Supplementary Material
to accompany
Top-k eXtreme *Contextual Bandits with Arm Hierarchy*

A. Extended Related Work

The relevant prior work can be broadly classified under the following three categories:

General Contextual Bandits: The general contextual bandit problem has been studied for more than two decades. In the agnostic setting where the mean reward of the arms given a context is not fully captured by the function class $\mathcal{F}$, the problem was studied in the adversarial setting leading to the well-known EXP-4 class of algorithms (Auer et al., 2002; McMahan & Streeter, 2009; Beygelzimer et al., 2011). These algorithms can achieve the optimal $\tilde{O}(\sqrt{AT\log(T|\mathcal{F}|)})$ regret bound but the computational cost per time-step can be $O(|\mathcal{F}|)$. This paved the way for oracle-based contextual bandit algorithms in the stochastic setting (Agarwal et al., 2014; Langford & Zhang, 2007). The algorithm in (Agarwal et al., 2014) can achieve optimal regret bounds while making only $\tilde{O}(\sqrt{AT})$ calls to a cost-sensitive classification oracle, however the algorithm and the oracle are not easy to implement in practice. In more recent work, it has been shown that algorithms that use regression oracles work better in practice (Foster et al., 2018). In this paper we will be focused on the realizable (or near-realizable) setting, where there exists a function in the function class, which can model the expected reward of arms given context. This setting has been studied with great practical success under specific instances of the function classes, such as linear. Most of the successful approaches are based on Upper Confidence Bound strategies or Thompson Sampling (Filippi et al., 2010; Chu et al., 2011; Krause & Ong, 2011; Agrawal & Goyal, 2013), both of which lead to algorithms which are heavily tailored to the specific function class. The general realizable case was modeled in (Agarwal et al., 2012) and recently there has been exciting progress in this direction. The authors in (Foster & Rakhlin, 2020) identified that a particular exploration scheme that dates back to (Abe & Long, 1999) can lead to a simple algorithm that reduces the contextual bandit problem to online regression and can achieve optimal regret guarantees. The same idea was extended for the stochastic realizable contextual bandit problem with an offline batch regression oracle (Simchi-Levi & Xu, 2020; Foster et al., 2020b). We build on the techniques introduced in these works. However all the literature discussed so far only address the problem of selecting one arm per time-step, while we are interested in selection the top-k arms at each time step.

Exploration in Combinatorial Action Spaces: In (Qin et al., 2014) authors study the k-arm selection problem in contextual bandits where the function class is linear and the utility of a set of arms chosen is a set function with some monotonicity and Lipschitz continuity properties. In (Yue & Guestrin, 2011) the authors study the problem of retrieving k-arms in contextual bandits in the context of a linear function class and the assumption that the utility of a set of arms is sub-modular. Both these approaches do not extend to general function classes and are not applicable to the extreme setting. In the context of off-policy learning from logged data there are several works that address the top-k arms selection problem under the context of slate recommendations (Swaminathan et al., 2017; Narita et al., 2019). We will now review the combinatorial action space literature in multi-armed bandit (MAB) problems. Most of the work in this space deals with semi-bandit feedback (Chen et al., 2016; Combes et al., 2015; Kveton et al., 2015; Merlis & Mannor, 2019). This is also our feedback model, but we work in a contextual setting. There is also work in the full-bandit feedback setting, where one gets to observe only one representative reward for the whole set of arms chosen. This body of literature can be divided into the adversarial setting (Merlis & Mannor, 2019; Cesa-Bianchi & Lugosi, 2012) and the stochastic setting (Dani et al., 2008; Agarwal & Aggarwal, 2018; Lin et al., 2014; Rejwan & Mansour, 2020).

Learning in eXtreme Output Spaces: The problem of learning from logged bandit feedback when the number of arms is extreme was studied recently in (Lopez et al., 2020). In (Majzoubi et al., 2020) the authors address the contextual bandit problem for continuous action spaces by using a cost sensitive classification oracle for large number of classes, which is itself implemented as a hierarchical tree of binary classifiers. In the context of supervised learning the problem of learning under large but correlated output spaces has been studied under the banner of eXtreme Multi-Label Classification/Ranking (XMC/ XMR) (see (Bhatia et al., 2016) and references). Tree based methods for XMR have been extremely successful (Jasinska et al., 2016; Prabhu et al., 2018;

Khandagale et al., 2020; Wydmuch et al., 2018; You et al., 2019; Yu et al., 2020). In particular our assumptions about arm hierarchy and the implementation of our algorithms have been motivated by (Prabhu et al., 2018; Yu et al., 2020).

B. Continuation of the Motivating Example

We introduce a hierarchical decomposition $\mathcal{T}$ for $\mathcal{A}$, which in this case is a balanced 2^d-ary tree. At the leaf level, each tree node has a maximum number of m arms from the extreme arm space $\mathcal{A}$. The height of such a tree is $H \approx \lceil \log \lceil A/m \rceil \rceil$ under some mild assumptions on the distributions of the arms in $\mathcal{A}$. For a specific depth h, we use $e_{h,i}$ to denote a node with index i at depth h and $\mathcal{C}_{h,i}$ to denote the 2^d children of $e_{h,i}$ at depth $h+1$. Each node $e_{h,i}$ of the tree is further equipped with a *routing function* $g_{h,i}(x) = \frac{\text{rad}_{h,i}}{\|x - \text{ctr}_{h,i}\|}$, where $\text{ctr}_{h,i}$ is the center of the node $e_{h,i}$ and $\text{rad}_{h,i}$ is the radius of the smallest ball at $\text{ctr}_{h,i}$ that contains $e_{h,i}$. The center $\text{ctr}_{h,i}$ serves as a representative for the set of arms in $e_{h,i}$. Figure 1 (left) illustrates the hierarchical decomposition for the 1D case.

Given a context x, we perform an *adaptive search* through this hierarchical decomposition $\mathcal{T}$, parameterized by a constant $\beta \in (0,1)$. Initially, the sets I_x and S_x are set to be empty and the search starts from the root of the tree. When a node $e_{h,i}$ is visited, it is considered *far from* x if $g_{h,i}(x) = \frac{\text{rad}_{h,i}}{\|x - \text{ctr}_{h,i}\|} \leq \beta$ and *close to* x if $g_{h,i}(x) = \frac{\text{rad}_{h,i}}{\|x - \text{ctr}_{h,i}\|} > \beta$. If $e_{h,i}$ is far from x, we simply place it in I_x. If $e_{h,i}$ close to x, we visit its children in $\mathcal{C}_{h,i}$ recursively if $e_{h,i}$ is an internal node or place it in S_x if it is a leaf. At the end of the search, I_x consists of a list of nodes and S_x is a list of *singleton* arms.

We claim that the union of the singleton arms in S_x and the nodes in I_x form an x-dependent decomposition $\mathcal{A}_x$. First, the disjoint union of I_x and S_x covers the whole arm space $\mathcal{A}$. S_x contains only $O(1)$ singleton arms with arm features close to the context feature x while the size of I_x is bounded by $O(\log A)$ as there are at most $O(1)$ nodes $e_{h,i}$ inserted into I_x at each of the $O(\log A)$ levels. Hence, the sum of the cardinalities of S_x and I_x is bounded by $Z = O(\log A)$, i.e., logarithmic in the size A of the extreme arm space $\mathcal{A}$.

Second, for any two original arms a_1, a_2 corresponding to a node $e_{h,i} \in I_x$,

$$|r(x,a_1) - r(x,a_2)| \leq \|\partial_l r(x,a')\| \cdot \|a_1 - a_2\| \leq \frac{\eta}{\|x - a'\|} \cdot (2 \text{ rad}_{h,i}),$$

where a' lies on the segment between a_1 and a_2. Since

$$\|x - a'\| \geq \|x - \text{ctr}_{h,i}\| - \|a' - \text{ctr}_{h,i}\| \geq (1/\beta - 1)\text{rad}_{h,i}$$

holds for $e_{h,i} \in I_x$,

$$|r(x,a_1) - r(x,a_2)| \leq \frac{\eta}{(1/\beta - 1)\text{rad}_{h,i}} \cdot (2 \text{ rad}_{h,i}) = \frac{2\eta\beta}{1-\beta}.$$

Hence, if one chooses β so that $2\eta\beta/(1-\beta) \leq \epsilon$, then $|r(x,a_1) - r(x,a_2)| \leq \epsilon$ for any two arms a_1, a_2 in any $e_{h,i} \in I_x$.

