at least m times. Note that for Y_m to occur, it must be that Y_{m-1} occurred. Therefore, by expliciting writing out the expectation, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}(A_t^{\text{UCB}}(g) = j, N_t(j) = n) \,\middle|\, \bar{G}_k\right] \le \sum_{m\ge1} m\Pr(Y_m \mid \bar{G}_k)$$
$$= \sum_{m\ge1} m\Pr(Y_m \mid Y_{m-1}, \bar{G}_k)\Pr(Y_{m-1} \mid \bar{G}_k).$$

We claim that there exists a $\lambda \in (0, 1)$ such that $\Pr(Y_m \mid Y_{m-1}, \bar{G}_k) \le \lambda$. Let τ be the time when L_s occurred for the $m-1$'th time, which exists since Y_{m-1} is true. For Y_m to occur, it must be that arm j was not pulled at time τ, even though arm j is the UCB. Given that j is the UCB, there exists a group g in which $N_\tau^g(a) \le \hat{q}_t^g(a)N_\tau(a)$. If such a group arrives, it will pull j with probability at least $\frac{1}{K}$. Therefore, at time τ, the probability that arm j will be pulled is at least $\min_{g\in G}\frac{p_g}{K}$. Then, $\lambda = 1 - \min_{g\in G}\frac{p_g}{K}$ satisfies $\Pr(Y_m \mid Y_{m-1}, \bar{G}_k) \le \lambda$.

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}(A_t^{\text{UCB}} = j, N_t(j) = n) \,\middle|\, \bar{G}_k\right] = \sum_{m\ge1} m\Pr(Y_m \mid Y_{m-1}, \bar{G}_k)\Pr(Y_{m-1} \mid \bar{G}_k)$$
$$\le \sum_{m\ge1} m\lambda^m$$
$$= O(1).$$

Substituting back into (22) gives

$$\sum_{t=1}^{T}\Pr(A_t^{\text{UCB}} = j, \text{UCB}_t(j) \ge \mathsf{OPT}(g), \hat{\theta}_t(j) \le \theta(a) + \delta \mid \bar{G}_k) \le \sum_{n=1}^{c'\log T} O(1) = O(\log T).$$

□

Proof of Lemma E.6. Let $a \notin \mathcal{A}_{\text{sub}}$, let G be the set of groups in which a is an optimal arm. We condition on whether a is the UCB for some group in G.

First, suppose $a = A_t^{\text{UCB}}(g)$ for some group $g \in G$, implying $\theta(a) = \mathsf{OPT}(g)$. We can assume $\hat{\theta}_t(a) > \mathsf{OPT}(g) - \delta$ from Lemma B.4. Then, if a is not the greedy arm for g, there exists a suboptimal arm $j \in \mathcal{A}_{\text{sub}}^g$ with higher mean but lower UCB than a. This implies that the UCB radius of j is smaller than the UCB radius of a, implying that j was pulled more times: $N_t(j) \ge N_t(a)$. We show that this event cannot happen often. Let $E_t = \{\text{Pull}_t(a), A_t^{\text{greedy}}(g) \ne a, a \in \mathcal{A}_t^{\text{UCB}}, a = A_t^{\text{UCB}}(g), \hat{\theta}_t(a) > \mathsf{OPT}(g) - \delta\}$. For any $j \in \mathcal{A}_{\text{sub}}^g$,

$$\sum_{t=1}^{T}\mathbf{1}(E_t, N_t(j) \ge N_t(a), \hat{\theta}_t(j) > \mathsf{OPT}(g) - \delta)$$
$$\le \sum_{t=1}^{T}\sum_{n=1}^{t}\sum_{n_j=n}^{t}\mathbf{1}(E_t, \hat{\theta}_{n_j}(j) > \mathsf{OPT}(g) - \delta, N_t(j) = n_j, N_t(a) = n)$$
$$\le \sum_{n_j=1}^{T}\mathbf{1}(\hat{\theta}_{n_j}(j) > \mathsf{OPT}(g) - \delta)\sum_{n=1}^{n_j}\sum_{t=n}^{T}\mathbf{1}(E_t, N_t(a) = n)$$
$$\le \sum_{n_j=1}^{T}\mathbf{1}(\hat{\theta}_{n_j}(j) > \mathsf{OPT}(g) - \delta)n_j,$$

where the last inequality uses $\sum_{t=n}^{T} \mathbf{1}(E_t, N_t(a) = n) \leq 1$ (since pulling arm a increasing $N_t(a)$ by 1). Since $\Pr(\hat{\theta}_n(j) > \mathsf{OPT}(g) - \delta) \leq \exp(-cn)$ for some constant $c > 0$, $\sum_{t=1}^{T} \Pr(E_t, N_t(j) \geq N_t(a), \hat{\theta}_t(j) > \mathsf{OPT}(g) - \delta) = O(1)$. Taking a union bound over actions $j \in \mathcal{A}_{\text{sub}}^g$ gives us the desired result:

$$\sum_{t=1}^{T} \Pr(\text{Pull}_t(a), A_t^{\text{greedy}}(g) \neq a \; \forall g \in G, a \in \mathcal{A}_t^{\text{UCB}}, \exists g \in G : a = A_t^{\text{UCB}}(g)) = O(\log \log T).$$

Now, suppose $a \notin A_t^{\text{UCB}}(g)$ for all $g \in G$. This means that there is another group h where $a = A_t^{\text{UCB}}(h)$, but a is suboptimal for h. We assume Λ_t holds. Let a_h be an optimal arm for h. Since Λ_t, $\text{UCB}_t(a_h) \geq \mathsf{OPT}(h)$. Therefore, it must be that $\text{UCB}_t(a) \geq \mathsf{OPT}(h)$. By Lemma C.4,

$$\sum_{t=1}^{T} \Pr(\text{Pull}_t(a), \text{UCB}_t(a) \geq \mathsf{OPT}(h)) = O(\log \log T).$$

This finishes the proof. □

Proof of Lemma E.7. Fix $a \in \mathcal{A}$ and time t. We will show $\Pr(\hat{\theta}_{s_k}(a) \in [\theta(a) - \delta/2, \theta(a) + \delta/2], N_{s_k}(a) \geq c_a \log s_k) \geq 1 - O\left(\frac{1}{\log t}\right)$. Then the result follows from taking a union bound over actions. We first show that PF-UCB is log-consistent.

