Fair and Risk-Averse Worker Selection in Mobile Crowdsourcing via Mean-Variance Bandits

Ziqun Chen, Kechao Cai, *Member, IEEE*, and Jinbei Zhang, *Member, IEEE*

***Abstract*—Mobile crowdsourcing is recognized as a promising paradigm for harnessing collective wisdom to solve problems. The crowdsourcing platform assigns tasks to a group of workers (e.g., vehicles, sensors) and receives the outcomes from them. In this paper, we develop a novel combinatorial mean-variance bandit model for worker selection and characterize the quality of workers with mean-variance measures to consider both the reward and risk in task assignment. Moreover, we introduce a utility-based fairness constraint to ensure a fair selection of workers. We consider two types of feedback: quantitative feedback, where the platform can receive numerical rewards from the selected workers, and preference feedback, where the platform only receives a qualitative performance ranking of the selected workers. We define the mean-variance regret and fairness regret and propose novel bandit algorithms that balance the fairness guarantee and mean-variance optimization in worker selection under both feedback settings. We prove that our algorithms achieve sublinear expected mean-variance regret and fairness regret. Through extensive simulations, we validate that our algorithms can fairly select workers while maximizing the mean-variance of selected workers.**

***Index Terms*—Mobile crowdsourcing, mean-variance bandit, utility-based fairness, quantitative feedback, preference feedback.**

I. Introduction

With the rapid development of mobile devices and wireless networks, mobile crowdsourcing has come into focus as a novel problem-solving paradigm in recent years [1]. The crowdsourcing platform assigns tasks like traffic detection [2], air quality measurement [3], and street view collection [4] to a group of workers (e.g., vehicles, sensors) selected from the candidates, who are willing to participate in the crowdsourcing. The selected workers then execute the tasks and send the outcomes to the platform. It is favorable for the platform to select workers with higher quality to improve task performance and user satisfaction. However, practically, the quality of workers is usually unknown in advance. Therefore, the platform has to experience a learning process to estimate the quality of workers. At this point, the platform faces a dilemma: explore various workers to learn about their quality or exploit the historical observations by selecting those with previously demonstrated high performance.

The emerging field of combinatorial multi-armed bandit (CMAB) offers a promising framework for worker selection problems in mobile crowdsourcing. Specifically, we view the platform as a player and the workers as arms. Each arm generates a random reward that reflects the corresponding worker's task performance and follows an unknown distribution. The player's goal is to maximize the cumulative reward based on the previous observations and decision history.

However, directly applying the standard CMAB model to mobile crowdsourcing can be problematic, as it fails to address the multifaceted challenges inherent in real-world task assignment. Firstly, it does not consider the risk of assigning tasks to workers with unstable performance. When some workers' task performance fluctuates significantly, they may perform poorly even with high average quality. Selecting such workers would degrade the overall platform experience and introduce unacceptable risks and costs. In particular, it may result in frequent worker switches, where tasks are repeatedly reassigned to different workers, thereby incurring significant communication overhead and switching costs [5]. Moreover, the conventional CMAB model ignores the interests of some workers, resulting in an unfair worker selection [6]. Consider a bandit algorithm that tries to maximize the worker's utility: it simply learns which worker has the highest quality and constantly assigns the task to that worker, even if other workers are almost equally good. This approach leads to a winner-takes-all allocation where many skillful workers will not receive sufficient tasks, and may eventually lose interest in the platform. Thus, to build a sustainable platform, a good policy should guarantee fairness among workers and ensure that workers with comparable skill levels have similar probabilities of being assigned tasks. Finally, for tasks such as content moderation or image annotation, it is challenging for the crowdsourcing platform to figure out the exact numerical rewards, known as quantitative feedback, from workers' outcomes, as it requires domain experts to establish detailed evaluation metrics beforehand [7]. In contrast, eliciting preference feedback through qualitative ranking of workers' task performance is generally more practical and easier than direct quantitative evaluation.

Main Contributions. Motivated by the above discussions, in this paper, we develop a novel combinatorial mean-variance bandit model (CMVB) for mobile crowdsourcing that considers both reward and risk in task assignment, while guaranteeing utility-based fairness among workers.

We measure worker quality using the *mean-variance* of their performance and define the utility of a worker as a function of its mean-variance. Then we impose a *utility-based fairness* constraint to ensure each worker is selected with a probability proportional to its utility. We design two fair mean-variance bandit algorithms, MVFQ and MVFP, tailored for *quantitative feedback* and *preference feedback*, respectively.

Ziqun Chen, Kechao Cai, and Jinbei Zhang are with the School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China, 518107. (e-mail: chenzq35@mail2.sysu.edu.cn; caikch3@mail.sysu.edu.cn; zhjinbei@mail.sysu.edu.cn). (Corresponding author: Kechao Cai).

We also introduce *mean-variance regret* and *fairness regret* as performance metrics to evaluate the effectiveness of our algorithms.

Our theoretical analysis and experimental results demonstrate that the proposed algorithms achieve sublinear upper bounds for both types of expected regret, outperforming existing methods by fairly selecting workers based on their utility while maximizing the mean-variance of selected workers.

II. Fair Mean-Variance Bandit with Quantitative Feedback

In this section, we present the details of our CMVB model with quantitative feedback and design a CMVB algorithm for mobile crowdsourcing to ensure utility-based fairness among workers.

A. Problem Formulation

We consider a mobile crowdsourcing system with a crowdsourcing platform and a set of workers $[K] = \{1, 2, ..., K\}$, which can provide computing and sensing service. Let $[T] := \{1, 2, ..., T\}$ denote the set of decision rounds. At round $t \in [T]$, the platform selects a subset S_t of L ($L \leq K$) workers from $[K]$ to execute the tasks. Once a selected worker $i \in S_t$ has returned the outcome, the platform will generate a numerical reward $r_{i,t}$ as quantitative feedback based on the worker's task performance. Due to the system uncertainty (e.g., worker mobility, transmission/processing oscillation), the reward $r_{i,t} \in [0, 1]$ is a random variable that follows an unknown distribution with expectation $\mu_i = \mathbb{E}[r_{i,t}]$ and variance $\sigma_i^2 = \mathrm{Var}[r_{i,t}]$. The μ_i and σ_i^2 characterize the reliability and stability of worker i, respectively. When the reward variance σ_i^2 is large, the worker could still perform poorly even with a high expected reward μ_i. To capture the tradeoff between a worker's reliability and stability, we formulate the worker selection problem as a CMVB model. The quality of each worker is measured by a *mean-variance* metric, defined for worker i as $w_i = \rho\mu_i - \sigma_i^2$, where $\rho > 0$ is the risk tolerance parameter that controls the balances the two objectives of high expected reward and low variance.

To ensure that similar levels of workers obtain comparable treatment, we define a utility function $f(\cdot) > 0$ that maps the mean-variance of a worker to a positive utility value. Then we enforce a utility-based fairness constraint that the probability p_i of selecting worker i is proportional to its utility $f(w_i)$. Formally, we have

$$\frac{p_i}{f(w_i)} = \frac{p_{i'}}{f(w_{i'})}, \quad \forall i \neq i', i, i' \in [K]. \tag{1}$$

The introduced utility function $f(\cdot)$ allows us to tailor the fairness criterion for different scenarios. We have two assumptions on the utility function $f(\cdot)$.

Assumption II.1. The utility of each worker is bounded such that (i) $\exists\ \lambda > 0$ and $\min_w f(w) \geq \lambda$, (ii) $\forall w_1, w_2, \frac{f(w_1)}{f(w_2)} \leq \frac{K-1}{L-1}$ for $L > 1$.

Assumption II.2. The utility function f is M-Lipschitz continuous, i.e., there exists a positive constant $M > 0$, such that $\forall w_1, w_2, |f(w_1) - f(w_2)| \leq M\,|w_1 - w_2|$.

We now show that there is a unique optimal fair policy that fulfills the fairness constraints in (1) in the following theorem.

Theorem II.3. *For any $w_i, i \in [K]$ and any choice of utility function $f(\cdot) > 0$ under Assumption II.1 and II.2, there exist a unique optimal fair policy $\boldsymbol{p}^* = \{p_1^*, p_2^*, ..., p_K^*\}$ such that*

$$p_i^* = \frac{Lf(w_i)}{\sum_{i'=1}^{K} f(w_{i'})}, \quad \forall i \in [K], \tag{2}$$

that satisfies the utility-based fairness constraints in (1).

Due to the space limit, all the proofs of the theorems are provided in the Appendix of the full version of the paper. Theorem II.3 implies that the optimal fair policy in our model is no longer selecting a fixed optimal set of L workers as in classical bandit problems, but a probability distribution on all the possible sets $S_t \subseteq [K]$, $|S_t| = L$. To be more specific, we characterize a worker selection policy with a probabilistic selection vector $\boldsymbol{p}_t = \{p_{1,t}, p_{2,t}, ..., p_{K,t}\}$ where $p_{i,t} \in [0, 1]$ is the probability of selecting worker $i \in [K]$ at round t, and $\sum_{i=1}^{K} p_{i,t} = L$ since only L workers can be selected at each round.