Therefore for each x, the union of the singleton arms in S_x and the nodes in I_x form an x-dependent decomposition of $\mathcal{A}$ that satisfies the conditions (2) and (3). Figure 1 (middle) shows the decomposition for a given context x, while Figure 1 (right) shows how the decomposition varies with the context x. In what follows, we shall refer to the members of I_x *node effective arms* and the ones of S_x *singleton effective arms*.

C. Top-k Analysis

Notation: Let l denote epoch index with n_l time steps. Define $N_l = \sum_{i=1}^{l} n_i$. At the beginning of each epoch l, we compute $\widehat{y}_l(x,a)$ as regression with respect to past data,

$$\widehat{y}_l = \underset{f \in \mathcal{F}}{\text{argmin}} \sum_{t=1}^{N_{l-1}} \sum_{a \in \Phi_t} (f(x_t,a) - r_t(a))^2,$$

where Φ_t is the subset for which the learner receives feedback.

Let $\{\phi_l\}_{l\geq 2}$ be a sequence of numbers. The analysis in this section will be carried out under the event

$$\mathcal{E} = \left\{ l \geq 2 : \frac{2}{N_{l-1}} \sum_{s=1}^{N_{l-1}} \mathbb{E}_{x_s, \mathcal{A}_s} \left\{ \frac{1}{k} \sum_{a \in \mathcal{A}_s} (\widehat{y}_l(x_s, a) - f^*(x_s, a))^2 | \mathcal{H}_{s-1} \right\} \leq \phi_l^2 \right\} \tag{4}$$

Lemmas 5 and 7 compute ϕ_l for finite class $\mathcal{F}$, such that event $\mathcal{E}$ holds with high probability.

We define $\gamma_l = \sqrt{A-k+1}/(32\phi_l)$, the scaling parameter used by Algorithm 1. In this paper, we analyze Algorithm 1 with $r = 1$, i.e. our procedure deterministically selects top $k-1$ actions of $\widehat{y}_l$ and selects the remaining action according to Inverse Gap Weighting on the remaining coordinates.

A deterministic strategy α is a map $\alpha : \mathcal{X} \to \mathcal{A}$. Throughout the proofs, we employ the following shorthand to simplify the presentation. We shall write $\widehat{y}_i(x, \alpha)$ and $f^*(x, \alpha)$ in place of $\widehat{y}_i(x, \alpha(x))$ and $f^*(x, \alpha(x))$. We reserve the letter α for a strategy and a for an action.

Given x, we let $\widehat{\alpha}_l^j(x)$ be the j-th highest action according to $\widehat{y}_l(x, \cdot)$. Similarly, $\alpha^{*,j}(x)$ is the j-th highest action according to $f^*(x, \cdot)$. We say that the set of strategies $\alpha^1, \ldots, \alpha^k$ is non-overlapping if for any x the set $\{\alpha^1(x), \ldots, \alpha^k(x)\}$ is a set of distinct actions. Let $e(s)$ denote the epoch corresponding to time step s.

Our argument is based on the beautiful observation of (Simchi-Levi & Xu, 2020) that one can analyze IGW inductively, by controlling the differences between estimated gaps (to the best estimated action) and the true gaps (to the best true action in the given context), with a *mismatched factor of* 2. We extend this technique to top-k selection, which introduces a number of additional difficulties in the analysis.

Induction hypothesis (l): For any epoch $i < l$, and all non-overlapping strategies $\alpha^1, \ldots, \alpha^k \in \mathcal{A}^{\mathcal{X}}$,

$$\mathbb{E}_x \left\{ \sum_{j=1}^k [\widehat{y}_i(x, \widehat{\alpha}_i^j) - \widehat{y}_i(x, \alpha^j)] - 2 \sum_{j=1}^k [f^*(x, \alpha^{*,j}) - f^*(x, \alpha^j)] \right\} \leq \frac{k(A-k+1)}{\gamma_i}$$

and

$$\mathbb{E}_x \left\{ \sum_{j=1}^k [f^*(x, \alpha^{*,j}) - f^*(x, \alpha^j)] - 2 \sum_{j=1}^k [\widehat{y}_i(x, \widehat{\alpha}_i^j) - \widehat{y}_i(x, \alpha^j)] \right\} \leq \frac{k(A-k+1)}{\gamma_i}.$$

Lemma 1. *Suppose event (4) holds. For all non-overlapping strategies $\alpha^1, \ldots, \alpha^k$,*

$$\mathbb{E}_x \frac{1}{k} \sum_{j=1}^k |\widehat{y}_l(x, \alpha^j) - f^*(x, \alpha^j)| \leq \phi_l \cdot \left((A-k+1) + \sum_{i=1}^{l-1} \frac{n_i}{N_{l-1}} \gamma_i \mathbb{E}_x \frac{1}{k} \sum_{j=1}^k \left[\widehat{y}_i(x, \widehat{\alpha}_i^j) - \widehat{y}_i(x, \alpha^j) \right] \right)^{1/2}$$

Hence, by the induction hypothesis (l),

$$\mathbb{E}_x \frac{1}{k} \sum_{j=1}^k |\widehat{y}_l(x, \alpha^j) - f^*(x, \alpha^j)| \leq \sqrt{2}\phi_l \cdot \left((A-k+1) + \gamma_l \mathbb{E}_x \frac{1}{k} \sum_{j=1}^k \left[f^*(x, \alpha^{*,j}) - f^*(x, \alpha^j) \right] \right)^{1/2}$$

assuming γ_i are non-decreasing.

Proof. Given x, let $T_x(\widehat{y}_i) \subset [A]$ denote the indices of top $k-1$ actions according to $\widehat{y}_i(x, \cdot)$. Let $p_i(\cdot|x)$ denote the IGW distribution on epoch i, with support on the remaining $A-k+1$ actions. On round s in epoch $e(s)$, given x_s, Algorithm 1 with $r = 1$ chooses $\mathcal{A}_s$ by selecting $T_{x_s}(\widehat{y}_{e(s)})$ determistically and selecting the last action according to $p_{e(s)}(\cdot|x_s)$. We write $p_{e(s)}(\alpha|x_s)$ as a shorthand for $p_{e(s)}(\alpha(x_s)|x_s)$.

For non-overlapping strategies $\alpha^1, \ldots, \alpha^k$,

$$\mathbb{E}_x \frac{1}{k} \sum_{j=1}^k |\widehat{y}_l(x, \alpha^j) - f^*(x, \alpha^j)| = \frac{1}{N_{l-1}} \sum_{s=1}^{N_{l-1}} \mathbb{E}_{x_s} \left\{ \frac{1}{k} \sum_{j=1}^k |\widehat{y}_l(x_s, \alpha^j) - f^*(x_s, \alpha^j)| \right\}.$$

This sum can be written as

$$\frac{1}{N_{l-1}}\sum_{s=1}^{N_{l-1}}\mathbb{E}_{x_s}\Bigg[\frac{1}{k}\sum_{j=1}^{k}\big|\widehat{y}_l(x_s,\alpha^j)-f^*(x_s,\alpha^j)\big|\cdot\mathbf{1}\big\{\alpha^j(x_s)\in T_{x_s}(\widehat{y}_{e(s)})\big\}$$
$$+\frac{1}{k}\sum_{j=1}^{k}|\widehat{y}_l(x_s,\alpha^j)-f^*(x_s,\alpha^j)|\sqrt{p_{e(s)}(\alpha^j|x_s)}\frac{1}{\sqrt{p_{e(s)}(\alpha^j|x_s)}}\cdot\mathbf{1}\big\{\alpha^j(x_s)\notin T_{x_s}(\widehat{y}_{e(s)})\big\}\Bigg].$$