Lemma E.9. *PF-UCB is log-consistent.*

Let $g \in \Gamma(a)$. Since $\Pr(M_t(a) < \frac{p_g}{2} t) \leq \exp(-\frac{1}{2} p_g t)$, we can assume that there have been at least $\frac{p_g}{2} t$ arrivals of g by time t. Then, using Proposition C.5 and Proposition C.6, we know that at time t, $\Pr(N_t(a) < c_a \log t | M_t(a) \geq \frac{p_g}{2} t) \leq O\left(\frac{1}{\log t}\right)$. Next, we show that the probability of the event $\hat{\theta}_t(a) \notin [\theta(a) - \delta/2, \theta(a) + \delta/2]$ given that we have more than $c_a \log t$ pulls of a is small.

$$\begin{aligned}
&\Pr(\hat{\theta}_t(a) \notin [\theta(a) - \delta/2, \theta(a) + \delta/2] \mid N_t(a) \geq c_a \log t) \\
= &\sum_{n = c_a \log t}^{t} \Pr(\hat{\theta}_n(a) \notin [\theta(a) - \delta/2, \theta(a) + \delta/2] \mid N_t(a) = n) \Pr(N_t(a) = n) \\
\leq &\sum_{n = c_a \log t}^{t} \exp(-c_1 n) \Pr(N_t(a) = n) \\
\leq &c_3 \exp(-c_2 \log t) \\
\leq &\frac{c_3}{t^{c_2}},
\end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$ that depends on the instance, a, and δ. Combining, we have that for any action a, $\Pr(\hat{\theta}_{s_k}(a) \in [\theta(a) - \delta/2, \theta(a) + \delta/2], N_{s_k}(a) \geq c_a \log s_k) \geq 1 - O\left(\frac{1}{\log t}\right)$.

□

Proof of Lemma E.8. Let $U_a = \theta(a) + \delta$ and $U_a^I = \theta(a) + \delta/2$. Let $\eta = U_a - U_a^I$. Since F_k is true, $N_{s_k}(a) \geq c_a \log s_k$. Let $n_1 = N_{s_k}(a)$. Let $\theta^n(a)$ be the empirical average of arm a after n pulls. We will bound

$$\Pr(\cup_{n_2 = n_1 + 1}^{\infty} \{\hat{\theta}^{n_2}(a) \notin [L_a, U_a]\} \mid \hat{\theta}^{n_1}(a) \in [L_a^I, U_a^I]).$$

For any n_2, $\hat{\theta}^{n_2}(a) > U_a$ implies $\hat{\theta}^{n_2}(a) > \hat{\theta}^{n_1}(a) + \eta$. Fix $n_2 > n_1$. Let $m = n_2 - n_1$.

$$\begin{aligned}
\left\{\hat{\theta}^{n_2}(a) > U_a\right\} &= \left\{\sum_{i=1}^{n_2} X_i > n_2 U_a\right\} \\
&= \left\{n_1 \hat{\theta}^{n_1}(a) + \sum_{i=n_1+1}^{n_2} X_i > n_2 U_a\right\} \\
&= \left\{\sum_{j=1}^{m} X_{n_1+j} > n_1(U_a - \hat{\theta}^{n_1}(a)) + m U_a\right\} \\
&= \left\{\sum_{j=1}^{m} (X_{n_1+j} - \mu) > n_1(U_a - \hat{\theta}^{n_1}(a)) + m(U_a - \mu)\right\}
\end{aligned}$$

Case $m \le n_1$: Since $U_a - \mu > 0$ and $U_a - \hat{\theta}^{n_1}(a) > \eta$ if F_k is true,

$$\begin{aligned}
\Pr\left(\bigcup_{m=1}^{n_1} \{\hat{\theta}^{n_1+m}(a) > U_a\} \;\middle|\; F_k\right) &\le \Pr\left(\bigcup_{m=1}^{n_1} \left\{\sum_{j=1}^{m} (X_{n_1+j} - \mu) > n_1 \eta\right\} \;\middle|\; F_k\right) \\
&\le \Pr\left(\max_{m=1,\ldots,n_1} S_m > n_1 \eta \;\middle|\; F_k\right),
\end{aligned}$$

where $S_m = \sum_{j=1}^{m}(X_{n_1+j} - \mu)$. Given that $X_{n_1+j} - \mu$ are zero mean independent random variables, by Kolomogorov's inequality, we have

$$\begin{aligned}
\Pr\left(\bigcup_{m=1}^{n_1} \{\hat{\theta}^{n_1+m}(a) > U_a\} \;\middle|\; F_k\right) &\le \frac{1}{n_1^2 \eta^2} \mathrm{Var}(S_{n_1}) \\
&= \frac{\sigma^2}{n_1 \eta^2} \\
&= \frac{\sigma^2}{\eta^2} \cdot \frac{1}{c_a \log s_k},
\end{aligned}$$

where $\sigma_2 = \mathrm{Var}(X_1)$.

Case $m > n_1$:

$$\begin{aligned}
\Pr\left(\bigcup_{m=n_1}^{\infty} \{\hat{\theta}^{n_1+m}(a) > U_a\} \;\middle|\; F_k\right) &\le \Pr\left(\bigcup_{m=n_1}^{\infty} \left\{\frac{\sum_{j=1}^{m}(X_{n_1+j} - \mu)}{m} > U_a - \mu\right\} \;\middle|\; F_k\right) \\
&\le \sum_{m=n_1}^{\infty} \Pr\left(\frac{\sum_{j=1}^{m}(X_{n_1+j} - \mu)}{m} > U_a - \mu \;\middle|\; F_k\right) \\
&\le \sum_{m=n_1}^{\infty} \exp(-mD) \\
&= \frac{\exp(-n_1 D)}{1 - \exp(-D)} \\
&= \frac{1}{s_k^{c_a D}(1 - \exp(-D))},
\end{aligned}$$

for a constant $D > 0$ that depends on $U_a - \mu$ and σ^2.