To measure the gap in the expected mean-variance of workers between the optimal fair policy and the deployed policy, we define the *mean-variance regret* as follows:

$$\mathrm{MR}_T = \sum_{t=1}^{T} \max\left\{\sum_{i=1}^{K} p_i^* w_i - \sum_{i=1}^{K} p_{i,t} w_i, 0\right\}, \tag{3}$$

which quantifies the speed of the mean-variance optimization of the deployed policy. Especially, we only consider the non-negative part at each round in (3) as a less fair policy could have a larger mean-variance than the optimal fair policy and cause a negative mean-variance gap. Moreover, we also require a measure to quantify its fairness guarantee. We define the *fairness regret* that measures the cumulative 1-norm distance between the optimal fair policy $\boldsymbol{p}^*$ and the deployed policy $\boldsymbol{p}_t$ as follows:

$$\mathrm{FR}_T = \sum_{t=1}^{T} \sum_{i=1}^{K} |p_i^* - p_{i,t}|. \tag{4}$$

The fairness regret measures the overall violation of the utility-based fairness constraints. Our objective is to design selection policies that have both *sublinear expected mean-variance regret* and *sublinear expected fairness regret* with respect to the number of rounds T, where the expectations are taken over the randomness in both the worker selections and the mean-variance measure. By doing so, we can approach the optimal fair policy and select the workers with high quality while ensuring fairness among all the workers in the long run.

It is important to point out that both Assumption II.1 and Assumption II.2 are necessary for designing CMVB algorithms as stated in the following theorem and remark.

Theorem II.4. *For any bandit algorithm, if either Assumption II.1 (i) or Assumption II.2 does not hold, the lower bound of the fairness regret is linear; in other words, there exists a CMVB instance with linear expected fairness regret $O(T)$.*

Algorithm 1 MVFQ Algorithm

Input: $f(\cdot)$, T, L, K, $\rho > 0$, $\delta = \frac{1}{LT}$.
Init: Select each worker in $[K]$ once with $\lceil K/L \rceil$ rounds.
Initial selection policy: $p_{i,\lceil K/L \rceil + 1} = 1/K$, $\forall i \in [K]$.

1: **for** $t = \lceil K/L \rceil + 1$ to T **do**
2: Select workers in $S_t = \text{RRS}(L, \boldsymbol{p}_t)$
3: Receive reward $r_{i,t}$ from $i \in S_t$
4: **for** $i \in [K]$ **do**
5: $n_{i,t} = \sum_{\tau=1}^{t} \mathbb{1}_{\{i \in S_\tau\}}$
6: $\hat{\mu}_{i,t} = \sum_{\tau=1}^{t} \mathbb{1}_{\{i \in S_\tau\}} r_{i,\tau} / n_{i,t}$
7: $\hat{\sigma}^2_{i,t} = \sum_{\tau=1}^{t} \mathbb{1}_{\{i \in S_\tau\}} \left(r_{i,\tau} - \hat{\mu}_{i,t}\right)^2 / n_{i,t}$
8: $\hat{w}_{i,t} = \rho \hat{\mu}_{i,t} - \hat{\sigma}^2_{i,t}$
9: $u_{i,t} = \hat{w}_{i,t} + (\rho + 3) \sqrt{\frac{\log(8KT/\delta)}{2n_{i,t}}}$
10: $l_{i,t} = \hat{w}_{i,t} - (\rho + 3) \sqrt{\frac{\log(8KT/\delta)}{2n_{i,t}}}$
11: **end for**
12: $\mathcal{C}_t = \{\widetilde{\boldsymbol{w}} \in \mathbb{R}^K \mid \forall i \in [K],\ \widetilde{w}_i \in [l_{i,t}, u_{i,t}]\}$
13: $\widetilde{\boldsymbol{w}}_t = \arg\max_{\widetilde{\boldsymbol{w}} \in \mathcal{C}_t} \sum_{i \in [K]} \frac{Lf(\widetilde{w}_i)}{\sum_{i' \in [K]} f(\widetilde{w}_{i'})} \widetilde{w}_i$
14: Compute $p_{i,t+1} = \frac{Lf(\widetilde{w}_{i,t})}{\sum_{i' \in [K]} f(\widetilde{w}_{i',t})}$ for $i \in [K]$
15: **end for**

Remark II.5. Assumption II.1 (ii) ensures that the selection probability $p_{i,t}$ in the form of $Lf(\cdot)/\sum_{a=1}^{K} f(\cdot)$ is constrained in $[0, 1]$.

B. Algorithm Description

Algorithm 1 shows the details of our *Mean-Variance Fair learning with Quantitative feedback* (MVFQ) algorithm, which follows the principle of optimism in the face of uncertainty. At each round t, we incorporate a randomized rounding scheme (RRS) from [8] to stochastically select workers with the selection vector $\boldsymbol{p}_t$, to ensure fairness. In line 2, RRS takes $\boldsymbol{p}_t$ as input and generates a set of selected workers S_t such that $\mathbb{E}[\mathbb{1}_{\{i \in S_t\}}] = p_{i,t}$, where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. Given the observed reward and the number of samples $n_{i,t}$ up to round t, we compute the empirical mean-variance estimate $\hat{w}_{i,t}$ for each worker $i \in [K]$, using empirical expected reward $\hat{\mu}_{i,t}$ and variance $\hat{\sigma}^2_{i,t}$. Then, using both UCB (Upper Confidence Bound) estimates $u_{i,t}$ and LCB (Lower Confidence Bound) estimates $l_{i,t}$ of all workers, we can construct a confidence region $\mathcal{C}_t$ (see Line 12) which contains the actual mean-variance vector $\boldsymbol{w} := (w_i)_{i \in [K]}$ with high probability. We find a vector $\tilde{\boldsymbol{w}}_t := (\tilde{w}_{i,t})_{i \in [K]}$ in the confidence region $\mathcal{C}_t$ that maximizes the expected mean-variance of a fair policy as shown in Line 13. Finally, according to Theorem II.3, we update the selection policy $p_{i,t+1}$ of worker $i \in [K]$ as $\frac{Lf(\tilde{w}_{i,t})}{\sum_{i'=1}^{K} f(\tilde{w}_{i',t})}$ to satisfy the utility-based fairness constraints, which is limited to the interval $[0, 1]$ under Assumption II.1 (ii).

We present the sublinear expected mean-variance regret and fairness regret upper bounds of MVFQ in the following theorem.

Theorem II.6. *For MVFQ in Algorithm 1, the expected mean-variance regret is upper bounded by* $O\left(\rho\sqrt{LKT\log T}\right)$ *and the expected fairness regret of MVFQ is upper bounded by* $O\left(\frac{\rho ML}{\lambda}\sqrt{LKT\log T}\right)$.

In Theorem II.6, the factor $\frac{ML}{\lambda}$ in the fairness regret bound comes from Assumption II.1 (i) and Assumption II.2 on the utility function $f(\cdot)$, while the mean-variance regret bound does not depend on Assumption II.1 (i) and Assumption II.2, making it tighter up to logarithmic factors.

III. Fair Mean-Variance Bandit with Preference Feedback

In this section, we consider the more challenging CMVB model with preference feedback, where the platform receives only the qualitative ranking of workers based on task performance, instead of numerical rewards. To address this, we propose another CMVB algorithm to maximize the mean-variance of workers and ensure utility-based fairness among workers.

A. Problem Formulation

To characterize the relative preferences of workers in a given set, we adopt the Multinomial-Logit (MNL) probability model with unknown parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_i, ..., \theta_K)$ where each component $\theta_i \in (0, 1]$ corresponds to the potential strength of a worker i. At each round t, the platform assigns the tasks to a subset $S_t \subseteq [K]$ of L workers and receives as *preference feedback* a partial task performance ranking π_t of $1 \le m \le L-1$ workers from S_t. Let $\pi_{i,t}$ denote the worker at i-th position in π_t and $\mathcal{P}^m_{S_t}$ denote the set of permutations of any m-subset of S_t. The ranking π_t, of the form $(\pi_{1,t} \succ ... \succ \pi_{m,t})$, is drawn without replacement from the MNL probability model on S_t, where $\pi_{i,t} \succ \pi_{j,t}, 1 \le i < j \le m$ means that the i-th is preferred to the j-th worker in π_t. More formally, the probability of a partial ranking $\pi_t \in \mathcal{P}^m_{S_t}$ under MNL model is

$$\mathbb{P}(\pi_t \in \mathcal{P}^m_{S_t} \mid S_t, \boldsymbol{\theta}) = \prod_{i=1}^{m} \frac{\theta_{\pi_{i,t}}}{\sum_{j \in S_t \setminus \{\theta_{\pi_{1,t}}, ..., \theta_{\pi_{i-1,t}}\}} \theta_{\pi_{j,t}}}. \quad (5)$$

In particular, preference feedback reduces to winner feedback, i.e., a single worker who is preferred to all other workers in S_t when $m = 1$ and a full performance ranking of workers in S_t when $m = L - 1$. Our model extends the original dueling bandits problem by simultaneously providing preference feedback on several workers rather than just two.

To model the quality of workers under preference feedback, we maintain a pairwise preference matrix $\boldsymbol{Q}$ where each element $q_{ij} = \frac{\theta_i}{\theta_i + \theta_j}$ denotes the pairwise probability of worker i being preferred to worker j. Then we introduce the Borda score [9] to measure the reliability of worker $i \in [K]$: $b_i = \frac{1}{K-1} \sum_{j \in [K]} q_{ij} \mathbb{1}_{\{j \neq i\}}$, which is the average preference probability that a worker is preferred to another worker in $[K]$ chosen uniformly at random. Analogous to the quantitative feedback setting, to incorporate both the reliability and stability of workers, we define the *Borda mean-variance* of worker i as $w^b_i = \rho b_i - \sigma^{2,b}_i$, where $\sigma^{2,b}_i$ denotes the variance of the Borda score b_i and $\rho > 0$ is a risk tolerance parameter.