By the Cauchy-Schwartz inequality, the last expression is upper-bounded by

$$\Bigg(\frac{1}{N_{l-1}}\sum_{s=1}^{N_{l-1}}\mathbb{E}_{x_s}\frac{1}{k}\sum_{j=1}^{k}|f^*(x_s,\alpha^j)-\widehat{y}_l(x_s,\alpha^j)|^2\mathbf{1}\big\{\alpha^j(x_s)\in T_{x_s}(\widehat{y}_{e(s)})\big\}\Bigg)^{1/2}$$
$$+\Bigg(\frac{1}{N_{l-1}}\sum_{s=1}^{N_{l-1}}\mathbb{E}_{x_s}\frac{1}{k}\sum_{j=1}^{k}|f^*(x_s,\alpha^j)-\widehat{y}_l(x_s,\alpha^j)|^2p_{e(s)}(\alpha^j|x_s)\mathbf{1}\big\{\alpha^j(x_s)\notin T_{x_s}(\widehat{y}_{e(s)})\big\}\Bigg)^{1/2}$$
$$\times\Bigg(\frac{1}{N_{l-1}}\sum_{s=1}^{N_{l-1}}\mathbb{E}_{x_s}\frac{1}{k}\sum_{j=1}^{k}\frac{1}{p_{e(s)}(\alpha^j|x_s)}\mathbf{1}\big\{\alpha^j(x_s)\notin T_{x_s}(\widehat{y}_{e(s)})\big\}\Bigg)^{1/2}$$
$$\le\Bigg(\frac{1}{N_{l-1}}\sum_{s=1}^{N_{l-1}}\mathbb{E}_{x_s}\frac{1}{k}\sum_{a\in T_{x_s}(\widehat{y}_{e(s)})}|f^*(x_s,a)-\widehat{y}_l(x_s,a)|^2\Bigg)^{1/2}$$
$$+\Bigg(\frac{1}{N_{l-1}}\sum_{s=1}^{N_{l-1}}\frac{1}{k}\mathbb{E}_{x_s,a\sim p_{e(s)}(\cdot|x_s)}|f^*(x_s,a)-\widehat{y}_l(x_s,a)|^2\Bigg)^{1/2}$$
$$\times\Bigg(\sum_{i=1}^{l-1}\frac{n_i}{N_{l-1}}\mathbb{E}_x\frac{1}{k}\sum_{j=1}^{k}\frac{1}{p_i(\alpha^j|x)}\mathbf{1}\big\{\alpha^j(x)\notin T_x(\widehat{y}_i)\big\}\Bigg)^{1/2}.$$

We further upper bound the above by

$$\Bigg\{\Bigg(\frac{1}{N_{l-1}}\sum_{s=1}^{N_{l-1}}\mathbb{E}_{x_s}\frac{1}{k}\sum_{a\in T_{x_s}(\widehat{y}_{e(s)})}|f^*(x_s,a)-\widehat{y}_l(x_s,a)|^2\Bigg)^{1/2}$$
$$+\Bigg(\frac{1}{N_{l-1}}\sum_{s=1}^{N_{l-1}}\frac{1}{k}\mathbb{E}_{x_s,a\sim p_{e(s)}(\cdot|x_s)}|f^*(x_s,a)-\widehat{y}_k(x_s,a)|^2\Bigg)^{1/2}\Bigg\}$$
$$\times\Bigg(1\vee\sum_{i=1}^{l-1}\frac{n_i}{N_{l-1}}\mathbb{E}_x\frac{1}{k}\sum_{j=1}^{k}\frac{1}{p_i(\alpha^j|x)}\mathbf{1}\big\{\alpha^j\notin T_x(\widehat{y}_i)\big\}\Bigg)^{1/2}$$
$$\le\Bigg(\frac{2}{N_{l-1}}\sum_{s=1}^{N_{l-1}}\frac{1}{k}\mathbb{E}_{x_s}\Bigg[\sum_{a\in T_{x_s}(\widehat{y}_{e(s)})}|f^*(x_s,a)-\widehat{y}_l(x_s,a)|^2+\mathbb{E}_{a\sim p_{e(s)}(\cdot|x_s)}|f^*(x_s,a)-\widehat{y}_l(x_s,a)|^2\Bigg]\Bigg)^{1/2} \tag{5}$$
$$\times\Bigg(1\vee\sum_{i=1}^{l-1}\frac{n_i}{N_{l-1}}\mathbb{E}_x\frac{1}{k}\sum_{j=1}^{k}\frac{1}{p_i(\alpha^j|x)}\mathbf{1}\big\{a^j(x)\notin T_x(\widehat{y}_i)\big\}\Bigg)^{1/2} \tag{6}$$

where we use $(\sqrt{a}+\sqrt{b})^2 \le 2(a+b)$ for nonnegative a, b. Now, observe that

$$\mathbb{E}_{x_s}\left[\sum_{a\in T_{x_s}(\widehat{y}_{e(s)})} |f^*(x_s,a)-\widehat{y}_l(x_s,a)|^2 + \mathbb{E}_{a\sim p_{e(s)}(\cdot|x_s)}|f^*(x_s,a)-\widehat{y}_l(x_s,a)|^2\right] \tag{7}$$

$$= \mathbb{E}_{x_s,\mathcal{A}_s}\left\{\sum_{a\in\mathcal{A}_s}(\widehat{y}_l(x_s,a)-f^*(x_s,a))^2|\mathcal{H}_{s-1}\right\} \tag{8}$$

by the definition of the selected set $\mathcal{A}_s$ in Algorithm 1 with $r=1$. Under the event (4), the expression in (5) is at most ϕ_l. We now turn to the expression in (6). Note that by definition, for any strategy α^j

$$\begin{aligned}\frac{1}{p_i(\alpha^j|x)}\mathbf{1}\{a^j(x)\notin T_x(\widehat{y}_i)\} &= \left[(A-k+1)+\gamma_i(\widehat{y}_i(x,\widehat{\alpha}_i^k)-\widehat{y}_i(x,\alpha^j))\right]\mathbf{1}\{\alpha^j(x)\notin T_x(\widehat{y}_i)\}\\ &\le (A-k+1)+\gamma_i\left[\widehat{y}_i(x,\widehat{\alpha}_i^k)-\widehat{y}_i(x,\alpha^j)\right]_+,\end{aligned}$$

where $[a]_+=\max\{a,0\}$. Therefore, by Lemma 3, for any non-overlapping strategies $\alpha^1,\dots,\alpha^k$,

$$\begin{aligned}\frac{1}{k}\sum_{j=1}^k\frac{1}{p_i(\alpha^j|x)}\mathbf{1}\{\alpha^j(x)\notin T_x(\widehat{y}_i)\} &\le (A-k+1)+\frac{1}{k}\sum_{j=1}^k\gamma_i\left[\widehat{y}_i(x,\widehat{\alpha}_i^k)-\widehat{y}_i(x,\alpha^j)\right]_+\\ &\le (A-k+1)+\frac{1}{k}\sum_{j=1}^k\gamma_i\left[\widehat{y}_i(x,\widehat{\alpha}_i^j)-\widehat{y}_i(x,\alpha^j)\right].\end{aligned}$$

Since the above expression is at least $(A-k+1)\ge 1$, we may drop the maximum with 1 in (6). Putting everything together,

$$\mathbb{E}_x\frac{1}{k}\sum_{j=1}^k|\widehat{y}_l(x,\alpha^j)-f^*(x,\alpha^j)| \le \phi_l\cdot\left((A-k+1)+\sum_{i=1}^{l-1}\frac{n_i}{N_{l-1}}\gamma_i\mathbb{E}_x\frac{1}{k}\sum_{j=1}^k\left[\widehat{y}_i(x,\widehat{\alpha}_i^j)-\widehat{y}_i(x,\alpha^j)\right]\right)^{1/2}$$

To prove the second statement, by induction we upper bound the above expression by

$$\begin{aligned}&\phi_l\cdot\left((A-k+1)+\max_{i<l}\gamma_i\left\{2\mathbb{E}_x\frac{1}{k}\sum_{j=1}^k\left[f^*(x,\alpha^{*,j})-f^*(x,\alpha^j)\right]+\frac{A}{\gamma_i}\right\}\right)^{1/2}\\ &\le \phi_l\cdot\left(2(A-k+1)+2\gamma_l\mathbb{E}_x\frac{1}{k}\sum_{j=1}^k\left[f^*(x,\alpha^{*,j})-f^*(x,\alpha^j)\right]\right)^{1/2}.\end{aligned}$$

□

We now prove that inductive hypothesis holds for each epoch l.

