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:

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

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: $\text{argmax}_{a \in \mathcal{A}(m_t)} \langle a, \hat{\theta}_t \rangle$.

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 $A_t = \text{argmax}_{a \in \mathcal{A}(m_t)} \langle a, \hat{\theta}_t \rangle + \sqrt{f_{T,1/s(t)^2}} ||a||_{G_t^{-1}}$.
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 $A_t = \text{argmin}_{a \in cA^{m_t}} N_t(a)$.
3. If the above two criteria are not true, then pull $A_t = \text{argmin}_{a \in \mathcal{A}^{m_t}} \frac{N_t(a)}{\min(\hat{Q}_t(m_t,a), f_T/(\hat{\Delta}_t^{\min})^2)}$.

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

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 $F_{1,n}, \ldots, F_{T,n}$ are disjoint. Then, by the law of total probability, $\Pr(\hat{\theta}^n(a) \notin [L, U]) \geq \sum_{t=1}^{T} \Pr(\hat{\theta}^n \notin [L, U] | F_{t,n}) \Pr(F_{t,n})$. Therefore,

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

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,*

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

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

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

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

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*

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

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 $a \notin \mathcal{A}_{\mathrm{sub}}$, 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

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

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

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

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

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

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.

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

We are done since $K^g_\varepsilon(T) = O(\log T)$.

D.2 Proof of Proposition C.2

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

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

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

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

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

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

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

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

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

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

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

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.*

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

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 $a \in \mathcal{A}_{\text{sub}}^g$,*

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

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, $\sum_{t=1}^{T} \mathbf{1}(\text{Pull}_t^g(a), a \in$

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

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

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

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

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 $\limsup_{T\to\infty} \sum_{g\neq g'} \frac{\mathbb{E}[N_T^{g'}(a)]}{\log T} \geq (1 - q^{g'}(a))J(a) + \alpha$, 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 $\sum_{t=1}^T \Pr(\mathrm{Pull}_t^g(a), a = A_t^{\text{greedy}}(g))$. 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 $\hat{\theta}_t(a') \leq \theta(a) + \delta$ for all $a' \in \mathcal{A}^g$.

Define the event

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

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

For R_t to occur, $\hat{\theta}_t(a') \leq \theta(a) + \delta$ (since a is the greedy arm) and $\mathrm{UCB}_t(a') \geq \mathsf{OPT}(g)$ (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, $\sum_{t=1}^{T} \Pr(R_t) = O(\log \log T)$.

□

Proof of Proposition E.3. Let $H_t = H_t(\delta)$. Fix a group g and an arm $a \in \mathcal{A}^g_{\mathrm{sub}}$. 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,* $\Pr\left(\hat{\theta}_t(a) \notin [\theta(a) - \delta, \theta(a) + \delta]\, \textit{for some}\ t \in \mathcal{T}_k \mid F_k\right) \leq O\left(\frac{1}{\log s_k}\right)$.

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.

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

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

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

We claim that there exists a $\lambda \in (0,1)$ such that $\Pr(Y_m \mid Y_{m-1}, \bar{G}_k) \leq \lambda$. 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 $N_\tau^g(a) \leq \hat{q}_t^g(a) N_\tau(a)$. 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 $\Pr(Y_m \mid Y_{m-1}, \bar{G}_k) \leq \lambda$.

Therefore,

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

Substituting back into (22) gives

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

□

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: $N_t(j) \geq N_t(a)$. We show that this event cannot happen often. Let $E_t = \{\text{Pull}_t(a), A_t^{\text{greedy}}(g) \neq a, a \in \mathcal{A}_t^{\text{UCB}}, a = A_t^{\text{UCB}}(g), \hat{\theta}_t(a) > \mathsf{OPT}(g) - \delta\}$. For any $j \in \mathcal{A}_{\text{sub}}^g$,

$$\begin{aligned}&\sum_{t=1}^{T} \mathbf{1}(E_t, N_t(j) \geq N_t(a), \hat{\theta}_t(j) > \mathsf{OPT}(g) - \delta) \\ &\leq \sum_{t=1}^{T} \sum_{n=1}^{t} \sum_{n_j=n}^{t} \mathbf{1}(E_t, \hat{\theta}_{n_j}(j) > \mathsf{OPT}(g) - \delta, N_t(j) = n_j, N_t(a) = n) \\ &\leq \sum_{n_j=1}^{T} \mathbf{1}(\hat{\theta}_{n_j}(j) > \mathsf{OPT}(g) - \delta) \sum_{n=1}^{n_j} \sum_{t=n}^{T} \mathbf{1}(E_t, N_t(a) = n) \\ &\leq \sum_{n_j=1}^{T} \mathbf{1}(\hat{\theta}_{n_j}(j) > \mathsf{OPT}(g) - \delta) n_j,\end{aligned}$$

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 $j \in \mathcal{A}^g_{\text{sub}}$ 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.

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

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

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

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

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

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

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

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

Case $m > n_1$:

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

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

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

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 $\{1,\dots,M\}$ 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, $\lim_{T\to\infty}\text{SYSTEM}_T(\mathcal{I}) = \lim_{T\to\infty}\sum_{g=2}^{G}\frac{\tilde{R}_T^g(\pi^{\text{KL-UCB}})}{\log T}$.

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:

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

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,

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

Since g' is the only group with access to arm a for every $a \in \mathcal{A}^{g'}_{\text{sub}}$, it must be that $q^{g'}_T(a, \pi) = 1$ 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

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

This implies that

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

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

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

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

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

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

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

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 $f_s(\hat{s}_t) > f_{\text{bd}}$. 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 $\varepsilon = f^* - f_{\text{bd}}$. Then, if $H_t(\delta)$ is true,

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

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 $\text{proj}(z, P) = z - A^\top(AA^\top)^{-1}(Az - b)$.

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 $||\tilde{q} - \text{proj}(\tilde{q}, P_{\mathcal{I}})||_2 \leq Cd(P_{\mathcal{I}}, P(\tilde{A}, \tilde{b}))$.*

Proof of Claim G.9. From Fact G.7, we have $||\tilde{q} - \text{proj}(\tilde{q}, P_{\mathcal{I}})||_2 = ||A_{\mathcal{I}}^\top(A_{\mathcal{I}}A_{\mathcal{I}}^\top)^{-1}(A_{\mathcal{I}}\tilde{q} - b_{\mathcal{I}})||_2$. 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,

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

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 $\tilde{q}^0 = \text{proj}(\hat{q}, Q^0)$. 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. □