Furthermore, we adopt the utility function $f(\cdot) > 0$ in Section II-A and reformulate the utility-based fairness constraint in (1) using Borda mean-variance:

$$\frac{p_i}{f(w^b_i)} = \frac{p_{i'}}{f(w^b_{i'})}, \quad \forall i \neq i', i, i' \in [K], \quad (6)$$

which ensures that the probability of selecting each worker is proportional to its utility. Similarly to Theorem II.3, there still exists a unique optimal fair policy $\boldsymbol{p}^{*,b} = \{p_1^{*,b}, p_2^{*,b}, ..., p_K^{*,b}\}$ in the form of

$$p_i^{*,b} = \frac{Lf(w_i^b)}{\sum_{i'=1}^{K} f(w_{i'}^b)}, \quad \forall i \in [K], \tag{7}$$

that satisfies the utility-based fairness constraints in (6). Moreover, we define the *Borda mean-variance regret* as the expected Borda mean-variance difference between the optimal policy and the deployed policy

$$\mathrm{BMR}_T = \sum_{t=1}^{T} \max\left\{ \sum_{i=1}^{K} p_i^{*,b} w_i^b - \sum_{i=1}^{K} p_{i,t} w_i^b, 0 \right\}, \tag{8}$$

and *Borda fairness regret* as the cumulative 1-norm distance between the optimal fair policy and the deployed policy

$$\mathrm{BFR}_T = \sum_{t=1}^{T} \sum_{i=1}^{K} |p_i^{*,b} - p_{i,t}|. \tag{9}$$

We aim to minimize both Borda mean-variance regret and Borda fairness regret. Note that Theorem II.4 and Remark II.5 still hold for the preference feedback setting as the Borda score can be interpreted as the expected reward of a worker. Therefore, we impose Assumption II.1 and Assumption II.2 on the utility function to design CMVB algorithms with sublinear Borda fairness regret.

B. Algorithm Description

To deal with preference feedback, we introduce a novel variant of MVFQ, named *Mean-Variance Fair learning with Preference feedback* (MVFP) algorithm, detailed in Algorithm 2. At each round t, we first construct an unbiased Borda score estimator for each worker $i \in [K]$. Denote $\boldsymbol{V}$ as a pairwise winning matrix, where each element $v_{ij,t}$ is the number of times worker i is preferred over j up to round t. Thanks to Lemma 1 in [10], we can extract pairwise comparisons from preference feedback and obtain the unbiased pairwise preference estimator $\hat{q}_{ij,t} = \frac{v_{ij,t}}{v_{ij,t}+v_{ji,t}}$ for $i, j \in [K]$ by treating each comparison independently as shown in line 4. Next, along with the selection policy $\boldsymbol{p}_t$, we construct an unbiased estimate $\tilde{b}_{i,t}$ of Borda score $b_{i,t}$ for each worker $i \in [K]$ (see line 5). Then we compute the empirical mean-variance estimate $\hat{w}_{i,t}^b$ of the Borda score using the time-average Borda score $\hat{b}_{i,t}$ and variance $\hat{\sigma}_{i,t}^{2,b}$. We further construct a confidence region $\mathcal{C}_t^b$ with UCB $u_{i,t}^b$ and LCB $l_{i,t}^b$ estimates of all workers in line 13. Similarly to the MVFQ algorithm, we identify a parameter $\widetilde{\boldsymbol{w}}_t^b = (\tilde{w}_{i,t}^b)_{i\in[K]}$ within the confidence region to maximize the expected mean-vairance subject to the utility-based fairness constraints. Finally, we update the selection policy $\boldsymbol{p}_{t+1}$ for the next round according to the optimal fair policy in (7).

We now state the sublinear expected Borda mean-variance regret and fairness regret upper bounds of MVFP algorithm in the following theorem.

Theorem III.1. *For MVFP in Algorithm 2, the expected Borda mean-variance regret is upper bounded by* $O\left(\alpha\rho L\sqrt{T\log T}\right)$ *and the expected Borda fairness regret of MVFP is upper bounded by* $O\left(\frac{\alpha\rho ML^2}{\lambda}\sqrt{T\log T}\right)$.

Algorithm 2 MVFP Algorithm

Input: $f(\cdot)$, T, L, K, $\rho > 0$, $\delta = \frac{1}{LT}$, $\alpha \geq \left(\frac{K^2-2K+L}{L^2-L}\right)^4$.
Init: Initial probability distribution: $p_{i,1} = 1/K$, $\forall i \in [K]$.
Pairwise winning matrix: $\boldsymbol{V} = [v_{ij,0}] \leftarrow [0]_{K\times K}$.

1: **for** $t = 1$ to T **do**
2: Select workers in $S_t = RRS(L, \boldsymbol{p}_t^b)$
3: Receive preference feedback π_t
4: For $i = 1, ..., m$, update

$$v_{\pi_{i,t}j,t} = v_{\pi_{i,t}j,t-1} + 1, \forall j \in S_t \backslash \{\pi_{1,t}, ..., \pi_{i,t}\}$$

5: Estimate Borda scores, for $i \in [K]$,

$$\tilde{b}_{i,t} = \frac{\mathbb{1}_{\{i\in S_t\}}}{(K-1)p_{i,t}} \sum_{j\in[K]} \frac{\mathbb{1}_{\{j\in S_t\}}\mathbb{1}_{\{j\neq i\}}}{p_{j,t}} \frac{v_{ij,t}}{v_{ij,t}+v_{ji,t}}$$

(Assuming $\frac{x}{0} = \frac{1}{2}, x \in \mathbb{R}$)
6: **for** $i \in [K]$ **do**
7: $\hat{b}_{i,t} = \sum_{\tau=1}^{t} \tilde{b}_{i,\tau}/t$
8: $\hat{\sigma}_{i,t}^{2,b} = \sum_{\tau=1}^{t} \left(\hat{b}_{i,t} - \tilde{b}_{i,\tau}\right)^2 /t$
9: $\hat{w}_{i,t}^b = \rho\hat{b}_{i,t} - \hat{\sigma}_{i,t}^2$
10: $u_{i,t}^b = \hat{w}_{i,t}^b + \alpha\left(\rho+3\right)\sqrt{\frac{2\log(4/\delta)}{t}}$
11: $l_{i,t}^b = \hat{w}_{i,t}^b - \alpha\left(\rho+3\right)\sqrt{\frac{2\log(4/\delta)}{t}}$
12: **end for**
13: $\mathcal{C}_t^b = \{\widetilde{\boldsymbol{w}}^b \in \mathbb{R}^K \mid \forall i \in [K],\ \widetilde{w}_i^b \in [l_{i,t}^b, u_{i,t}^b]\}$
14: $\widetilde{\boldsymbol{w}}_t^b = \arg\max_{\widetilde{\boldsymbol{w}}^b\in\mathcal{C}_t} \sum_{i\in[K]} \frac{Lf(\widetilde{w}_i^b)}{\sum_{i'\in[K]} f(\widetilde{w}_{i'}^b)} \widetilde{w}_i^b$
15: Compute $p_{i,t+1} = \frac{Lf(\widetilde{w}_{i,t}^b)}{\sum_{i'\in[K]} f(\widetilde{w}_{i',t}^b)}$ for $i \in [K]$
16: **end for**

Note that Theorem III.1 does not capture the impact of the number of ranked workers m on the regret bounds. Obtaining tight regret bounds that reflect this impact is challenging and remains an open problem. But the regret bounds in Theorem III.1 hold for all $1 \leq m \leq L-1$ and we demonstrate empirically in Section IV that the m only contributes a small additive penalty.

IV. Experiments

In this section, we conduct experiments to demonstrate the effectiveness of our algorithms.

We consider a mobile crowdsourcing system with $K = 6$ workers where the platform selects $L = 2$ workers at each round. The rewards of each worker follow a Gaussian distribution [1] with expectation in $\boldsymbol{\mu} = \{0.3, 0.5, 0.7, 0.9, 0.8, 0.6, 0.4\}$ and variance in $\boldsymbol{\sigma^2} = \{0.3, 0.5, 0.7, 0.9, 0.8, 0.6, 0.4\}$. We use the utility function $f(w) = \frac{K-L}{(1+e^{-cw})(L-1)} + 1$ which is adapted from Sigmoid function under Assumption II.1 and Assumption II.2. The parameter c controls the gradients of

[1] Although we assumed the reward distributions to be bounded in $[0, 1]$ in Theorem II.6, all the results could be extended to sub-Gaussian distributions.

the utility function. We set $c = 4$ and $\rho = 1$ in the following and all results are averaged over 100 runs.

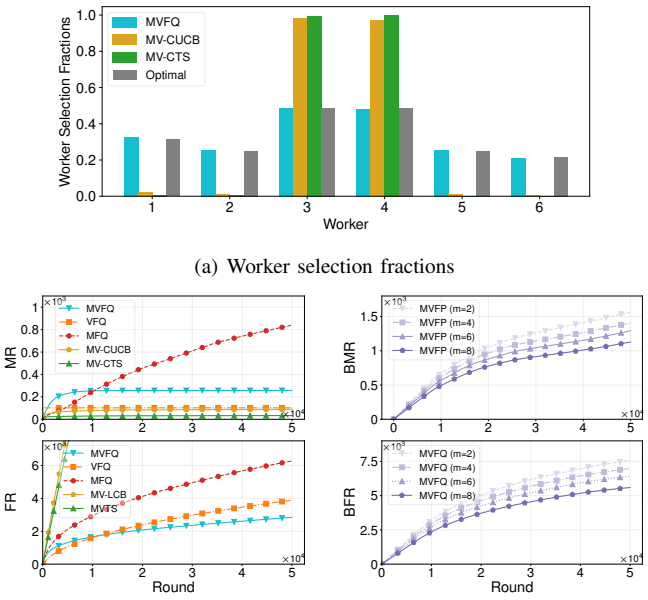
We first examine the fairness of different algorithms with quantitative feedback. For comparison, we implement two other mean-variance algorithms, MV-CUCB and MV-CTS which are adapted from MV-UCB [11] and MVTS [12], respectively, to accommodate the combinatorial selection structure. Figure 1(a) illustrates the average worker selection fractions of MV-CUCB, MV-CTS, the optimal fair policy, and our MVFQ algorithm. Each bar corresponds to the fraction of times a worker is chosen over $T = 5\times10^4$ rounds by a specific algorithm. As shown in Figure 1(a), MV-CUCB and MV-CTS are unfair by mainly selecting the workers (worker $3, 4$) with high mean-variances, neglecting the potential utility of other workers. In contrast, MVFQ algorithm can converge to the optimal fair policy. This observation shows the effectiveness of our algorithm in achieving utility-based fairness, ensuring that each worker receives a selection allocation proportional to its utility.