Lemma 2. *Suppose we set $\gamma_l=\sqrt{A-k+1}/(32\phi_l)$ for each l, and that event $\mathcal{E}$ in (4) holds. Then the induction hypothesis holds for each $l\ge 2$.*

Proof. The base of the induction ($l=2$) is satisfied trivially if $\gamma_2=O(1)$ since functions are bounded. Now suppose the induction hypothesis (l) holds for some $l\ge 2$. We shall prove it for $(l+1)$.

Denote by $\boldsymbol{\alpha}=(\alpha^1,\dots,\alpha^k)$ any set of non-overlapping strategies. We also use the shorthand $A'=A-k+1$ for the size of the support of the IGW distribution. Define

$$\mathsf{R}(\boldsymbol{\alpha})=\mathbb{E}_x\frac{1}{k}\sum_{j=1}^k\left[f^*(x,\alpha^{*,j})-f^*(x,\alpha^j)\right],\quad \widehat{\mathsf{R}}_l(\boldsymbol{\alpha})=\mathbb{E}_x\frac{1}{k}\sum_{j=1}^k\left[\widehat{y}_l(x,\widehat{\alpha}_l^j)-\widehat{y}_k(x,\alpha^j)\right].$$

Since

$$\big[f^*(x,\alpha^{*,j}) - f^*(x,a)\big] = \big[\widehat{y}_l(x,\alpha^{*,j}) - \widehat{y}_l(x,a)\big] + \big[f^*(x,\alpha^{*,j}) - \widehat{y}_l(x,\alpha^{*,j})\big] + [\widehat{y}_l(x,a) - f^*(x,a)],$$

it holds that

$$\begin{aligned}
&\sum_{j=1}^{k}\big[f^*(x,\alpha^{*,j}) - f^*(x,\alpha^j)\big] \\
&= \sum_{j=1}^{k}\big[\widehat{y}_l(x,\alpha^{*,j}) - \widehat{y}_l(x,\alpha^j)\big] + \sum_{j=1}^{k}\big[f^*(x,\alpha^{*,j}) - \widehat{y}_l(x,\alpha^{*,j})\big] + \sum_{j=1}^{k}\big[\widehat{y}_l(x,\alpha^j) - f^*(x,\alpha^j)\big] \\
&\le \sum_{j=1}^{k}\Big[\widehat{y}_l(x,\widehat{\alpha}_l^j) - \widehat{y}_l(x,\alpha^j)\Big] + \sum_{j=1}^{k}\big[f^*(x,\alpha^{*,j}) - \widehat{y}_l(x,\alpha^{*,j})\big] + \sum_{j=1}^{k}\big[\widehat{y}_l(x,\alpha^j) - f^*(x,\alpha^j)\big].
\end{aligned}$$

Therefore, for any $\boldsymbol{\alpha}$,

$$\begin{aligned}
\mathsf{R}(\boldsymbol{\alpha}) \le \mathbb{E}_x \frac{1}{k}\sum_{j=1}^{k}\big[\widehat{y}_l(x,\alpha^{*,j}) - \widehat{y}_l(x,\alpha^j)\big] \\
+ \mathbb{E}_x \frac{1}{k}\sum_{j=1}^{k}\big[f^*(x,\alpha^{*,j}) - \widehat{y}_l(x,\alpha^{*,j})\big] + \mathbb{E}_x \frac{1}{k}\sum_{j=1}^{k}\big[\widehat{y}_l(x,\alpha^j) - f^*(x,\alpha^j)\big].
\end{aligned} \tag{9}$$

For the middle term in (9), we apply the last statement of Lemma 1 to $\alpha^{*,1},\ldots,\alpha^{*,k}$. We have:

$$\mathbb{E}_x \frac{1}{k}\sum_{j=1}^{k}\big[f^*(x,\alpha^{*,j}) - \widehat{y}_l(x,\alpha^{*,j})\big] \le \sqrt{2A'}\phi_l.$$

For the last term in (9),

$$\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}_x\big[\widehat{y}_l(x,\alpha^j) - f^*(x,\alpha^j)\big] \le \sqrt{2}\phi_l \cdot \left(A' + \gamma_l \mathsf{R}(\boldsymbol{\alpha})\right)^{1/2}.$$

Hence, we have the inequality

$$\begin{aligned}
\mathsf{R}(\boldsymbol{\alpha}) &\le \widehat{\mathsf{R}}_l(\boldsymbol{\alpha}) + \sqrt{2A'}\phi_l + \sqrt{2}\phi_l \cdot \left(A' + \gamma_l \mathsf{R}(\boldsymbol{\alpha})\right)^{1/2} \\
&\le \widehat{\mathsf{R}}_l(\boldsymbol{\alpha}) + 2\phi_l\sqrt{2A'} + \phi_l\sqrt{2\gamma_l \mathsf{R}(\boldsymbol{\alpha})} \\
&\le \widehat{\mathsf{R}}_l(\boldsymbol{\alpha}) + 2\phi_l\sqrt{2A'} + \gamma_l\phi_l^2 + \frac{1}{2}\mathsf{R}(\boldsymbol{\alpha})
\end{aligned}$$

and thus

$$\mathsf{R}(\boldsymbol{\alpha}) \le 2\widehat{\mathsf{R}}_l(\boldsymbol{\alpha}) + 4\phi_l\sqrt{2A'} + 2\gamma_l\phi_l^2 \le 2\widehat{\mathsf{R}}_l(q) + A'/(2\gamma_l)$$

On the other hand,

$$\Big[\widehat{y}_l(x,\widehat{\alpha}_l^j) - \widehat{y}_l(x,\alpha^j)\Big] = \Big[f^*(x,\widehat{\alpha}_l^j) - f^*(x,\alpha^j)\Big] + \Big[\widehat{y}_l(x,\widehat{\alpha}_l^j) - f^*(x,\widehat{\alpha}_l^j)\Big] + \big[f^*(x,\alpha^j) - \widehat{y}_l(x,\alpha^j)\big]$$

and so

$$\begin{aligned}
&\sum_{j=1}^{k}\Big[\widehat{y}_l(x,\widehat{\alpha}_l^j) - \widehat{y}_l(x,\alpha^j)\Big] \\
&= \sum_{j=1}^{k}\Big[f^*(x,\widehat{\alpha}_l^j) - f^*(x,\alpha^j)\Big] + \sum_{j=1}^{k}\Big[\widehat{y}_l(x,\widehat{\alpha}_l^j) - f^*(x,\widehat{\alpha}_l^j)\Big] + \sum_{j=1}^{k}\big[f^*(x,\alpha^j) - \widehat{y}_l(x,\alpha^j)\big] \\
&\le \sum_{j=1}^{k}\big[f^*(x,\alpha^{*,j}) - f^*(x,\alpha^j)\big] + \sum_{j=1}^{k}\Big[\widehat{y}_l(x,\widehat{\alpha}_l^j) - f^*(x,\widehat{\alpha}_l^j)\Big] + \sum_{j=1}^{k}\big[f^*(x,\alpha^j) - \widehat{y}_l(x,\alpha^j)\big].
\end{aligned}$$

Therefore, for any $\boldsymbol{\alpha}$

$$\widehat{\mathsf{R}}_l(\boldsymbol{\alpha}) \le \mathsf{R}(\boldsymbol{\alpha}) + \mathbb{E}_x \frac{1}{k}\sum_{j=1}^{k}[\widehat{y}_l(x,\widehat{\alpha}_l^j) - f^*(x,\widehat{\alpha}_l^j)] + \mathbb{E}_x \frac{1}{k}\sum_{j=1}^{k}[f^*(x,\alpha^j) - \widehat{y}_l(x,\alpha^j)]. \tag{10}$$