Therefore,

$$\begin{aligned}
&\Pr\left(\bigcup_{m=1}^{\infty}\{\hat{\theta}^{N_{s_k}(a)+m}(a) > U_a\} \,\middle|\, F_k\right) \\
\leq &\Pr\left(\bigcup_{m=1}^{n_1}\{\hat{\theta}^{N_{s_k}(a)+m}(a) > U_a\} \,\middle|\, F_k\right) + \Pr\left(\bigcup_{m=n_1}^{\infty}\{\hat{\theta}^{N_{s_k}(a)+m}(a) > U_a\} \,\middle|\, F_k\right) \\
\leq &\frac{\sigma^2}{\eta^2} \cdot \frac{1}{c_a \log s_k} + \frac{1}{s_k^{c_a D}(1-\exp(-D))} \\
= &O\left(\frac{1}{\log s_k}\right),
\end{aligned}$$

as desired. ☐

Proof of Lemma E.9. Fix a group g. At time t, if group g arrives, the PF-UCB pulls either the UCB arm or the greedy arm. The original regret analysis of KL-UCB from [28] shows that

$$\sum_{t=1}^{T} \Pr(A_t \notin \mathcal{A}_{\text{opt}}^g, A_t = A_t^{\text{UCB}}, g_t = g) = O(\log T).$$

Proposition E.2 shows that the number of times the greedy arm is pulled and incurs regret is $O(\log\log T)$. Combining, the total regret is $O(\log T)$. ☐

F Price of Fairness Proofs

F.1 Proof of Theorem 4.2

Proof. Consider the set of profiles $(s^g)_{g\in\mathcal{G}}$ that are in the feasible region of the polytope defined by the constraints of $(P(\theta))$. Refer to this polytope as the "utility set", in the language of [29]. This utility set is compact and convex, and therefore we can apply Theorem 2 of [29], which gives us the desired inequality. It is easy to see that the point in this utility set that maximizes total utility corresponds to a regret-optimal policy, and the point in the utility set that maximizes proportional fairness corresponds to PF-UCB (by definition, since PF-UCB maximizes proportional fairness within this set). ☐

F.2 Proof of Proposition 4.3

Proof. In this proof, for convenience, we use subscripts instead of superscript to refer to groups g since we do not need to refer to time steps.

Let $\{1, \ldots, M\}$ be the set of shared arms, where $\theta_1 \leq \cdots \leq \theta_M$. Let $\mathcal{G} = [G]$ be the set of groups, where $\mathsf{OPT}(1) \leq \cdots \leq \mathsf{OPT}(G)$. We assume that $\theta_M < \mathsf{OPT}(1)$. (If there is a shared arm whose reward is as large as $\mathsf{OPT}(1)$, then neither policy will incur any regret from this arm, and hence this arm is irrelevant.) In this case, all of the regret in the regret-optimal solution goes to group 1, and the other groups incur no regret. Therefore, the total utility gain of the regret-optimal solution is the sum of the regret at the disagreement point for groups 2 to G. Specifically, $\lim_{T\to\infty} \mathrm{SYSTEM}_T(\mathcal{I}) = \lim_{T\to\infty} \sum_{g=2}^{G} \frac{\tilde{R}_T^g(\pi^{\text{KL-UCB}})}{\log T}$.

We will show that for each group $g \geq 2$, the regret incurred from PF-UCB is less than half of the regret at the disagreement point — i.e. $R_T^g(\pi^{\text{PF-UCB}}, \mathcal{I}) \leq \frac{1}{2}\tilde{R}_T^g(\mathcal{I})$. Then, the utility gain for the group reduces by at most a half from the regret-optimal solution, which is our desired result.

Let $R_g = \lim_{T\to\infty} \frac{R_T^g(\pi^{\text{PF-UCB}},\mathcal{I})}{\log T}$ and $\tilde{R}_g = \lim_{T\to\infty} \frac{\tilde{R}_T^g(\mathcal{I})}{\log T}$ for all $g \in \mathcal{G}$. Recall that the proportionally fair solution comes out of the optimal solution to the following optimization problem:

$$
(P(\theta)) \qquad \begin{aligned} \max_{q\geq 0} \quad & \sum_{g\in\mathcal{G}} \log\left(\sum_{a\in\mathcal{A}^g_{\text{sub}}} \Delta^g(a)\left(J^g(a) - q^g(a)J(a)\right)\right)^+ \\ \text{s.t.} \quad & \sum_{g\in\mathcal{G}} q^g(a) = 1 \quad \forall a \in \mathcal{A}_{\text{sub}} \\ & q^g(a) = 0 \quad \forall g \in G, a \notin \mathcal{A}_{\text{sub}} \cap \mathcal{A}_g. \end{aligned}
$$

We first show a structural result of the optimal solution. Note that in terms of minimizing total regret, it is optimal for group 1 to pull all suboptimal arms. Therefore, if $q_g(a) > 0$ for some $g > 1$, we think of this as "transferring" pulls of arm a from group 1 to group g. This transfer increases the regret by a factor of $\frac{\Delta_g(a)}{\Delta_1(a)}$. We prove the following property that these transfers must satisfy:

Claim F.1 (Structure of Optimal Solution)**.** *For $g \in [M]$, let $b = \max\{a : q_g(a) > 0\}$. If $h < g$, then $q_h(a) = 0$ for all $a < b$.*

Writing out the KKT conditions of the optimization problem gives us the following result.

Claim F.2 (KKT conditions)**.** *Let $g, h \in \mathcal{G}$, $a \in \mathcal{A}$ such that $q_g(a) > 0$ and $h < g$. Then, $s_g \geq s_h \frac{\Delta_g(a)}{\Delta_h(a)}$. Moreover, if $q_1(a) > 0$, $s_g \leq \frac{\Delta_2(a)}{\Delta_1(a)} s_1$ for any $g > 1$.*

The next claim is immediate from Claim F.2.

Claim F.3. *If $h < g$ and there exists an arm a such that $q_g(a) > 0$, then $s_g \leq s_h$.*

Regret is minimized if $q_1(a) = 1$ for all a, in which case $s_1 = 0$. If $s_1 \neq 0$, then we think of this as pulls from group 1 that are re-allocated to other groups $g \neq 1$. This re-allocation increases total regret, since other groups incur more regret from pulling any arm compared to group 1.