In Figure 1(b), it is evident that MV-CUCB and MV-CTS consistently exhibit smaller mean-variance regret and larger fairness regret when compared to MVFQ. This observation suggests that MV-CUCB and MV-CTS outperform MVFQ in mean-variance regret but substantially violate the utility-based fairness constraints, incurring higher fairness regret. Moreover, the mean-variance/fairness regrets of our algorithms increase sublinearly in T, aligning with the bounds we derived in Theorem II.6 Then we introduce two other algorithms, *Variance Fair learning with Quantitative feedback* (VFQ) and *Mean Fair learning with Quantitative feedback* (MFQ), to model extreme cases where ρ is small and large. Specifically, VFQ estimates the quality of workers based only on the expected reward, while MFQ relies on the variance of rewards. We show that the VFQ and MFQ algorithms still achieve both sublinear mean-variance and fairness regret.

Finally, we evaluate the performance of MVFP algorithm with preference feedback in Figure 1(c). With the same utility function, we consider the case $K = 10$, $L = 8$ and the MNL model with the parameters $\boldsymbol{\theta} = \{0.1, 0.3, 0.45, 0.6, 0.7, 0.95, 0.9, 0.75, 0.4, 0.15\}$. We show MVFP with different m achieve sublinear Borda mean-variance/fairness regret, which is consistent with the regret bounds we derived in Theorem III.1. As expected, the Borda mean-variance/fairness regret of MVFP scales down as m increases. This is because MVFP provides a more accurate estimate of Borda scores given more observations about relative preferences among workers.

V. Conclusion

In this paper, we develop a fair and risk-averse worker selection framework in mobile crowdsourcing via the CMVB model. We consider quantitative feedback and preference feedback settings, respectively and design novel algorithms that achieve both sublinear expected mean-variance regret and fairness regret. Our experimental results demonstrate that the proposed algorithms effectively balance utility-based fairness and mean-variance optimization of worker selection under both types of feedback.



(a) Worker selection fractions

(b) MVFQ algorithm

(c) MVFP algorithm

Fig. 1. Experiment results of MVFQ and MVFP algorithms.

References

[1] N. Zhao, Y. Pei, Y.-C. Liang, and D. Niyato, "A deep reinforcement learning-based contract incentive mechanism for mobile crowdsourcing networks," *IEEE Transactions on Vehicular Technology*, 2023.

[2] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson, "Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones," in *Proceedings of the 7th ACM conference on embedded networked sensor systems*, 2009, pp. 85–98.

[3] J. Xu, Z. Luo, C. Guan, D. Yang, L. Liu, and Y. Zhang, "Hiring a team from social network: Incentive mechanism design for two-tiered social mobile crowdsourcing," *IEEE Transactions on Mobile Computing*, vol. 22, no. 8, pp. 4664–4681, 2022.

[4] Y. Zhang, Q. Liu, H. Wang, D. Chen, and K. Han, "Crowdsourcing live high definition map via collaborative computation in automotive edge computing," *IEEE Transactions on Vehicular Technology*, 2024.

[5] X. Liu, M. Derakhshani, L. Mihaylova, and S. Lambotharan, "Risk-aware contextual learning for edge-assisted crowdsourced live streaming," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 3, pp. 740–754, 2022.

[6] M. Mansoury, B. Mobasher, and H. van Hoof, "Exposure-aware recommendation using contextual bandits," in *5th FAccTRec Workshop: Responsible Recommendation*. Association for Computing Machinery (ACM), 2022.

[7] S. Wang and Z. Shao, "Green dueling bandits," in *ICC 2023 - IEEE International Conference on Communications*, 2023, pp. 5129–5134.

[8] R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan, "Dependent rounding and its applications to approximation algorithms," *Journal of the ACM (JACM)*, vol. 53, no. 3, pp. 324–360, 2006.

[9] K. Jamieson, S. Katariya, A. Deshpande, and R. Nowak, "Sparse dueling bandits," in *Artificial Intelligence and Statistics*. PMLR, 2015, pp. 416–424.

[10] A. Saha and A. Gopalan, "Pac battling bandits in the plackett-luce model," in *Algorithmic Learning Theory*. PMLR, 2019, pp. 700–737.

[11] S. Vakili and Q. Zhao, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1093–1111, 2016.

[12] Q. Zhu and V. Tan, "Thompson sampling algorithms for mean-variance bandits," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 599–11 608.

[13] J. Bretagnolle and C. Huber, "Estimation des densités: risque minimax," *Séminaire de probabilités de Strasbourg*, vol. 12, pp. 342–363, 1978.

APPENDIX

A. Proof of Theorem II.3

Proof. According to (1), the optimal fair policy $\boldsymbol{p}^*$ satisfies the following merit-based fairness constraints:

$$\frac{p_i^*}{f(w_i)} = \frac{p_{i'}^*}{f(w_{i'})}, \quad \forall i \neq i', i, i' \in [K], \tag{10}$$

which correspond to $K-1$ linearly independent equations of $\boldsymbol{p}^*$. Moreover, because only L arms can be selected at each round, there is an additional linear equation $\sum_{i=1}^{K} p_i^* = L$ that is linearly independent of the other $K-1$ ones. Then we have K linearly independent equations on K unknowns in $\boldsymbol{p}^* = \{p_1^*, p_2^*, ..., p_K^*\}$. Therefore, the optimal fair policy $\boldsymbol{p}^*$ is unique. By solving this system of linear equations, we have

$$p_i^* = \frac{Lf(w_i)}{\sum_{i'=1}^{K} f(w_{i'})}, \quad \forall i \in [K]. \tag{11}$$

This completes the proof of Theorem II.3. □

B. *Proof of Theorem II.4*

Proof. We first prove that *the lower bound on fairness regret is linear without Assumption II.1 (i)* by constructing two CMVB instances and a 1-Lipschitz utility function $f(\cdot)$. For any bandit algorithm, we show that the sum of the expected fairness regrets of the two CMVB instances increases linearly in T. Consequently, we conclude that any bandit algorithm will incur a linear regret in T for at least one of the two CMVB instances.

The two instances can be defined as $x^1 = (\nu_1^1, \nu_2^1, \nu_3^1)$ and $x^2 = (\nu_1^2, \nu_2^2, \nu_3^2)$, where ν_i is the reward distribution of worker i. Each instance consists of three workers and the crowdsourcing platform selects a subset S_t of $L = 2$ workers at each round t. We assume that the reward of each worker in the two instances follows a Bernoulli distribution. The mean-variances of three workers in the first instance are $3\eta, 2\eta, 2\eta$, and the mean-variances of three workers in the second instance are $2\eta, 2\eta, 2\eta$, where $\eta \in (0, 1/3]$. The utility function $f(\cdot)$ is defined as an identity function. i.e., $f(w) = w$. Therefore, the optimal fair policy for the first instance is $\boldsymbol{p}^{*,1} = \{6/7, 4/7, 4/7\}$, and the optimal fair policy for the second instance is $\boldsymbol{p}^{*,2} = \{2/3, 2/3, 2/3\}$. For any bandit algorithm $\mathcal{A}$, the platform selects the workers stochastically according to a selection policy $\boldsymbol{p}_t$ at each round t based on the history $\mathcal{H}_t$, which consists of all the previous selection vectors, selected worker sets, and received feedback. We have $i \sim \boldsymbol{p}_t$, $r_{i,t} \sim \nu_i$ for $i \in S_t$. Then we can derive the lower bound of the expected fairness regret for the two instances as follows. For the first instance x^1, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^1\right] &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\left|p_{1,t} - \frac{6}{7}\right| + \left|p_{2,t} - \frac{4}{7}\right| + \left|p_{3,t} - \frac{4}{7}\right|\right)\right] \\
&\geq \mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T} p_{1,t} - \frac{6}{7}\right| + \left|\frac{1}{T}\sum_{t=1}^{T} p_{2,t} - \frac{4}{7}\right| + \left|\frac{1}{T}\sum_{t=1}^{T} p_{3,t} - \frac{4}{7}\right|\right] \\
&= 2\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T} p_{1,t} - \frac{6}{7}\right|\right].
\end{aligned} \tag{12}$$

Similarly, for the second instance x^2, we can derive the fairness regret lower bound as follows.

$$\mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^2\right] \geq 2\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T} p_{1,t} - \frac{2}{3}\right|\right]. \tag{13}$$

We consider a worker selection trace during the T rounds as $h = (\boldsymbol{p}_1, S_1, \boldsymbol{r}_1, ..., \boldsymbol{p}_T, S_T, \boldsymbol{r}_T)$, where $\boldsymbol{r}_t := (r_{i,t})_{i\in[K]}$. Denote $\mathbb{H}^1, \mathbb{H}^2$ as the distributions of h for first CMVB instance x^1, x^2 using the algorithm $\mathcal{A}$, respectively. Then we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^1\right] + \mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^2\right] &\geq \frac{4}{21}\mathbb{P}^1\left(\frac{1}{T}\sum_{t=1}^{T} p_{1,t} \leq \frac{16}{21}\right) + \frac{4}{21}\mathbb{P}^2\left(\frac{1}{T}\sum_{t=1}^{T} p_{1,t} > \frac{16}{21}\right) \\
&\overset{(a)}{\geq} \frac{2}{21}\exp\left(-\mathrm{KL}\left(\mathbb{H}^1, \mathbb{H}^2\right)\right),
\end{aligned} \tag{14}$$

where (a) follows from the Bretagnolle-Huber inequality [13]. We can derive the upper bound of the KL divergence between $\mathbb{H}^1$ and $\mathbb{H}^2$, $\mathrm{KL}\left(\mathbb{H}^1, \mathbb{H}^2\right)$, as follows:

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{H}^1,\mathbb{H}^2\right) &= \mathbb{E}_{h\sim\mathbb{H}^1}\left[\log\frac{\mathbb{H}^1(h)}{\mathbb{H}^2(h)}\right] \le \mathbb{E}_{h\sim\mathbb{H}^1}\left[\sum_{t=1}^{T}\sum_{i\in S_t}\log\frac{\nu_i^1\left(r_{i,t}\right)}{\nu_i^2\left(r_{i,t}\right)}\right] = \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{p}_t\sim\mathcal{A}^1}\mathbb{E}_{i\sim\boldsymbol{p}_t}\left[\mathrm{KL}\left(\nu_i^1,\nu_i^2\right)\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{p}_t\sim\mathcal{A}^1}\left[p_{1,t}\,\mathrm{KL}\left(\nu_1^1,\nu_1^2\right)\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{p}_t\sim\mathcal{A}^1}\left[p_{1,t}\left(3\eta\log\frac{3}{2}+(1-3\eta)\log\frac{1-3\eta}{1-2\eta}\right)\right] \\
&\le 3T\eta\log\frac{3}{2},
\end{aligned} \tag{15}$$

where $\boldsymbol{p}_t \sim \mathcal{A}^1$ means that $\boldsymbol{p}_t$ is sampled from the process of the algorithm $\mathcal{A}^1$ applied to the first CMVB instance.

Combining (15) with (14) and setting $\eta = 1/3T$, we have

$$\mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^1\right] + \mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^2\right] \ge \frac{2}{21}\exp\left(-3T\eta\log\frac{3}{2}\right) \ge 0.06, \tag{16}$$

which implies that at least one of the two CMVB instances incurs linear expected fairness regret. Therefore, we infer that at least one of the two CMVB instances incurs linear expected fairness regret.

Next, we prove that *the lower bound on fairness regret is linear without Assumption II.2* by constructing two CMVB instances and a utility function $f(\cdot)$ where $\min_w f(w) = 1$. For any bandit algorithm, we demonstrate that the expected fairness regrets of the two CMVB instances grow linearly with respect to T. Thus, any bandit algorithm will result in linear regret in T for at least one of the two CMVB instances.

The two instances can be defined as $x^1 = (\nu_1^1, \nu_2^1, \nu_3^1)$ and $x^2 = (\nu_1^2, \nu_2^2, \nu_3^2)$, where ν_i is the reward distribution of worker i. Each instance consists of three workers and the platform selects a subset S_t of $L = 2$ workers at each round t. We assume that the reward of each worker in the two instances follows a Bernoulli distribution. The mean-variances of three workers in the first instance are $2\eta-1, \eta-1, \eta-1$, and the mean-variances of three workers in the second instance are $\eta-1, \eta-1, \eta-1$, where $\eta \in (0, 1/2)$. We use the utility function $f(\cdot)$ with the form

$$f(w) = \begin{cases} 1 & w \le -1 \\ M(w+1)+1 & w > -1 \end{cases}$$

where $M > 0$ is a positive constant to be defined later. Therefore, the optimal fair policy for the first instance is $\boldsymbol{p}^{*,1} = \{(4\eta M+2)/(4\eta M+3), (2\eta M+2)/(4\eta M+3), (2\eta M+2)/(4\eta M+3)\}$, and the optimal fair policy for the second instance is $\boldsymbol{p}^{*,2} = \{2/3, 2/3, 2/3\}$. For any bandit algorithm $\mathcal{A}$, the platform selects the workers with a probabilistic selection vector $\boldsymbol{p}_t$ at each round t based on the observation and decision history $\mathcal{H}_t$. Then we have $i \sim \boldsymbol{p}_t$, $r_{i,t} \sim \nu_i$ for $i \in S_t$.

For any algorithm, we can lower bound the expected fairness regret for the two instances as follows. For the first instance x^1, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^1\right] &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\left|p_{1,t}-\frac{4\eta M+2}{4\eta M+3}\right| + \left|p_{2,t}-\frac{2\eta M+2}{4\eta M+3}\right| + \left|p_{3,t}-\frac{2\eta M+2}{4\eta M+3}\right|\right)\right] \\
&\ge \mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T}p_{1,t}-\frac{4\eta M+2}{4\eta M+3}\right| + \left|\frac{1}{T}\sum_{t=1}^{T}p_{2,t}-\frac{2\eta M+2}{4\eta M+3}\right| + \left|\frac{1}{T}\sum_{t=1}^{T}p_{3,t}-\frac{2\eta M+2}{4\eta M+3}\right|\right] \\
&\ge 2\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T}p_{1,t}-\frac{4\eta M+2}{4\eta M+3}\right|\right].
\end{aligned} \tag{17}$$

Similarly, for the second instance x^2, we can derive the fairness regret lower bound as follows,

$$\mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^2\right] \ge 2\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T}p_{1,t}-\frac{2}{3}\right|\right]. \tag{18}$$

We consider a worker selection trace during the T rounds as $h = (\boldsymbol{p}_1, S_1, \boldsymbol{r}_1, ..., \boldsymbol{p}_T, S_T, \boldsymbol{r}_T)$. Denote $\mathbb{H}^1$ as the distribution of h when the algorithm $\mathcal{A}$ is applied to the first MAB instance x^1, while $\mathbb{H}^2$ as the distribution of h when the algorithm $\mathcal{A}$ is applied to the second MAB instance x^2. Then we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^1\right] + \mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^2\right] &\ge \frac{4M\eta}{12M\eta+9}\mathbb{P}^1\left(\frac{1}{T}\sum_{t=1}^{T}p_{1,t} \le \frac{10M\eta+6}{12M\eta+9}\right) + \frac{4M\eta}{12M\eta+9}\mathbb{P}^2\left(\frac{1}{T}\sum_{t=1}^{T}p_{1,t} > \frac{10M\eta+6}{12M\eta+9}\right) \\
&\overset{(a)}{\ge} \frac{2M\eta}{12M\eta+9}\exp\left(-\mathrm{KL}\left(\mathbb{H}^1,\mathbb{H}^2\right)\right),
\end{aligned} \tag{19}$$

where (a) follows from the Bretagnolle-Huber inequality [13]. Then we can upper bound the KL divergence $\mathrm{KL}\left(\mathbb{H}^1, \mathbb{H}^2\right)$ between $\mathbb{H}^1$ and $\mathbb{H}^2$ as follows,

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{H}^1, \mathbb{H}^2\right) &= \mathbb{E}_{h\sim\mathbb{H}^1}\left[\log\frac{\mathbb{H}^1(h)}{\mathbb{H}^2(h)}\right] \leq \mathbb{E}_{h\sim\mathbb{H}^1}\left[\sum_{t=1}^{T}\sum_{a\in A_t}\log\frac{\nu_i^1\left(r_{i,t}\right)}{\nu_i^2\left(r_{i,t}\right)}\right] = \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{p}_t\sim\mathcal{A}^1}\mathbb{E}_{i\sim\boldsymbol{p}_t}\left[\mathrm{KL}\left(\nu_i^1, \nu_i^2\right)\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{p}_t\sim\mathcal{A}^1}\left[p_{1,t}\,\mathrm{KL}\left(\nu_1^1, \nu_1^2\right)\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{p}_t\sim\mathcal{A}^1}\left[p_{1,t}\left(2\eta\log 2 + (1-2\eta)\log\frac{1-2\eta}{1-\eta}\right)\right] \\
&\leq 2T\eta\log 2,
\end{aligned} \tag{20}$$

where $\boldsymbol{p}_t \sim \mathcal{A}^1$ means that $\boldsymbol{p}_t$ is sampled from the process of the algorithm $\mathcal{A}^1$ applied to the first MAB instance.

According to (19) and (20), set $\eta = \frac{1}{2T}, M = T$, we have

$$\mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^1\right] + \mathbb{E}\left[\frac{1}{T}\mathrm{FR}_T^2\right] \geq \frac{2M\eta}{12M\eta+9}\exp\left(-2T\eta\log 2\right) \geq 0.03, \tag{21}$$

which implies that at least one of the two CMVB instances incurs linear expected fairness regret.

This completes the proof of Theorem II.4. □

C. *Proof of Theorem II.6*

Proof. We first prove the expected mean-variance upper bound and then prove the expected fairness regret upper bound of MVFQ.

Part 1: Proof of the Expected Mean-variance Regret Upper Bound of MVFQ

We first prove the following lemmas that will be used in the proofs.