The last term in (10) is bounded by Lemma 1 by

$$\begin{aligned}
\mathbb{E}_x \frac{1}{k}\sum_{j=1}^{k}|f^*(x,\alpha^j) - \widehat{y}_l(x,\alpha^j)| &\le \sqrt{2}\phi_l \cdot (A' + \gamma_l \mathsf{R}(\boldsymbol{\alpha}))^{1/2} \\
&\le \sqrt{2}\phi_l \cdot \left(A' + 2\gamma_l \widehat{\mathsf{R}}_l(\boldsymbol{\alpha}) + A'/2\right)^{1/2} \\
&\le 2\phi_l\sqrt{A'} + 2\phi_l^2\gamma_l + \frac{1}{2}\widehat{\mathsf{R}}_l(\boldsymbol{\alpha}) \\
&\le \frac{A'}{4\gamma_l} + \frac{1}{2}\widehat{\mathsf{R}}_l(\boldsymbol{\alpha}).
\end{aligned}$$

Now, for the middle term in (10), we use the above inequality with $\widehat{\boldsymbol{\alpha}}_l = (\widehat{\alpha}_l^1, \dots, \widehat{\alpha}_l^k)$:

$$\mathbb{E}_x \frac{1}{k}\sum_{j=1}^{k}[\widehat{y}_l(x,\widehat{\alpha}_l^j) - f^*(x,\widehat{\alpha}_l^j)] \le \frac{A'}{4\gamma_l} + \frac{1}{2}\widehat{\mathsf{R}}_l(\widehat{\boldsymbol{\alpha}}_l) = \frac{A'}{4\gamma_l}.$$

Putting the terms together,

$$\widehat{\mathsf{R}}_l(\boldsymbol{\alpha}) \le 2\mathsf{R}(\boldsymbol{\alpha}) + \frac{A'}{\gamma_l}.$$

Since $\boldsymbol{\alpha}$ is arbitrary, the induction step follows. □

Lemma 3. *For $v \in \mathbb{R}^A$, let $\widehat{a}^1, \dots, \widehat{a}^k$ be indices of largest k coordinates of v in decreasing order. Let $a^1, \dots, a^k$ be any other set of distinct coordinates. Then*

$$\sum_{j=1}^{k}[v(\widehat{a}^k) - v(a^j)]_+ \le \sum_{j=1}^{k} v(\widehat{a}^j) - v(a^j)$$

Proof. We prove this by induction on r. For $r = 1$,

$$[v(\widehat{a}^1) - v(a^1)]_+ = v(\widehat{a}^1) - v(a^1)$$

Induction step: Suppose

$$\sum_{j=1}^{k-1}[v(\widehat{a}^k) - v(b^j)]_+ \le \sum_{j=1}^{k-1} v(\widehat{a}^j) - v(b^j)$$

for any $b^1, \dots, b^{k-1}$. Let $a^m = \operatorname{argmin}_{j=1,\dots,k} v(a^j)$. Since all the values are distinct, it must be that $v(\widehat{a}^k) \ge v(a^m)$. Applying the induction hypothesis to $\{a^1, \dots, a^k\} \setminus \{a^m\}$ and adding

$$[v(\widehat{a}^k) - v(a^m)]_+ = v(\widehat{a}^k) - v(a^m)$$

to both sides concludes the induction step. □

Proof of Theorem 1. Recall that on epoch l, the strategy is $\alpha_l^1 = \widehat{\alpha}_l^1, \dots, \alpha_l^{k-1} = \widehat{\alpha}_l^{k-1}$ for the first $k-1$ arms, and then sampling $\alpha_l^k(x)$ from IGW distribution p_l. Observe that for any x and any draw $\alpha_l^k(x)$, the set of k arms is distinct (i.e. the

strategies are non-overlapping), and thus under the event $\mathcal{E}$ in (4), Lemma 1 and inductive statements hold. Hence, expected regret per step in epoch l is bounded as

$$\mathbb{E}_{x,\alpha_l^k(x)} \sum_{j=1}^{k} [f^*(x, \alpha^{*,j}) - f^*(x, \alpha_l^j)] \tag{11}$$

$$\leq \frac{k(A-k+1)}{\gamma_l} + 2\mathbb{E}_{x,\alpha_l^k(x)} \sum_{j=1}^{k} [\widehat{y}_l(x, \widehat{\alpha}_l^j)) - \widehat{y}_l(x, \alpha^j)]$$

$$= \frac{k(A-k+1)}{\gamma_l} + 2\mathbb{E}_{x,\alpha_l^k(x)} [\widehat{y}_l(x, \widehat{\alpha}_l^k)) - \widehat{y}_l(x, \alpha_l^k)]$$

$$\leq \frac{k(A-k+1)}{\gamma_l} + 2\mathbb{E}_x \sum_{a \notin T_x(\widehat{y}_l)} \frac{\widehat{y}_l(x, \widehat{\alpha}_l^k) - \widehat{y}_l(x, a)}{(A-k+1) + \gamma_l[\widehat{y}_l(x, \widehat{\alpha}_l^k) - \widehat{y}_l(x, a)]}$$

$$\leq \frac{k(A-k+1)}{\gamma_l} + \frac{2(A-k+1)}{\gamma_l} \tag{12}$$

From Lemma 5, the event $\mathcal{E}$ in (4) holds with probability at least $1-\delta$ if we set

$$\phi_l = \sqrt{\frac{162}{cN_{l-1}} \log\left(\frac{|\mathcal{F}|N_{l-1}^3}{\delta}\right)}.$$

Now recall that we set $N_l = 2^l \leq 2T$ and $\gamma_l = \sqrt{A-k+1}/(32\phi_l)$. Combining this with equation (12), we find that the cumulative regret is bounded with probability at least $1-\delta$ by

$$R(T) \leq \sum_{l=2}^{e(T)} \frac{(k+2)(A-k+1)N_{l-1}}{\gamma_l}$$

$$\leq c^{-1/2} 408(k+2) \sqrt{(A-k+1)\log\left(\frac{|\mathcal{F}|T^3}{\delta}\right)} \sum_{l=2}^{\log_2(2T)} 2^{(l-1)/2}$$

$$\leq c^{-1/2} 2308(k+2) \sqrt{(A-k+1)T\log\left(\frac{|\mathcal{F}|T^3}{\delta}\right)}.$$

□

Proof of Theorem 2. The proof is essentially the same as the proof of Theorem 1.

From Lemma 7, the event $\mathcal{E}$ in (4) holds with probability at least $1-\delta$ if we set

$$\phi_l = \sqrt{\frac{420}{cN_{l-1}} \log\left(\frac{|\mathcal{F}|N_{l-1}^3}{\delta}\right) + 2\epsilon^2}.$$

Combining this with equation (12) we get that the regret is bounded by,

$$R(T) \leq \sum_{l=2}^{e(T)} \frac{(k+2)(A-k+1)N_{l-1}}{\gamma_l}$$

$$\leq c^{-1/2} 656(k+2) \sqrt{(A-k+1)\log\left(\frac{|\mathcal{F}|T^3}{\delta}\right)} \sum_{l=2}^{\log_2(2T)} 2^{(l-1)/2} + 46(k+2)\sqrt{(A-k+1)\epsilon^2} \sum_{l=2}^{e(T)} N_{l-1}$$

$$\leq c^{-1/2} 3711(k+2) \sqrt{(A-k+1)T\log\left(\frac{|\mathcal{F}|T^3}{\delta}\right)} + 46(k+2)T\sqrt{(A-k+1)\epsilon^2}$$

given $\mathcal{E}$ is true. □

D. Regression Martingale Bound

Recall that we have the following dependence structure in our problem. On each round s, context x_s is drawn independently of the past $\mathcal{H}_{s-1}$ and rewards $\boldsymbol{r}_s = \{r_s(a)\}_{a\in\mathcal{A}}$ are drawn from the distribution with mean $f^*(x_s, a)$. The algorithm selects a random set $\mathcal{A}_s$ given x_s, and feedback is provided for a (possibly random) subset $\Phi_s \subseteq \mathcal{A}$. Importantly, $\mathcal{A}_s$ and Φ_s are independent of $\boldsymbol{r}_s$ given x_s.