Let $a_0 = \max\{a : q_g(a) \neq 1\}$. All pulls for any action $a > a_0$ come from group 1. We claim that $q_2(a_0) > 0$. Suppose not. Let $a' > 2$ such that $q_2(a_0) > 0$. Then, by Claim F.1, $q_2(a) = 0$ for all a. This implies that $s_2 = r_2 > r_{a'} \geq s_{a'}$, which contradicts Claim F.3. Then, by Claim F.2, $s_2 = s_1 \frac{\Delta_2(a_0)}{\Delta_1(a_0)}$.

Next, we claim that $s_2 \geq \frac{\tilde{R}_2}{2}$, which proves the desired result for $g = 2$. Note that s_1 represents the amount of regret that was "transferred" from group 1 to other groups, which increases the total regret. If *all* of this was transferred to group 2, the total regret from group 2 would be at most $s_1 \frac{\Delta_2(a_2)}{\Delta_1(a_2)} \leq s_2$. Therefore, $R_2 \leq s_2$. Since $R_2 + s_2 = \tilde{R}_2$, $s_2 \geq \frac{\tilde{R}_2}{2}$.

For $g > 2$, Claim F.2 shows $s_g \geq s_2$. Moreover, since $\mathsf{OPT}(g) \geq \mathsf{OPT}(2)$, $\tilde{R}_g \leq \tilde{R}_2$. Therefore, $s_g \geq s_2 \geq \frac{\tilde{R}_2}{2} \geq \frac{\tilde{R}_g}{2}$ as desired.

□

F.3 Proof of Claims

Proof of Claim F.1. Suppose not. Let $g \in \mathcal{G}$ and $b = \max\{a : q_g(a) > 0\}$. Let $a < b$ such that $q_h(a) > 0$. Then, since $\sum_{g'} q_{g'}(a) = 1$, $q_g(a) < 1$. By the ordering of arms and groups, we have

$$
\frac{\Delta_h(a)}{\Delta_g(a)} > \frac{\Delta_h(b)}{\Delta_g(b)}. \tag{23}
$$

We essentially show, using this inequality, that if we want to "transfer" pulls from group h to g, it is more efficient to do so using arm a rather than arm b, and hence it is a contradiction that $q_h(b)$ is positive.

We construct a "swap" that will strictly increase the objective function. Let $\varepsilon = \min\{q_h(a), q_g(b), 1 - q_g(a), 1 - q_h(b)\}$.

- Decrease $q_h(a)$ by ε, and increase $q_h(b)$ by $\frac{\Delta_h(a)J(a)}{\Delta_h(b)J(b)}\varepsilon \leq \varepsilon$, where the last inequality follows from the convexity of $\mathrm{KL}(\theta_b, \cdot)$. By construction, s_h does not change.
- Increase $q_g(a)$ by ε, and decrease $q_g(b)$ by $\frac{\Delta_h(a)J(a)}{\Delta_h(b)J(b)}\varepsilon$. The first operation decreases s_g by $\Delta_g(a)J(a)\varepsilon$, while the second operation increases s_g by $\frac{\Delta_h(a)J(a)\Delta_g(b)}{\Delta_h(b)}\varepsilon$. By (23), this strictly increases s_g overall.

This is a contradiction. □

Proof of Claim F.2. From the stationarity KKT condition, we have that

$$\frac{\Delta_g(a)J(a)}{s_g} + \lambda(a) - \mu_g(a) = 0,$$
$$\frac{\Delta_h(a)J(a)}{s_h} + \lambda(a) - \mu_h(a) = 0,$$

for some $\lambda_a \in \mathbb{R}$ and $\mu_g(a), \mu_h(a) \geq 0$. From complementary slackness, $\mu_g(a)q_g(a) = 0$. Since $q_g(a) > 0$, it must be that $\mu_g(a) = 0$. Since $\mu_h(a) \geq 0$, $\frac{\Delta_g(a)J(a)}{s_g} \leq \frac{\Delta_h(a)J(a)}{s_h}$. □

G Other Proofs

G.1 Proof that Nash Solution is Unique Under Grouped Bandit Model

The uniqueness of the Nash bargaining solution in the general bargaining problem requires that the set U is convex. In the grouped bandit model, it is not clear that the set $U(\mathcal{I}) = \{(\mathrm{UtilGain}^g(\pi, \mathcal{I}))_{g\in\mathcal{G}} : \pi \in \Psi\}$ is convex. In this section, we show that the uniqueness theorem still holds in the grouped bandit setting.

Let G be the number of groups. Let $W(u) = \sum_{g\in\mathcal{G}} \log u_g$, and let $f(U) = \mathrm{argmax}_{u\in U} W(u)$ for $U \subseteq \mathbb{R}^G$. Fix a grouped bandit instance $\mathcal{I}$, and let $u^* = f(U(\mathcal{I}))$. We first show that u^* is unique (i.e. $\mathrm{argmax}_{u\in U(\mathcal{I})} W(u)$ is unique). Suppose there was another $u' \in U(\mathcal{I})$ with the same welfare. Then, let $\bar{u} \in U(\mathcal{I})$ be the policy that runs u' with probability 50%, and u^* with probability 50%. Using the fact that $\liminf_{T\to\infty}(a_T + b_T) \geq \liminf_{T\to\infty} a_T + \liminf b_T$ implies that $\bar{u}_g \geq \frac{1}{2}(u^*_g + u'_g)$ for all g. Since $\log$ is strictly concave, $\log \bar{u}_g > \frac{1}{2}(\log u^*_g + \log u'_g)$. This implies $W(\bar{u}) > W(u^*)$, which is a contradiction.

Next, we show that f is the unique solution that satisfies the four axioms. Let $U = U(\mathcal{I})$. It is easy to see that this solution satisfies the axioms. We need to show that no other solution satisfies them. Suppose $g(\cdot)$ satisfies the axioms. We need to show $g(U) = f(U)$. Let $U' = \{(\alpha_g u_g)_{g\in\mathcal{G}} : u \in U; \alpha_g u^*_g = 1, \alpha_g > 0\}$. U' is the translated utility set so that u^* becomes the 1 vector. Then, the optimal welfare is $W(\mathbf{1}) = 0$. We need to show $g(U') = \mathbf{1}$. We claim that there is no $v \in U'$ such that $\sum_{g\in\mathcal{G}} v_g > G$. Assume that such a v exists. For $\lambda \in (0, 1)$, let t be the utilities from the policy that runs the policy induced by v with probability λ, and the policy induced by $\mathbf{1}$ with probability $1 - \lambda$. Then, by the same argument with $\liminf$ to prove uniqueness, $t_g \geq \lambda v_g + (1 - \lambda)1$. If λ is small enough, then $\sum_{g\in\mathcal{G}} \log t_g > 0$. This is a contradiction to $\mathbf{1}$ maximizing $W(\cdot)$.