Lemma A.1. *For any $\delta \in (0,1)$, with probability at least $1-\frac{\delta}{2}$, $\forall t > \lceil K/L \rceil$, $i \in [K]$, $r_{i,t} \in [0,1]$, the expected mean-variance vector $\boldsymbol{w} \in \mathcal{C}_t$.*

Proof. According to Hoeffding's inequality, for $t \in [T]$, $i \in [K]$, with probability at least $1 - \frac{\delta}{4KT}$,

$$|\hat{\mu}_{i,t} - \mu_i| \leq \sqrt{\frac{\log(8KT/\delta)}{2n_{i,t}}}. \tag{22}$$

Let $\hat{\mu}_{i,t}^{(2)} = \frac{\sum_{\tau=1}^{t}\mathbb{1}_{\{i\in S_\tau\}}r_{i,\tau}^2}{n_{i,t}}$ and $\mu_i^{(2)} = \mathbb{E}\left[r_{i,t}^2\right]$. Similarly, with probability at least $1 - \frac{\delta}{4KT}$,

$$\left|\hat{\mu}_{i,t}^{(2)} - \mu_i^{(2)}\right| \leq \sqrt{\frac{\log(8KT/\delta)}{2n_{i,t}}}. \tag{23}$$

Note that $\sigma_i^2 = \mu_i^{(2)} - \mu_i^2$ and $\hat{\sigma}_{i,t}^2 = \sum_{\tau=1}^{t}\frac{\mathbb{1}_{\{i\in S_\tau\}}(r_{i,\tau}-\hat{\mu}_{i,t})^2}{n_{i,t}} = \sum_{\tau=1}^{t}\frac{\mathbb{1}_{\{i\in S_\tau\}}r_{i,\tau}^2}{n_{i,t}} - \hat{\mu}_{i,t}^2 = \hat{\mu}_{i,t}^{(2)} - \hat{\mu}_{i,t}^2$. Then we have

$$\begin{aligned}
\left|\hat{\sigma}_{i,t}^2 - \sigma_i^2\right| &\leq \left|\hat{\mu}_{i,t}^{(2)} - \mu_i^{(2)}\right| + \left|\hat{\mu}_{i,t}^2 - \mu_i^2\right| \\
&\leq \left|\hat{\mu}_{i,t}^{(2)} - \mu_i^{(2)}\right| + \left(\hat{\mu}_{i,t} + \mu_i\right)\left|\hat{\mu}_{i,t} - \mu_i\right| \\
&\leq \left|\hat{\mu}_{i,t}^{(2)} - \mu_i^{(2)}\right| + 2\left|\hat{\mu}_{i,t} - \mu_i\right|
\end{aligned} \tag{24}$$

Fianlly, with probability at least $1 - \frac{\delta}{2KT}$,

$$\begin{aligned}
|\widehat{w}_{i,t} - w_i| &= \left|(\rho\hat{\mu}_{i,t} - \hat{\sigma}_{i,t}^2) - (\rho\mu_i - \sigma_i^2)\right| \leq \rho\left|\hat{\mu}_{i,t} - \mu_i\right| + \left|\hat{\sigma}_{i,t}^2 - \sigma_i^2\right| \\
&\leq (\rho+2)\left|\hat{\mu}_{i,t} - \mu_i\right| + \left|\hat{\mu}_{i,t}^{(2)} - \mu_i^{(2)}\right| \\
&\leq (\rho+3)\sqrt{\frac{\log(8KT/\delta)}{2n_{i,t}}}.
\end{aligned} \tag{25}$$

Using union bound over all round t and i, with probability at least $1-\frac{\delta}{2}$, $\forall t > \lceil K/L \rceil, a \in [K]$, $w_i \in [l_{i,t}, u_{i,t}]$.

This completes the proof of Lemma A.1. □

Lemma A.2. *For any $\delta \in (0,1)$, with probability at least $1-\frac{\delta}{2}$, it holds that*

$$\left| \sum_{t=\lceil K/L \rceil+1}^{T} \mathbb{E}_{i\sim \boldsymbol{p}_t} \left[\sqrt{\frac{1}{n_{i,t}}} \right] - \sum_{t=\lceil K/L \rceil+1}^{T} \sum_{i\in S_t} \sqrt{\frac{1}{n_{i,t}}} \right| \leq L\sqrt{2T\log\frac{4}{\delta}}. \tag{26}$$

Proof. We first construct a martingale difference sequence

$$\sum_{i\in S_t} \sqrt{\frac{1}{n_{i,t}}} - \mathbb{E}_{I\sim \boldsymbol{p}_t} \left[\sqrt{\frac{1}{n_{i,t}}} \right]. \tag{27}$$

Then we have

$$\left| \sum_{i\in S_t} \sqrt{\frac{1}{n_{i,t}}} - \mathbb{E}_{i\sim \boldsymbol{p}_t} \left[\sqrt{\frac{1}{n_{i,t}}} \right] \right| < L. \tag{28}$$

By Azuma-Hoeffding's inequality, with probability at least $1-\frac{\delta}{2}$, we have

$$\left| \sum_{t=\lceil K/L \rceil+1}^{T} \mathbb{E}_{i\sim \boldsymbol{p}_t} \left[\sqrt{\frac{1}{n_{i,t}}} \right] - \sum_{t=\lceil K/L \rceil+1}^{T} \sum_{i\in S_t} \sqrt{\frac{1}{n_{i,t}}} \right| \leq L\sqrt{2T\log\frac{4}{\delta}}. \tag{29}$$

This completes the proof of Lemma A.2. □

Based on the lemmas above, with probability at least $1-\delta$, the mean-variance regret can be bounded as follows:

$$\begin{aligned}
\mathrm{MR}_T &= \sum_{t=1}^{T} \max\left\{ \sum_{i=1}^{K} p_i^* w_i - \sum_{i=1}^{K} p_{i,t} w_i, 0 \right\} \\
&\overset{(a)}{\leq} \left(\frac{K}{L} + 1 \right) L + \sum_{t=\lceil \frac{K}{L} \rceil+1}^{T} \sum_{i=1}^{K} (p_{i,t}\tilde{w}_{i,t} - p_{i,t} w_i) \\
&= K + L + \sum_{t=\lceil \frac{K}{L} \rceil+1}^{T} \sum_{i=1}^{K} p_{i,t}(\tilde{w}_{i,t} - \hat{w}_{i,t} + \hat{w}_{i,t} - w_i) \\
&\overset{(b)}{\leq} K + L + \sum_{t=\lceil \frac{K}{L} \rceil+1}^{T} \sum_{i=1}^{K} p_{i,t} 2\left(\rho+3\right) \sqrt{\frac{\log(8KT/\delta)}{2n_{i,t}}} \\
&= K + L + (\rho+3)\sqrt{2\log\frac{8KT}{\delta}} \sum_{t=\lceil \frac{K}{L} \rceil+1}^{T} \mathbb{E}_{i\sim \boldsymbol{p}_t} \left[\sqrt{\frac{1}{n_{i,t}}} \right] \\
&\overset{(c)}{\leq} K + L + (\rho+3)\sqrt{2\log\frac{8KT}{\delta}} \left(L\sqrt{2T\log\frac{4}{\delta}} + \sum_{t=\lceil K/L \rceil+1}^{T} \sum_{i\in S_t} \sqrt{\frac{1}{n_{i,t}}} \right) \\
&\leq K + L + (\rho+3)\sqrt{2\log\frac{8KT}{\delta}} \left(L\sqrt{2T\log\frac{4}{\delta}} + K\int_0^{\frac{LT}{K}} \sqrt{\frac{1}{x}} dx \right) \\
&\leq K + L + (\rho+3)\sqrt{2\log\frac{8KT}{\delta}} \left(L\sqrt{2T\log\frac{4}{\delta}} + 2\sqrt{LKT} \right),
\end{aligned} \tag{30}$$

where (a) is from Line 13 in Algorithm 1, (b) is from Lemma A.1 and (c) is from Lemma A.2. Setting $\delta = \frac{1}{LT}$, the expected mean-variance regret can be upper bounded as

$$\begin{aligned}
\mathbb{E}\left[\mathrm{MR}_T\right] &\leq K + L + (\rho+3)\sqrt{2\log\frac{8KT}{\delta}} \left(L\sqrt{2T\log\frac{4}{\delta}} + 2\sqrt{LKT} \right) + LT\delta \\
&\leq K + L + (\rho+3)\sqrt{4\log(8LKT)} \left(L\sqrt{2T\log(4LT)} + 2\sqrt{LKT} \right) + 1 \\
&= O\left(\rho\sqrt{LKT\log T} \right).
\end{aligned} \tag{31}$$

This completes the proof of the mean-variance regret upper bound of MVFQ.

Part 2: Proof of the Expected Fairness Regret Upper Bound of MVFQ

For any $\delta \in (0,1)$, with probability $1-\delta$,

$$\begin{aligned}
&\sum_{i=1}^{K}|p_{i,t}-p_i^*| = \sum_{i=1}^{K}\left|\frac{Lf(\tilde{w}_{i,t})}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t})}-\frac{Lf(w_i)}{\sum_{i'=1}^{K}f(w_{i'})}\right| \\
&= \sum_{i=1}^{K}\frac{L\left|f(\tilde{w}_{i,t})\sum_{i'=1}^{K}f(w_{i'})-f(w_i)\sum_{i'=1}^{K}f(\tilde{w}_{i',t})\right|}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t})\sum_{i'=1}^{K}f(w_{i'})} \\
&= \sum_{i=1}^{K}\frac{L\left|f(\tilde{w}_{i,t})\sum_{i'=1}^{K}f(w_{i'})-f(w_i)\sum_{i'=1}^{K}f(w_{i'})+f(w_i)\sum_{i'=1}^{K}f(w_{i'})-f(w_i)\sum_{i'=1}^{K}f(\tilde{w}_{i',t})\right|}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t})\sum_{i'=1}^{K}f(w_{i'})} \\
&\le \frac{L\sum_{i=1}^{K}|f(\tilde{w}_{i,t})-f(w_i)|\sum_{i'=1}^{K}f(w_{i'})+L\sum_{i=1}^{K}f(w_i)\sum_{i'=1}^{K}|f(w_{i'})-f(\tilde{w}_{i,t})|}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t})\sum_{i'=1}^{K}f(w_{i'})} \\
&= \frac{2L\sum_{i=1}^{K}\frac{f(\tilde{w}_{i,t})}{f(\tilde{w}_{i,t})}|f(\tilde{w}_{i,t})-f(w_i)|}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t})} \\
&\overset{(a)}{\le} \sum_{i=1}^{K}\frac{2Mp_{i,t}L}{\lambda}\left(\tilde{w}_{i,t}-w_i\right) \\
&\overset{(b)}{\le} \sum_{i=1}^{K}\frac{4ML\left(\rho+3\right)p_{i,t}}{\lambda}\sqrt{\frac{\log(8KT/\delta)}{2n_{i,t}}} \\
&= \frac{4ML\left(\rho+3\right)}{\lambda}\sqrt{\frac{1}{2}\log\frac{8KT}{\delta}}\mathbb{E}_{i\sim\boldsymbol{p}_t}\left[\sqrt{\frac{1}{n_{i,t}}}\right],
\end{aligned} \tag{32}$$