The next lemma considers a single time step s, conditionally on the past $\mathcal{H}_{s-1}$.

Lemma 4. *Let $x_s, \boldsymbol{r}_s = \{r_s(a)\}_{a\in\mathcal{A}}$ be sampled from the data distribution, and let $\mathcal{A}_s \subseteq \mathcal{A}$ be conditionally independent of $\boldsymbol{r}_s$ given x_s. Let $\Phi_s \subseteq \mathcal{A}_s$ be a random subset given $\mathcal{A}_s$ and x_s, but independent of $\boldsymbol{r}_s$. Fix an arbitrary $f : \mathcal{X} \times \mathcal{A} \to [0, 1]$ and define the following random variable,*

$$Y_s = \frac{1}{k}\sum_{a\in\mathcal{A}}\big((f(x_s, a) - r_s(a))^2 - (f^*(x_s, a) - r_s(a))^2\big) \times \mathbf{1}\{a \in \Phi_s\}.$$

Then, under the realizability assumption (Assumption 1), we have the following,

$$\mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s] = \frac{1}{k}\sum_{a\in\mathcal{A}}\mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\big\{(f(x_s, a) - f^*(x_s, a))^2 \times \mathbf{1}\{a \in \Phi_s\}\big\}$$

and

$$\mathrm{Var}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s] \le 4\mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s].$$

Proof. By the conditional independence assumptions,

$$\begin{aligned}\mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s] &= \frac{1}{k}\sum_{a\in\mathcal{A}}\mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}\{(f(x_s, a) - f^*(x_s, a))(f(x_s, a) + f^*(x_s, a) - 2r_s(a)) \times \mathbf{1}\{a \in \Phi_s\}\}\\ &= \frac{1}{k}\sum_{a\in\mathcal{A}}\mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\big\{(f(x_s, a) - f^*(x_s, a))^2 \times \mathbf{1}\{a \in \Phi_s\}\big\}.\end{aligned}$$

We also have

$$\begin{aligned}Y_s^2 &\le \frac{1}{k}\sum_{a\in\mathcal{A}}(f(x_s, a) - f^*(x_s, a))^2(f(x_s, a) + f^*(x_s, a) - 2r_s(a))^2 \times \mathbf{1}\{a \in \Phi_s\}\\ &\le \frac{4}{k}\sum_{a\in\mathcal{A}}(f(x_s, a) - f^*(x_s, a))^2 \times \mathbf{1}\{a \in \Phi_s\}.\end{aligned}$$

□

Lemma 5. *Let $\widehat{y}_l$ be the estimate of the regression function f^* at epoch l. Assume the conditional independence structure in Lemma 4 and suppose Assumption 1 holds. Let $\mathcal{H}_{t-1}$ denote history (filtration) up to time $t-1$. Then for any $\delta < 1/e$,*

$$\mathcal{E} = \left\{l \ge 2 : \sum_{s=1}^{N_{l-1}}\mathbb{E}_{x_s,\mathcal{A}_s}\left\{\frac{1}{k}\sum_{a\in\mathcal{A}_s}(\widehat{y}_l(x_s, a) - f^*(x_s, a))^2|\mathcal{H}_{s-1}\right\} \le c^{-1}81\log\left(\frac{|\mathcal{F}|N_{l-1}^3}{\delta}\right)\right\}$$

holds with probability at least $1 - \delta$.

Proof. Following Lemma 4, let

$$Y_s(f) = \frac{1}{k}\sum_{a\in\mathcal{A}}\big((f(x_s, a) - r_s(a))^2 - (f^*(x_s, a) - r_s(a))^2\big) \times \mathbf{1}\{a \in \Phi_s\}.$$

The argument proceeds as in (Agarwal et al., 2012). Let $\mathbb{E}_s$ and Var_s denote the conditional expectation and conditional variance given $\mathcal{H}_{s-1}$. By Freedman's inequality (Bartlett et al., 2008), for any t, with probability at least $1-\delta'\log t$, we have

$$\sum_{s=1}^{t}\mathbb{E}_s[Y_s(f)]-\sum_{s=1}^{t}Y_s(f)\le 4\sqrt{\sum_{s=1}^{t}\mathrm{Var}_s[Y_s(f)]\log(1/\delta')}+2\log(1/\delta')$$

Let $X(f)=\sqrt{\sum_{s=1}^{t}\mathbb{E}_s[Y_s(f)]}$, $Z(f)=\sum_{s=1}^{t}Y_s(f)$ and $C=\sqrt{\log(1/\delta')}$. In view of Lemma 4, with probability at least $1-\delta'\log t$,

$$X(f)^2-Z(f)\le 8CX(f)+2C^2$$

and hence

$$(X(f)-4C)^2\le Z(f)+18C^2.$$

Consequently, with the aforementioned probability, for all functions $f\in\mathcal{F}$ (and, in particular, for $\widehat{y}_l$),

$$(X(f)-4C')^2\le Z(f)+18C'^2$$

where $C'=\sqrt{\log(|\mathcal{F}|/\delta')}$. Now recall that

$$\widehat{y}_l=\operatorname*{argmin}_{f\in\mathcal{F}}\sum_{t=1}^{N_{l-1}}\sum_{a\in\Phi_t}(f(x_t,a)-r_t(a))^2$$

where Φ_t is a random feedback set satisfying Assumption 3. Hence, $Z(\widehat{y}_l)\le 0$ for $t=N_{l-1}$, implying that with probability at least $1-\delta'/(N_{l-1}^2)$,

$$\sum_{s=1}^{N_{l-1}}\frac{1}{k}\sum_{a\in\mathcal{A}}\mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\left\{(\widehat{y}_l(x_s,a)-f^*(x_s,a))^2\times\mathbf{1}\{a\in\Phi_s\}|\mathcal{H}_{s-1}\right\}\le 81\log\left(\frac{|\mathcal{F}|N_{l-1}^2\log(N_{l-1})}{\delta'}\right).$$

We now take a union bound over l and recall that $\sum_{i\ge1}1/i^2=\pi^2/6<2$.

Finally, observe that by Assumption 3,

$$\begin{aligned}&\mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\left\{(\widehat{y}_l(x_s,a)-f^*(x_s,a))^2\times\mathbf{1}\{a\in\Phi_s\}|\mathcal{H}_{s-1}\right\}\\&=\mathbb{E}_{x_s,\mathcal{A}_s}\left\{(\widehat{y}_l(x_s,a)-f^*(x_s,a))^2\times\mathbf{1}\{a\in\mathcal{A}_s\}\times\mathbb{P}(a\in\Phi_s|x_s,\mathcal{A}_s)|\mathcal{H}_{s-1}\right\}\\&\ge c\cdot\mathbb{E}_{x_s,\mathcal{A}_s}\left\{(\widehat{y}_l(x_s,a)-f^*(x_s,a))^2\mathbf{1}\{a\in\mathcal{A}_s\}|\mathcal{H}_{s-1}\right\}.\end{aligned}$$

We conclude that with probability at least $1-2\delta'$, for all $l\ge2$,

$$\sum_{s=1}^{N_{l-1}}\frac{1}{k}\mathbb{E}_{x_s,\mathcal{A}_s}\left\{\sum_{a\in\mathcal{A}_s}(\widehat{y}_l(x_s,a)-f^*(x_s,a))^2|\mathcal{H}_{s-1}\right\}\le c^{-1}81\log\left(\frac{|\mathcal{F}|N_{l-1}^2\log(N_{l-1})}{\delta'}\right).$$

□

E. Regression Martingale Bound with Misspecification

Lemma 6. *Under the notation and assumptions of Lemma 4, but in the case of misspecified model (Assumption 2 replacing Assumption 1), it holds that*

$$\mathrm{Var}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s]\le 8\mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s]+16\epsilon^2.$$