Consider the symmetric set $U'' = \{u \in \mathbb{R}^G : u \geq 0, \sum_g u_g \leq G\}$. We have shown that $U' \subseteq U''$. By Pareto efficiency and symmetry, it must be that $g(U'') = \mathbf{1}$. By independence of irrelevant alternatives, $g(U') = \mathbf{1}$, and we are done.

G.2 Proof that Assumption 2.2 is Sufficient

Proposition G.1. *If an instance $\mathcal{I}$ satisfies Assumption 2.2, then there exists a consistent policy π such that $f(\pi) > -\infty$. Otherwise, $f(\pi) = -\infty$ for all $\pi \in \Psi$.*

Proof. First, suppose $\mathcal{I}$ satisfies Assumption 2.2. We need to show that there exists a consistent policy such that $f(\pi) > -\infty$. We will construct a feasible solution to the optimization problem $(P(\theta))$ with a strictly positive objective value. This will imply that the objective value Y^* is strictly larger than 0, and hence the social welfare of PF-UCB is higher than $-\infty$.

For each arm $a \in \mathcal{A}$, let $g(a) \in \Gamma(a)$. Start with $q^{g(a)}(a) = 1$ for all a and $q^g(a) = 0$ for $g \neq g(a)$. We will modify these values for suboptimal arms $\mathcal{A}_{\text{sub}}$. For arm $a \in \mathcal{A}_{\text{sub}}$, let $g'(a) \neq g(a)$ be another group with access to arm a. We will "split" the pulls of arm a between groups $g(a)$ and $g'(a)$ in a way that both groups benefit from the disagreement point. Let $p(a) \in [0, 1]$ such that $p(a)J(a) = J^{g'(a)}(a)$. Let $q^{g'(a)} = p(a)/2$ and $q^{g(a)} = 1 - p(a)/2$. Then, $J^g(a) - q^g(a)J(a) > 0$ for $g \in \{g(a), g'(a)\}$. This implies that $s^g > 0$ for all g, and therefore $Y^* > 0$. This proves the first part of the proposition.

For the second statement, suppose $\mathcal{I}$ does not satisfy Assumption 2.2. Let g' be the group that does not have a suboptimal arm that is shared with another group. First, suppose g' does not have any suboptimal arms. Then, all arms available to group g' is optimal, so group g' will incur zero regret regardless of the algorithm. Hence, the utility gain for group g' is exactly 0, and therefore $W(\pi, \mathcal{I}) = -\infty$ for any π.

Next, suppose g' does have a suboptimal arm but it is not shared. Let π be a consistent policy. Then from the following upper bound on Nash SW from Section 3.2,

$$W(\pi, \mathcal{I}) \leq \liminf_{T\to\infty} \sum_{g\in\mathcal{G}} \log \left(\sum_{a\in\mathcal{A}^g} \Delta^g(a) \left(J^g(a) - q_T^g(a, \pi)J(a)\right) \right)^+ .$$

Since g' is the only group with access to arm a for every $a \in \mathcal{A}^{g'}_{\text{sub}}$, it must be that $q_T^{g'}(a, \pi) = 1$ for every $a \in \mathcal{A}^{g'}_{\text{sub}}$. Moreover, $J^{g'}(a) = J(a)$ for every $a \in \mathcal{A}^{g'}_{\text{sub}}$. This implies that the term corresponding to g' in the sum equals $\log 0 = -\infty$. Therefore, $W(\pi, \mathcal{I}) = -\infty$ for any $\pi \in \Psi$. $\square$

G.3 Omitted Details of Theorem 3.2

We provide details on the two steps in Section 3.2 starting from (9). (4) implies that for every $\varepsilon > 0$, there exists a T_ε such that if $T \geq T_\varepsilon$, then

$$\frac{\mathbb{E}[N_T(a)]}{\log T} \geq (1 - \varepsilon)J(a).$$

Therefore, for large enough T, plugging into (9), we get

$$\frac{R_T^g(\pi, \mathcal{I})}{\log T} \geq \sum_{a\in\mathcal{A}_{\text{sub}}} \Delta^g(a)q_T^g(a, \pi)J(a)(1 - \varepsilon).$$

This implies that

$$\limsup_{T\to\infty} \frac{R_T^g(\pi, \mathcal{I})}{\log T} \geq \limsup_{T\to\infty}(1 - \varepsilon) \sum_{a\in\mathcal{A}_{\text{sub}}} \Delta^g(a)q_T^g(a, \pi)J(a).$$

Since this holds for every $\varepsilon > 0$ and the RHS is continuous in ε,

$$\limsup_{T\to\infty} \frac{R_T^g(\pi, \mathcal{I})}{\log T} \geq \limsup_{T\to\infty} \sum_{a\in\mathcal{A}_{\text{sub}}} \Delta^g(a)q_T^g(a, \pi)J(a). \tag{24}$$

Plugging in (24) into the definition of $\text{UtilGain}^g(\pi, \mathcal{I})$ gives

$$\text{UtilGain}^g(\pi, \mathcal{I}) \leq \liminf_{T\to\infty} \sum_{a\in\mathcal{A}^g_{\text{sub}}} \Delta^g(a) \left(J^g(a) - q_T^g(a, \pi)J(a)\mathbf{1}\{a \in \mathcal{A}_{\text{sub}}\}\right).$$

Using the definition of $W(\pi, \mathcal{I})$ and taking the $\liminf$ outside of the sum gives

$$W(\pi, \mathcal{I}) \leq \liminf_{T\to\infty} \sum_{g\in\mathcal{G}} \log \left(\sum_{a\in\mathcal{A}^g_{\text{sub}}} \Delta^g(a) \left(J^g(a) - q_T^g(a, \pi)J(a)\mathbf{1}\{a \in \mathcal{A}_{\text{sub}}\}\right) \right)^+ .$$

G.4 Proof of Proposition E.1

Proof. First, we prove the statement with respect to the variables $(s^g)_{g\in\mathcal{G}}$. Let $f_s(s) = \sum_{g\in\mathcal{G}} \log s^g$, and let $s^g_* = \sum_{a\in\mathcal{A}^g} \Delta^g(a) \left(J^g(a) - q^g_*(a)J(a)\right)$ and $\hat{s}^g_t = \sum_{a\in\mathcal{A}^g} \hat{\Delta}^g(a) \left(\hat{J}^g(a) - \hat{q}^g_t(a)\hat{J}(a)\right)$. Since f_s is strictly concave with respect to s, s^g_* is unique. Define the event $H_t(\delta) = \{\hat{\theta}_t(a) \in [\theta(a) - \delta, \theta(a) + \delta]$ for all $a \in \mathcal{A}\}$.