where (a) follows from the Assumption II.1 (i) and Assumption II.2, (b) follows from Lemma A.1. When $T > K$, with probability at least $1-\delta$, the fairness regret can be upper bounded as follows:

$$\begin{aligned}
\mathrm{FR}_T &= \sum_{t=1}^{T}\sum_{i=1}^{K}|p_i^*-p_{i,t}| \le \left(\frac{K}{L}+1\right)L + \sum_{t=\lceil\frac{K}{L}\rceil+1}^{T}\sum_{i=1}^{K}|p_i^*-p_{i,t}| \\
&\le K+L+\frac{4ML\left(\rho+3\right)}{\lambda}\sqrt{\frac{1}{2}\log\frac{8KT}{\delta}}\sum_{t=\lceil\frac{K}{L}\rceil+1}^{T}\mathbb{E}_{i\sim\boldsymbol{p}_t}\left[\sqrt{\frac{1}{n_{i,t}}}\right] \\
&\overset{(a)}{\le} K+L+\frac{4ML\left(\rho+3\right)}{\lambda}\sqrt{\frac{1}{2}\log\frac{8KT}{\delta}}\left(L\sqrt{2T\log\frac{4}{\delta}}+2\sqrt{LKT}\right),
\end{aligned} \tag{33}$$

where (a) follows from the result in (30). Furthermore, setting $\delta = \frac{1}{LT}$, the expected fairness regret can be upper bounded as

$$\begin{aligned}
\mathbb{E}\left[\mathrm{FR}_T\right] &\le K+L+\frac{4ML\left(\rho+3\right)}{\lambda}\sqrt{\frac{1}{2}\log\frac{8KT}{\delta}}\left(L\sqrt{2T\log\frac{4}{\delta}}+2\sqrt{LKT}\right)+LT\delta \\
&\le K+L+\frac{4ML\left(\rho+3\right)}{\lambda}\sqrt{\log(8LKT)}\left(L\sqrt{2T\log(4LT)}+2\sqrt{LKT}\right)+1 \\
&= O\left(\frac{\rho ML}{\lambda}\sqrt{LKT\log T}\right).
\end{aligned} \tag{34}$$

This completes the proof of fairness regret upper bound of MVFQ.

Combining Part 1 and Part 2 of the proof, we complete the proof of Theorem II.6.

□

*D. **Proof of Theorem III.1***

Proof. We first prove the expected mean-variance upper bound and then prove the expected fairness regret upper bound of MVFP.

Part 1: Proof of the Expected Mean-variance Regret Upper Bound of MVFP

Before presenting our theoretical analysis and results, we need two technical lemmas that will be used in the proofs. We first prove $\tilde{b}_{i,t}$ is an unbiased estimate of the Borda score b_i in the following lemma.

Lemma A.3. *At any round t, for all $i \in [K]$, it holds that $\mathbb{E}[\tilde{b}_{i,t}] = b_i$.*

Proof. Note that

$$\begin{aligned}
\mathbf{E}\left[\tilde{b}_{i,t}\right] &= \mathbf{E}\left[\frac{\mathbb{1}_{\{i\in S_t\}}\mathbb{1}_{\{j\neq i\}}}{p_{i,t}(K-1)}\sum_{j\in[K]}\frac{\mathbb{1}_{\{j\in S_t\}}}{p_{j,t}}\frac{v_{ij,t}}{v_{ij,t}+v_{ji,t}}\right] = \frac{1}{K-1}\mathbf{E}\left[\sum_{j\in[K]}\frac{\mathbb{1}_{\{i\in S_t\}}\mathbb{1}_{\{j\in S_t\}}\mathbb{1}_{\{j\neq i\}}}{p_{i,t}p_{j,t}}\frac{v_{ij,t}}{v_{ij,t}+v_{ji,t}}\right] \\
&= \frac{1}{K-1}\mathbf{E}_{\mathcal{H}_{t-1}}\left[\mathbf{E}\left[\frac{\mathbb{1}_{\{i\in S_t\}}}{p_{i,t}}\sum_{j\in[K]}\frac{\mathbb{1}_{\{j\in S_t\}}\mathbb{1}_{\{j\neq i\}}}{p_{j,t}}\frac{v_{ij,t}}{v_{ij,t}+v_{ji,t}}\middle|\mathcal{H}_{t-1}\right]\right] \\
&= \frac{1}{K-1}\mathbf{E}_{\mathcal{H}_{t-1}}\left[\mathbf{E}_i\left[\frac{\mathbb{1}_{\{i\in S_t\}}}{p_{i,t}}\sum_{j\in[K]}\mathbf{E}_j\left[\frac{\mathbb{1}_{\{j\in S_t\}}\mathbb{1}_{\{j\neq i\}}}{p_{j,t}}\mathbf{E}\left[\frac{v_{ij,t}}{v_{ij,t}+v_{ji,t}}\right]\right]\middle|\mathcal{H}_{t-1}\right]\right] \\
&= \frac{1}{K-1}\mathbf{E}_{\mathcal{H}_{t-1}}\left[\mathbf{E}_i\left[\frac{\mathbb{1}_{\{i\in S_t\}}}{p_{i,t}}\sum_{j\in[K]}\mathbf{E}_j\left[\frac{\mathbb{1}_{\{j\in S_t\}}\mathbb{1}_{\{j\neq i\}}q_{ij}}{p_{j,t}}\right]\middle|\mathcal{H}_{t-1}\right]\right] \\
&= \frac{1}{K-1}\mathbf{E}_{\mathcal{H}_{t-1}}\left[\mathbf{E}_i\left[\frac{\mathbb{1}_{\{i\in S_t\}}}{p_{i,t}}\sum_{j\in[K]}\sum_{j'=1}^{K}\frac{\mathbb{1}_{\{j=j'\}}\mathbb{1}_{\{j\in S_t\}}\mathbb{1}_{\{j\neq i\}}q_{ij}p_{j',t}}{p_{j,t}}\middle|\mathcal{H}_{t-1}\right]\right] \\
&= \frac{1}{K-1}\mathbf{E}_{\mathcal{H}_{t-1}}\left[\mathbf{E}_i\left[\frac{\mathbb{1}_{\{i\in S_t\}}}{p_{i,t}}\sum_{j\in[K]}q_{ij}\mathbb{1}_{\{j\neq i\}}\middle|\mathcal{H}_{t-1}\right]\right] \\
&= \frac{1}{K-1}\mathbf{E}_{\mathcal{H}_{t-1}}\left[\sum_{i'=1}^{K}\frac{\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{i\in S_t\}}p_{i',t}}{p_{i,t}}\sum_{j\in[K]}q_{ij}\mathbb{1}_{\{j\neq i\}}\right] \\
&= \frac{1}{K-1}\sum_{j\in[K]}q_{ij}\mathbb{1}_{\{j\neq i\}} = b_i,
\end{aligned} \tag{35}$$

which concludes the proof. □

Next, we prove that the constructed confidence region $\mathcal{C}_t^b$ contains the actual Borda score mean-variance with high probability in the following lemma.

Lemma A.4. *For any $\delta \in (0,1)$, $\alpha \geq \left(\frac{K^2-2K+L}{L^2-L}\right)^4$, with probability at least $1-\delta$, the expected Borda score mean-variance vector $\boldsymbol{w}^b \in \mathcal{C}_t^b$.*

Proof. By Lemma A.3, we have $\mathbb{E}\left[\tilde{b}_{i,t}\right] = b_i$. We can construct a martingale difference sequence $\tilde{b}_{i,\tau} - b_i$ for $\tau \leq t$. Note that $\left|\tilde{b}_{i,\tau} - b_i\right| \leq \left(\frac{K^2-2K+L}{L^2-L}\right)^2$ since $p_{i,t} \geq \frac{L^2-L}{K^2-2K+L}$ under Assumption II.1. Then by Azuma-Hoeffding's inequality, with probability at least $1-\delta/2$, we have

$$\left|\hat{b}_{i,t} - b_i\right| = \left|\frac{1}{t}\sum_{\tau=1}^{t}\tilde{b}_{i,\tau} - \frac{1}{t}\sum_{\tau=1}^{t}b_i\right| \leq \left(\frac{K^2-2K+L}{L^2-L}\right)^2\sqrt{\frac{2\log 4/\delta}{t}} \leq \sqrt{\frac{2\alpha\log 4/\delta}{t}}. \tag{36}$$