Proof. The proof is along the lines of Lemma 4 (see also (Foster & Rakhlin, 2020)). We have for any $f : \mathcal{X} \times \mathcal{A} \to [0, 1]$,

$$\begin{aligned}\mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s] &= \frac{1}{k}\sum_{a\in\mathcal{A}} \mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}\{(f(x_s,a) - f^*(x_s,a))(f(x_s,a) + f^*(x_s,a) - 2r_s(a)) \times \mathbf{1}\{a \in \Phi_s\}\} \\ &= \frac{1}{k}\sum_{a\in\mathcal{A}} \mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\big\{(f(x_s,a) - f^*(x_s,a))^2 \times \mathbf{1}\{a \in \Phi_s\}\big\} \\ &+ \frac{2}{k}\sum_{a\in\mathcal{A}} \mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\{(f(x_s,a) - f^*(x_s,a))(f^*(x_s,a) - \mathbb{E}_{\boldsymbol{r}_s}[r(a)|x_s]) \times \mathbf{1}\{a \in \Phi_s\}\}.\end{aligned}$$

Rearranging, using AM-GM inequality, and Assumption 2,

$$\begin{aligned}&\frac{1}{k}\sum_{a\in\mathcal{A}} \mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\big\{(f(x_s,a) - f^*(x_s,a))^2 \times \mathbf{1}\{a \in \Phi_s\}\big\} \\ &= \mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s] - \frac{2}{k}\sum_{a\in\mathcal{A}} \mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\{(f(x_s,a) - f^*(x_s,a))(f^*(x_s,a) - \mathbb{E}_{\boldsymbol{r}_s}[r(a)|x_s]) \times \mathbf{1}\{a \in \Phi_s\}\} \\ &\le \mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s] + \frac{1}{2k}\sum_{a\in\mathcal{A}} \mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\big\{(f(x_s,a) - f^*(x_s,a))^2 \times \mathbf{1}\{a \in \Phi_s\}\big\} + 2\epsilon^2.\end{aligned}$$

Rearranging,

$$\frac{1}{k}\sum_{a\in\mathcal{A}} \mathbb{E}_{x_s,\mathcal{A}_s,\Phi_s}\big\{(f(x_s,a) - f^*(x_s,a))^2 \times \mathbf{1}\{a \in \Phi_s\}\big\} \le 2\mathbb{E}_{x_s,\boldsymbol{r}_s,\mathcal{A}_s,\Phi_s}[Y_s] + 4\epsilon^2.$$

On the other hand,

$$\begin{aligned}Y_s^2 &\le \frac{1}{k}\sum_{a\in\mathcal{A}} (f(x_s,a) - f^*(x_s,a))^2 (f(x_s,a) + f^*(x_s,a) - 2r_s(a))^2 \times \mathbf{1}\{a \in \Phi_s\} \\ &\le \frac{4}{k}\sum_{a\in\mathcal{A}} (f(x_s,a) - f^*(x_s,a))^2 \times \mathbf{1}\{a \in \Phi_s\}.\end{aligned}$$

Combining the two inequalities concludes the proof. □

Lemma 7. *Let $\widehat{y}_l$ be the estimate of the regression function f^* at epoch l. Assume the conditional independence structure in Lemma 4 and suppose Assumption 2 holds. Let $\mathcal{H}_{t-1}$ denote history (filtration) up to time $t-1$. Then for any $\delta < 1/e$,*

$$\mathcal{E} = \left\{ l \ge 2 : \sum_{s=1}^{N_{l-1}} \mathbb{E}_{x_s,\mathcal{A}_s}\left\{\frac{1}{k}\sum_{a\in\mathcal{A}_s} (\widehat{y}_l(x_s,a) - f^*(x_s,a))^2 | \mathcal{H}_{s-1}\right\} \le c^{-1} 210 \log\left(\frac{|\mathcal{F}|N_{l-1}^3}{\delta}\right) + \epsilon^2 N_{l-1} \right\}$$

holds with probability at least $1 - \delta$.

Proof. We follow the proof of Lemma 5 to see how the misspecification level ϵ^2 enters the bounds.

Let $X(f) = \sum_{s=1}^t \mathbb{E}_s[Y_s(f)]$, $Z(f) = \sum_{s=1}^t Y_s(f)$, $C = \log(1/\delta')$ and $M = \epsilon^2 t$. Now using Lemma 6 and Freedman's inequality in the proof of Lemma 5, we find that with probability at least $1 - \delta' \log t$,

$$\begin{aligned}&X(f) - Z(f) \le 8\sqrt{C(2X(f) + 4\epsilon^2 t)} + 2C \\ &\implies (X(f) - Z(f) - 2C)^2 \le 128 X(f) C + 256 MC \\ &\implies (X(f) - 66C - Z(f))^2 \le 4352 C^2 + 256 MC + 128 Z(f) C.\end{aligned}$$

The above bound holds for a fixed function f. We now apply an union bound to conclude that for all functions $f \in \mathcal{F}$, with probability at least $1 - \delta' \log t$,

$$\begin{aligned}(X(f) - 66C' - Z(f))^2 &\le 4352 C'^2 + 256 MC' + 128 Z(f) C' \\ &\le 20736 C'^2 + M^2 + 128 Z(f) C'\end{aligned}$$

where $C' = \log(|\mathcal{F}|/\delta')$. As in Lemma 5, $Z(\widehat{y}_l) \leq 0$ when $t = N_{l-1}$ and thus with probability at least $1 - \delta' \log(N_{l-1})$,

$$X(\widehat{y}_l) \leq 210C' + \epsilon^2 N_{l-1}.$$

Hence, with probability at least $1 - \delta'/N_{l-1}^2$,

$$\sum_{s=1}^{N_{l-1}} \frac{1}{k} \sum_{a \in \mathcal{A}} \mathbb{E}_{x_s, \mathcal{A}_s, \Phi_s} \left\{ (\widehat{y}_l(x_s, a) - f^*(x_s, a))^2 \times \mathbf{1}\{a \in \Phi_s\} | \mathcal{H}_{s-1} \right\} \leq 210 \log\left(\frac{|\mathcal{F}| N_{l-1}^2 \log(N_{l-1})}{\delta'} \right) + \epsilon^2 N_{l-1}.$$

The rest of the proof proceeds exactly as in Lemma 5. □

F. Reduction from eXtreme to $\log(A)$-armed Contextual Bandits

In this section we will prove Corollary 1 which is a reduction style argument. We reduce the A armed top-k contextual bandit problem under Definition 1 to a Z armed top-k contextual bandit problem where $Z = O(\log A)$.

Proof of Corollary 1. Note that the proof of Theorem 1 does not require the physical definition of an arm being consistent across all contexts as long as realizability holds. Let us assume w.l.o.g that Algorithm 2 returns the internal and leaf effective arms for any context x in $\mathcal{A}_x$ in a deterministic ordering. Let us call the j-th effective arm in this ordering for any context as arm j. This defines a system with Z arms where $Z \leq (p-1)b(H-1) + bm$ as Z is the number of effective arms returned by the beam-search in Algorithm 2. Recall the definition of the new function class $\tilde{\mathcal{F}}$ from Section 4.1. We can thus say that when Definition 1 holds this new system is a Z armed top-k contextual bandit system with realizablity (Assumption 1) with function class $\tilde{\mathcal{F}}$. Therefore the first part of coroallary 1 is implied by Theorem 1. Similarly when Definition 1 holds along with Assumption 2, this new system is a Z armed top-k contextual bandit system with ϵ-realizablity (Assumption 2) with function class $\tilde{\mathcal{F}}$. Therefore the second part of corollary 1 is implied by Theorem 2. Note that we have used the fact $|\tilde{\mathcal{F}}| = |\mathcal{F}|$. □

G. More Experiments

In Figure 4 we plot the progressive mean rewards vs time for all the experiments using simulated bandit feedback on eXtreme datasets.