Lemma G.2. *For any $\varepsilon > 0$, there exists $\delta > 0$ such that if $H_t(\delta)$, then $\hat{s}_t^g \in [s_*^g - \varepsilon, s_*^g + \varepsilon]$ for all $g \in \mathcal{G}$.*

This shows that if $H_t(\delta)$, then the variables $\hat{s}_t^g$ are close to s_*^g for all g. Next, we need to show that the corresponding q's are also close. Let $\mathrm{proj}(z, P)$ be the projection of point z onto a polytope P.

Let $Q = \{q : \sum_{g\in G} q^g(a) = 1 \ \forall a \in \mathcal{A}_{\mathrm{sub}}, q^g(a) = 0 \ \forall g \in G, a \notin \mathcal{A}_{\mathrm{sub}}, q^g(a) \geq 0 \ \forall g \in G, a \in \mathcal{A}\}$ be the feasible space. Let $S^g(q, \tilde{\theta}) = \sum_{a\in\mathcal{A}^g} \tilde{\Delta}^g(a) \left(\tilde{J}^g(a) - q^g(a)\tilde{J}(a)\right)$, where $\tilde{\Delta}^g(a)$, $\tilde{J}^g(a)$, and $\tilde{J}(a)$ are computed with $\tilde{\theta}$.

Given $s = (s^g)_{g\in\mathcal{G}}$, let $Q(s, \tilde{\theta}) = \{q^g(a) \in Q : S^g(q, \tilde{\theta}) = s^g\}$ be the set of all feasible q's that corresponds to the solution s under the parameters $\tilde{\theta}$. Note that $Q(s, \tilde{\theta})$ is a linear polytope, and we can write it as $Q(s, \tilde{\theta}) = \{q : A(\tilde{\theta})q = b(s), q \geq 0\}$ for a matrix $A(\tilde{\theta})$ and a vector $b(s)$. We are interested in the polytopes $Q(s, \theta)$ and $Q(\hat{s}_t, \hat{\theta}_t)$, which correspond the optimal solutions of $(P(\theta))$ and $(\hat{P}_t)$ respectively. The next two lemmas state that these polytypes are close together:

Lemma G.3. *Let $\varepsilon > 0$. There exists $\delta > 0$ such that if $H_t(\delta)$, for any $\hat{q} \in Q(\hat{s}_t, \hat{\theta}_t)$, $||\mathrm{proj}(\hat{q}, Q(s, \theta)) - \hat{q}||_2 \leq \varepsilon$.*

Lemma G.4. *Let $\varepsilon > 0$. There exists $\delta > 0$ such that if $H_t(\delta)$, for any $q \in Q(s, \theta)$, $||\mathrm{proj}(q, Q(\hat{s}_t, \hat{\theta}_t)) - q||_2 \leq \varepsilon$.*

Let $q_* = \mathrm{argmin}_{q\in Q(s,\theta)} ||q||_2^2$, $\hat{q} = \mathrm{argmin}_{q\in Q(\hat{s}_t,\hat{\theta}_t)} ||q||_2^2$. Our goal is to show $||q_* - \hat{q}||_1 \leq \varepsilon$. Let $R(\eta) = \{q \in Q(s, \theta) : ||q||_2 \leq ||q_*||_2 + \eta\}$ for $\eta > 0$. Since the function $|| \cdot ||_2^2$ is strongly convex and q_* is minimizer, we have the following result:

Claim G.5. *For every $\varepsilon > 0$, there exists $\eta > 0$ such that if $q \in R(\eta)$, then $||q - q_*||_2 \leq \varepsilon$.*

First, assume $||\hat{q}_t||_2 \leq ||q_*||_2$. Let $\eta > 0$ be from Claim G.5 using $\varepsilon = \frac{\varepsilon}{2}$. Let $\delta > 0$ be from Lemma G.3 using $\varepsilon = \min\{\frac{\varepsilon}{2}, \eta\}$. Let $q' = \mathrm{proj}(\hat{q}, Q(s, \theta)) \in Q(s, \theta)$. From Lemma G.3, $||\hat{q}_t - q'||_2 \leq \eta$, implying $||q'||_2 \leq ||\hat{q}_t||_2 + \eta \leq ||q_*||_2 + \eta$. Therefore, $q' \in R(\eta)$. Claim G.5 implies $||q' - q_*|| \leq \frac{\varepsilon}{2}$. Let $\delta > 0$ correspond to $\frac{\varepsilon}{2}$ from Lemma G.3, so that $||\hat{q}_t - q'||_2 \leq \frac{\varepsilon}{2}$. Then,

$$||\hat{q}_t - q_*||_2 \leq ||\hat{q}_t - q'||_2 + ||q' - q_*||_2 \leq \varepsilon.$$

An analogous argument shows the same result in the case that $||q_*||_2 \leq ||\hat{q}_t||_2$ using Lemma G.4.

□

G.4.1 Proof of Lemmas

We first state an additional lemma:

Lemma G.6. *For any $\varepsilon > 0$ there exists a $\delta > 0$ such that if $H_t(\delta)$, then for any feasible solution q, $|f(q) - \hat{f}(q)| < \varepsilon$.*

Proof of Lemma G.6. Let q be a feasible solution. Let $S^g(q, \tilde{\theta}) = \sum_{a\in\mathcal{A}^g} \tilde{\Delta}^g(a) \left(\tilde{J}^g(a) - q^g(a)\tilde{J}(a)\right)$, where $\tilde{\Delta}^g(a)$, $\tilde{J}^g(a)$, and $\tilde{J}(a)$ are computed with $\tilde{\theta}$.