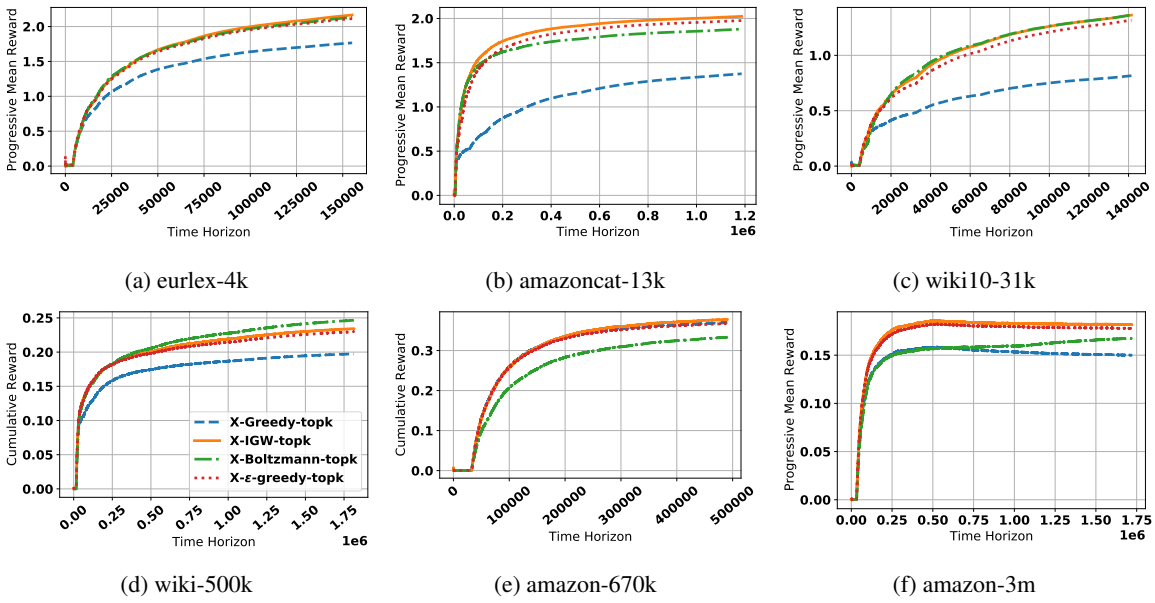
H. Implementation Details

For the realizable experiment on Eurlex-4k shown in Figure 3(a), the optimal weights ν^*'s are obtained by training ridge regression on the rewards vs context for each arm in the dataset. During the experiment we also use the same function class, that is one ridge regression is trained per arm on all collected data during the course of the algorithm. The reward for arm a given context x is chosen as $r_t(a) = [x; 1.0]^T \nu_a^* + \epsilon_t$, where ϵ_t is a zero-mean Gaussian noise.

Simulated Bandit Feedback: A sample in a multi-label dataset can be described as $(x, \mathbf{y})$ where $x \in \mathcal{X}$ can be thought of as the context while $\mathbf{y} \in \{0, 1\}^L$ denotes the correct classes. We can shuffle such a dataset into an ordering $\{(x_t, \mathbf{y}^{(t)})\}_{t=1}^T$. Then we feed one sample from the dataset at each time step to the contextual bandit algorithm that we are evaluating, in the following manner,

- at time t, send the input x_t to the contextual bandit algorithm,
- the contextual bandit algorithm then chooses an action corresponding to k arms $\boldsymbol{a}_t$,
- the environment then reveals the reward for **only** the k arms chosen $\boldsymbol{r}_t(\boldsymbol{a}_t)$, i.e. whether the arms chosen are among the correct classes or not.

Note that the algorithm is free to optimize its policy for choosing arms based on everything it has seen so far. In practice however, most contextual bandit algorithms will improve their policy (the $\widehat{y}$ it has learnt) in batches. The total number of positive classes selected by the algorithm in this process is the total reward collected by the algorithm.



(a) eurlex-4k (b) amazoncat-13k (c) wiki10-31k

(d) wiki-500k (e) amazon-670k (f) amazon-3m

Figure 4: We plot the progressive mean rewards collected by each algorithm as a function of time. All algorithms are implemented under our eXtreme reduction framework. The initialization held out set for each dataset is used to train the hierarchy and the routing functions. Then the regressors for all nodes are trained on collected data at the beginning of each epoch. In all our experiments we have $k = 5$. In Algorithm 3 we set the number of explore slots $r = 3$. The common legend for all the plots is provided in (d). The beam-size used is $b = 10$.

eXtreme Framework: We follow the framework described in Section 4. We first form the tree and the routing functions from the held out portion of each dataset. The assumption is that there is a small supervised dataset available to each algorithm before proceeding with the simulated bandit feedback experiment. This dataset is used to form a balanced binary tree over the labels till the penultimate level. The nodes in the penultimate level can have a maximum of m children which are the original arms. The value of m is specified in Table 1 for each dataset. The division of the labels in each level of the tree is done through hierarchical 2-means clustering over label embeddings, where at each clustering step we use the algorithm from (Dhillon, 2001). The specific label embedding technique that we use is called Positive Instance Feature Aggregation (PIFA) (see and (Prabhu et al., 2018) for more details). The routing functions for each internal node in the tree is essentially a one-vs-all linear classifier trained on the held out set. The classifiers are trained using a SVM ℓ_2-hinge loss. The positive and negative examples for each internal node is selected similar to the strategy in (Prabhu et al., 2018). Finally for the regression function $\tilde{f}(x, \tilde{a})$ where $\tilde{a}$ can be an original arm or an internal node in the tree, we train a linear regressor $\tilde{f}(x, \tilde{a}) = \nu_{\tilde{a}}^T [x; 1]$ as we progress through the experiment as in Algorithm 3. Note that the held out dataset is only used to train the tree and the routing function for each of the algorithms, while the regression functions are trained from scratch only using the samples observed during the bandit feedback experiment. The details are as follows:

- **Tree:** Initially a small part of the dataset is supplied to the algorithms in full-information mode. The size of this portion is captured in Table 1 in the Initialization Size column. This portion is used to construct an approximately balanced binary tree over the labels. A supervised multilabel dataset can be represented as (X, Y) where $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times L}$. We form an embedding for each label using PIFA (Jasinska et al., 2016; Prabhu et al., 2018; Yu et al., 2020). Essentially the embedding for each label is the average of all instances that the label is connected to, normalized to ℓ_2 norm 1. Then we use approximately balanced 2-means recursively to form the tree until each leaf has less than a predefined maximum number of labels. The exact clustering algorithm used at each step is (Dhillon, 2001).
- **Routing Functions:** The routing functions are essentially one-vs-all linear classifiers at each internal node of the tree. The positive examples for the classifier at an internal node are the input instances in the small supervised dataset that have a positive label in the subtree of that node. The negative instances are the set of all instances that has a positive label in the subtree of the parent of that node but not in that node's subtree. This is the same methodology as in (Prabhu et al., 2018). The routing functions are trained using

LinearSVC (Fan et al., 2008).

- **Regression Functions:** After creating the tree and the routing function from the small held out set, they are held fixed. The function class $\tilde{\mathcal{F}}$ as Algorithm 3 progresses is a set of linear regression functions at each internal and leaf node of the tree. They are trained on past data collected during the course of the previous epochs. Note that the examples for training the regressor for an internal node are only from the singleton arms that were shown when the algorithm selected that particular internal node in the IGW sampling. The regression functions are trained using LinearSVR (Fan et al., 2008).
- **Hyper-parameter Tuning:** For all the exploration algorithms in the eXtreme experiments the parameters are tuned over the eurlex-4k dataset and then held fixed. For the IGW scheme C is tuned over a grid of $\{1e-7, 1e-6, \cdots, 1e7\}$. The same is done for the β in the Boltzmann scheme. For ϵ-greedy the ϵ value is tuned between $[1e-7, 1.0]$ in a equally spaced grid in the logarithmic scale. The best parameters that are found are $\beta = 1.0, C = 1.0$ and $\epsilon = 0.167$.
- **Inference:** Inference using a trained model is done exactly according to Algorithm 3. The beam-search over the routing function yields effective arms. Then we evaluate the linear regression functions for each of the effective arms (singleton arms or the internal nodes in the tree). If a non-singleton effective arm is chosen among the k arms we randomly sample a singleton arm in it's subtree. The beam search and IGW sampling is implemented in C++ where the linear operations are implemented using the Eigen package (Guennebaud et al., 2010).