For each g, let $\varepsilon_g > 0$ be such that if $|\tilde{s}^g - s_*^g| \leq \varepsilon_g$, then $|\log s_*^g - \log \tilde{s}^g| \leq \frac{\varepsilon}{G}$. $\Delta^g(a)$, $J^g(a)$, and $J(a)$ are all differentiable functions of θ with finite derivatives around θ_*. Then, it is possible to find $\delta_g > 0$ such that if $H_t(\delta_g)$, $|\hat{\Delta}^g(a) \left(\hat{J}^g(a) - q^g(a)\hat{J}(a)\right) - \Delta^g(a) \left(J^g(a) - q^g(a)J(a)\right)| \leq \frac{\varepsilon_g}{|\mathcal{A}|}$. Summing over actions, $|S^g(q, \hat{\theta}_t) - S^g(q, \hat{\theta})| \leq \varepsilon_g$. Then, if $H_t(\delta_g)$, $|\log S^g(q, \hat{\theta}) - \log S^g(q, \theta)| \leq \frac{\varepsilon}{G}$. Take $\delta = \min_{g\in\mathcal{G}} \delta_g$. If $H_t(\delta)$ is true, $|f(q) - \hat{f}(q)| < \varepsilon$. □

Proof of Lemma G.2. Let $\varepsilon > 0$. Let $S_\varepsilon = \{s : |s^g - s_*^g| \leq \varepsilon \ \forall g\}$ be the set around s_* of interest. Our goal is to show that $f_s(\hat{s}) \in S_\varepsilon$. Let $f_{\mathrm{bd}} = \max\{f(s) : s \in \mathrm{bd}(S_\varepsilon)\} < f^*$ be the largest f on the boundary of S_ε. Then, if $f_s(s) > f_{\mathrm{bd}}$, it must be that $s \in S_\varepsilon$. (Since the entire line between s and s_* must have a value of f_s that is higher than $f_s(s)$ due to concavity, and it must cross the boundary.) Therefore, we need to show $f_s(\hat{s}_t) > f_{\mathrm{bd}}$. Let $\hat{q}_t$ be the corresponding solution to $\hat{s}_t$. Then, $f_s(\hat{s}_t) = \hat{f}_t(\hat{q}_t)$. Let $\delta > 0$ as in Lemma G.6 with $\varepsilon = f^* - f_{\mathrm{bd}}$. Then, if $H_t(\delta)$ is true,

$$f_s(\hat{s}_t) = \hat{f}_t(\hat{q}_t) \geq \hat{f}_t(q_*) \geq f(q_*) - (f^* - f_{\mathrm{bd}}) = f_{\mathrm{bd}},$$

where the second inequality follows from Lemma G.6.

□

Proof of Lemma G.3. Let $\varepsilon > 0$. Let n be the dimension of q. We will make use of the following closed form formula for the projection onto a linear subspace:

Fact G.7. Let $P = \{x : Ax = b\}$. The orthogonal projection of z onto P is $\mathrm{proj}(z, P) = z - A^\top(AA^\top)^{-1}(Az - b)$.

Let $Q = Q(s, \tilde{\theta})$, and let A, b be the corresponding parameters of the linear constraints; i.e. $Q = \{x : Ax = b, x \geq 0\}$. Similarly, let $\hat{Q} = Q(\hat{s}_t, \hat{\theta}_t)$, and let $\hat{A}, \hat{b}$ be defined similarly. Note that Fact G.7 only works with equality constraints.

We define a distance between two linear polytopes. We use the notation $P(D, f) = \{x : Dx = f\}$. Then, $Q = P(A, b)$, $\hat{Q} = P(\hat{A}, \hat{b})$.

Definition G.8. For two polytopes $P(A, b)$ and $P(A', b')$, the distance is defined as $d(P(A, b), P(A', b')) = \max\{||A - A'||_2, ||b - b'||_2\}$.

Note that for every $\alpha > 0$, there exists $\delta > 0$ such that $H_t(\delta)$ implies $d(Q, \hat{Q}) \leq \alpha$ using Lemma G.2. For any $\mathcal{I} \in 2^{[n]}$, let $P_{\mathcal{I}} = P(A_{\mathcal{I}}, b_{\mathcal{I}}) = \{x : Ax = b, x_i = 0 \; \forall i \in \mathcal{I}\}$.

Claim G.9. *There exists a constant $C \geq 1$ such that for any $\mathcal{I} \in 2^{[n]}$ and any $\tilde{A}, \tilde{b}$ of same dimensions as $A_{\mathcal{I}}, b_{\mathcal{I}}$, if $\tilde{q} \in P(\tilde{A}, \tilde{b})$ with $\tilde{q} \leq 1$ (for all elements), then $||\tilde{q} - \mathrm{proj}(\tilde{q}, P_{\mathcal{I}})||_2 \leq Cd(P_{\mathcal{I}}, P(\tilde{A}, \tilde{b}))$.*

Proof of Claim G.9. From Fact G.7, we have $||\tilde{q} - \mathrm{proj}(\tilde{q}, P_{\mathcal{I}})||_2 = ||A_{\mathcal{I}}^\top(A_{\mathcal{I}}A_{\mathcal{I}}^\top)^{-1}(A_{\mathcal{I}}\tilde{q} - b_{\mathcal{I}})||_2$. Since $\tilde{q} \in P(\tilde{A}, \tilde{b})$, $\tilde{A}\tilde{q} = \tilde{b}$. Let $\lambda = \max_{\mathcal{I}} ||A_{\mathcal{I}}^\top(A_{\mathcal{I}}A_{\mathcal{I}}^\top)^{-1}||_2$ and let $d = d(P_{\mathcal{I}}, P(\tilde{A}, \tilde{b}))$. Therefore,

$$\begin{aligned} ||\tilde{q} - \mathrm{proj}(\tilde{q}, P_{\mathcal{I}})||_2 &\leq \lambda||(A_{\mathcal{I}} - \tilde{A})\tilde{q} + (\tilde{b} - b_{\mathcal{I}})||_2 \\ &\leq \lambda\left(||A_{\mathcal{I}} - \tilde{A}||_2||\tilde{q}||_2 + ||\tilde{b} - b_{\mathcal{I}}||_2\right) \\ &\leq 2\lambda nd. \end{aligned}$$

Therefore, $C = 2\lambda n$. □

We now describe an iterative process to prove this result.

Let $Q^0 = \{q : Aq = b\}$ (Q without the non-negativity constraint), and same with $\hat{Q}^0 = \{q : \hat{A}q = \hat{b}\}$. Let $\alpha_0 = d(Q^0, \hat{Q}^0)$. Let $\tilde{q}^0 = \mathrm{proj}(\hat{q}, Q^0)$. By Claim G.9, $||\hat{q} - \tilde{q}^0||_2 \leq C\alpha_0$. If $\tilde{q}^0 \geq 0$, then STOP here.

Otherwise, find an index i which violates the non-negativity constraint using the following method:

- Let $q \in Q$ be an arbitrary feasible point ($q \geq 0$).
- From the point $\tilde{q}^0$, move along the direction towards q. Let p^0 be the first point on this line where p^0 is non-negative.
- Since Q is simply Q^0 with non-negativity constraints and both sets are convex, $p^0 \in Q$.
- Let i be an index where $\tilde{q}_i^0 < 0$ and $p_i^0 = 0$ (the last index to become non-negative).

Since $\hat{q} \geq 0$, it must be that $\hat{q}_i \leq C\alpha_0$ since $||\tilde{q}^0 - \hat{q}|| \leq C\alpha_0$.

Let Q^1 be the same polytope as Q^0, but with the additional constraint that $q_i = 0$ — call this constraint C. Let A^1, b^1 be the corresponding equality constraints for Q^1. Let $\hat{Q}^1$ be the same polytope as $\hat{Q}$, but with the additional equality constraint that $q_i = \hat{q}_i$ — call this constraint $\hat{C}$. Let $\hat{A}^1, \hat{b}^1$ be the equality constraints for $\hat{Q}^1$. Note that the only difference between constraints C and $\hat{C}$ is the right hand side, which differ by at most $C\alpha_0$. Therefore, $d(Q^1, \hat{Q}^1) \leq d(Q^0, \hat{Q}^0) + C\alpha_0 \leq 2C\alpha_0$. Clearly, $\hat{q} \in \hat{Q}^1$. Let $\tilde{q}^1 = \text{proj}(\hat{q}, Q^1)$. Applying Claim G.9 again, we have $||\hat{q} - \tilde{q}^1||_2 \leq C(2C\alpha_0) = 2C^2\alpha_0$. If $\tilde{q}^1 \geq 0$, then STOP here.

Otherwise, let j be the index which violates the non-negativity constraint found using the same method as before; except this time, we draw a line between $\tilde{q}^1$ towards $p^0 \in Q$. We let p^1 be the first point where $p^1 \geq 0$. Then, we repeat the above process. We define Q^2 to be the same polytope as Q^1, with the additional constraint that $q_j = 0$. $\hat{Q}^2$ is defined as $\hat{Q}^1$ with the additional constraint $q_j = \hat{q}_j$. Then, $\hat{q}_j \leq 2C^2\alpha_0$. Therefore, $d(Q^2, \hat{Q}^2) \leq d(Q^1, \hat{Q}^1) + 2C^2\alpha_0 \leq 2C\alpha_0 + 2C^2\alpha_0 \leq 4C^2\alpha_0$. Applying Claim G.9, we get $||\hat{q} - \tilde{q}^2||_2 \leq C(4C^2\alpha_0) = 4C^3\alpha_0$. If $\tilde{q}^2 \geq 0$, then STOP here.

After stopping: If this process stopped at iteration m, then $\tilde{q}^m \in Q$ and $||\hat{q} - \tilde{q}^m||_2 \leq 2^m C^{m-1}\alpha_0$. It must be that $m \leq n$. If $\alpha_0 = \frac{\varepsilon}{2^n C^{n-1}}$, then $||\hat{q} - \tilde{q}^m||_2 \leq \varepsilon$. Then, $||\text{proj}(\hat{q}, Q) - \hat{q}||_2 \leq \varepsilon$. Let $\delta > 0$ such that $H_t(\delta)$ implies $d(Q, \hat{Q}) \leq \alpha_0$. □

Proof of Lemma G.4. This proof follows essentially the same steps as the proof of Lemma G.3 by swapping Q and $\hat{Q}$. The main difference is that we are projecting q onto $Q(\hat{s}_t, \hat{\theta}_t)$, but this must hold for all possible values of $\hat{s}_t, \hat{\theta}_t$ (using a single δ). Due to this, the only thing we have to change from the proof of Lemma G.3 is Claim G.9. We must show that there exists a constant C where Claim G.9 is satisfied for all possible values of $\hat{s}_t, \hat{\theta}_t$. The only place where C relies on a property of the polytope $P_{\mathcal{I}}$ is in choosing λ. Therefore our goal is to uniformly upper bound $\max_{\mathcal{I}} ||\hat{A}_{\mathcal{I}}^\top (\hat{A}_{\mathcal{I}} \hat{A}_{\mathcal{I}}^\top)^{-1}||_2$ for all possible $\hat{A}_{\mathcal{I}}$ that can be induced by all possible $\hat{s}_t, \hat{\theta}_t$.

Note that since we assume that $H_t(\delta_0)$ holds, the possible matrices $\hat{A}$ lie in a compact space (since every element of the matrix $\hat{A}$ can be at most δ_0 apart). Since $||A^\top (AA^\top)^{-1}||_2$ is a continuous function of the elements of the matrix A, $\lambda_1 = \max_{\hat{A}} ||\hat{A}^\top (\hat{A}\hat{A}^\top)^{-1}||_2$ exists. Moreoever, for any $\mathcal{I}$, $||\hat{A}_{\mathcal{I}}^\top (\hat{A}_{\mathcal{I}} \hat{A}_{\mathcal{I}}^\top)^{-1}||_2 \leq C(n) ||\hat{A}^\top (\hat{A}\hat{A}^\top)^{-1}||_2$ for a constant $C(n)$. Therefore, by replacing λ with $\lambda_1 C(n)$, Claim G.9 holds. □