Let $\hat{b}_{i,t}^{(2)} = \frac{\sum_{\tau=1}^{t}\tilde{b}_{i,\tau}^2}{t}$ and $b_i^{(2)} = \mathbb{E}\left[\tilde{b}_{i,t}^2\right]$. Similarly, by Azuma-Hoeffding's inequality, with probability at least $1-\delta/2$,

$$\left|\hat{b}_{i,t}^{(2)} - b_i^{(2)}\right| = \left|\frac{1}{t}\sum_{\tau=1}^{t}\tilde{b}_{i,\tau}^{(2)} - \frac{1}{t}\sum_{\tau=1}^{t}b_i^{(2)}\right| \leq \left(\frac{K^2-2K+L}{L^2-L}\right)^4\sqrt{\frac{2\log 4/\delta}{t}} \leq \alpha\sqrt{\frac{2\log 4/\delta}{t}}. \tag{37}$$

Note that $\sigma_i^2 = b_i^{(2)} - b_i^2$ and $\hat{\sigma}_{i,t}^2 = \sum_{\tau=1}^{t}\frac{\left(\hat{b}_{i,t}-\tilde{b}_{i,\tau}\right)^2}{t} = \sum_{\tau=1}^{t}\frac{\tilde{b}_{i,\tau}^2}{t} - \hat{b}_{i,t}^2 = \hat{b}_{i,t}^{(2)} - \hat{b}_{i,t}^2$. Then we have

$$\begin{aligned}
\left|\hat{\sigma}_{i,t}^2 - \sigma_i^2\right| &\leq \left|\hat{b}_{i,t}^{(2)} - b_i^{(2)}\right| + \left|\hat{b}_{i,t}^2 - b_i^2\right| \\
&\leq \left|\hat{b}_{i,t}^{(2)} - b_i^{(2)}\right| + \left(\hat{b}_{i,t} + b_i\right)\left|\hat{b}_{i,t} - b_i\right| \\
&\leq \left|\hat{b}_{i,t}^{(2)} - b_i^{(2)}\right| + \left(\left(\frac{K^2-2K+L}{L^2-L}\right)^2 + 1\right)\left|\hat{b}_{i,t} - b_i\right| \\
&\leq \left|\hat{b}_{i,t}^{(2)} - b_i^{(2)}\right| + (\sqrt{\alpha}+1)\left|\hat{b}_{i,t} - b_i\right|
\end{aligned} \tag{38}$$

Finally, with probability at least $1-\delta$, $\alpha \geq \left(\frac{K^2-2K+L}{L^2-L}\right)^4$,

$$\begin{aligned}\left|\widehat{w}_{i,t}^b - w_i^b\right| &= \left|(\rho\hat{b}_{i,t} - \hat{\sigma}_{i,t}^2) - (\rho b_i - \sigma_i^2)\right| \leq \rho\left|\hat{b}_{i,t} - b_i\right| + \left|\hat{\sigma}_{i,t}^2 - \sigma_i^2\right| \\ &\leq \rho\left|\hat{b}_{i,t} - b_i\right| + \left|\hat{b}_{i,t}^{(2)} - b_i^{(2)}\right| + (\sqrt{\alpha}+1)\left|\hat{b}_{i,t} - b_i\right| \\ &\leq \left(\sqrt{\alpha}\,(\rho+1) + 2\alpha\right)\sqrt{\frac{2\log 4/\delta}{t}} \\ &\leq \alpha\,(\rho+3)\sqrt{\frac{2\log 4/\delta}{t}}.\end{aligned} \tag{39}$$

This completes the proof of Lemma A.4.

Based on the lemmas above, with probability at least $1-\delta$, the mean-variance regret can be bounded as follows:

$$\begin{aligned}\mathrm{BMR}_T &= \sum_{t=1}^{T}\max\left\{\sum_{i=1}^{K} p_i^* w_i^b - \sum_{i=1}^{K} p_{i,t} w_i^b, 0\right\} \\ &\overset{(a)}{\leq} \sum_{t=1}^{T}\sum_{i=1}^{K}(p_{i,t}\tilde{w}_{i,t}^b - p_{i,t} w_i^b) \\ &= \sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t}(\tilde{w}_{i,t}^b - \hat{w}_{i,t}^b + \hat{w}_{i,t}^b - w_i^b) \\ &\overset{(b)}{\leq} \sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t} 2\left(\sqrt{\alpha}\,(\rho+1) + 2\alpha\right)\sqrt{\frac{2\log 4/\delta}{t}} \\ &= 2\left(\sqrt{\alpha}\,(\rho+1) + 2\alpha\right)\sqrt{2\log\frac{4}{\delta}}\sum_{t=1}^{T}\mathbb{E}_{i\sim\boldsymbol{p}_t}\left[\sqrt{\frac{1}{t}}\right] \\ &\overset{(c)}{\leq} 4\left(\sqrt{\alpha}\,(\rho+1) + 2\alpha\right) L\sqrt{2T\log\frac{4}{\delta}}\end{aligned} \tag{40}$$

Setting $\delta = \frac{1}{LT}$, the expected mean-variance regret can be upper bounded as

$$\begin{aligned}\mathbb{E}\left[\mathrm{BMR}_T\right] &\leq 4\left(\sqrt{\alpha}\,(\rho+1) + 2\alpha\right) L\sqrt{2T\log\frac{4}{\delta}} + LT\delta \leq 4\left(\sqrt{\alpha}\,(\rho+1) + 2\alpha\right) L\sqrt{2T\log(4LT)} + 1 \\ &= O\left(\alpha\rho L\sqrt{T\log T}\right).\end{aligned} \tag{41}$$

This completes the proof of the mean-variance regret upper bound of MVFP.

□

Part 2: Proof of the Expected Fairness Regret Upper Bound of MVFP

For any $\delta \in (0,1)$, with probability $1-\delta$,

$$\begin{aligned}
&\sum_{i=1}^{K}|p_{i,t}-p_i^*| = \sum_{i=1}^{K}\left|\frac{Lf(\tilde{w}_{i,t}^b)}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t}^b)}-\frac{Lf(w_i^b)}{\sum_{i'=1}^{K}f(w_{i'}^b)}\right| \\
&=\sum_{i=1}^{K}\frac{L\left|f(\tilde{w}_{i,t}^b)\sum_{i'=1}^{K}f(w_{i'}^b)-f(w_i^b)\sum_{i'=1}^{K}f(\tilde{w}_{i',t}^b)\right|}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t}^b)\sum_{i'=1}^{K}f(w_{i'}^b)} \\
&=\sum_{i=1}^{K}\frac{L\left|f(\tilde{w}_{i,t}^b)\sum_{i'=1}^{K}f(w_{i'}^b)-f(w_i^b)\sum_{i'=1}^{K}f(w_{i'}^b)+f(w_i^b)\sum_{i'=1}^{K}f(w_{i'}^b)-f(w_i^b)\sum_{i'=1}^{K}f(\tilde{w}_{i',t}^b)\right|}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t}^b)\sum_{i'=1}^{K}f(w_{i'}^b)} \\
&\le \frac{L\sum_{i=1}^{K}\left|f(\tilde{w}_{i,t}^b)-f(w_i^b)\right|\sum_{i'=1}^{K}f(w_{i'}^b)+L\sum_{i=1}^{K}f(w_i^b)\sum_{i'=1}^{K}\left|f(w_{i'}^b)-f(\tilde{w}_{i,t}^b)\right|}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t}^b)\sum_{i'=1}^{K}f(w_{i'}^b)} \\
&=\frac{2L\sum_{i=1}^{K}\frac{f(\tilde{w}_{i,t}^b)}{f(\tilde{w}_{i,t}^b)}\left|f(\tilde{w}_{i,t}^b)-f(w_i^b)\right|}{\sum_{i'=1}^{K}f(\tilde{w}_{i',t}^b)} \\
&\overset{(a)}{\le}\sum_{i=1}^{K}\frac{2Mp_{i,t}L}{\lambda}\left(\tilde{w}_{i,t}^b-w_i^b\right) \\
&\overset{(b)}{\le}\sum_{i=1}^{K}\frac{4ML\alpha\left(\rho+3\right)p_{i,t}}{\lambda}\sqrt{\frac{2\log 4/\delta}{t}} \\
&=\frac{4ML\alpha\left(\rho+3\right)}{\lambda}\sqrt{2\log\frac{4}{\delta}}\mathbb{E}_{i\sim\boldsymbol{p}_t}\left[\sqrt{\frac{1}{t}}\right],
\end{aligned} \tag{42}$$

where (a) follows from the Assumption II.1 (i) and Assumption II.2, (b) follows from Lemma A.1. When $T > K$, with probability at least $1-\delta$, the fairness regret can be upper bounded as follows:

$$\begin{aligned}
\mathrm{BFR}_T &= \sum_{t=1}^{T}\sum_{i=1}^{K}|p_i^*-p_{i,t}| \le \sum_{t=\lceil\frac{K}{L}\rceil+1}^{T}\sum_{i=1}^{K}|p_i^*-p_{i,t}| \\
&\le \frac{4ML\left(\sqrt{\alpha}\left(\rho+1\right)+2\alpha\right)}{\lambda}\sqrt{2\log\frac{4}{\delta}}\sum_{t=1}^{T}\mathbb{E}_{i\sim\boldsymbol{p}_t}\left[\sqrt{\frac{1}{t}}\right] \\
&\overset{(a)}{\le} \frac{8ML^2\left(\sqrt{\alpha}\left(\rho+1\right)+2\alpha\right)}{\lambda}\sqrt{2T\log\frac{4}{\delta}},
\end{aligned} \tag{43}$$

where (a) follows from the result in (30). Furthermore, setting $\delta = \frac{1}{LT}$, the expected fairness regret can be upper bounded as

$$\begin{aligned}
\mathbb{E}\left[\mathrm{BFR}_T\right] &\le \frac{8ML^2\left(\sqrt{\alpha}\left(\rho+1\right)+2\alpha\right)}{\lambda}\sqrt{2T\log\frac{4}{\delta}}+LT\delta \\
&\le \frac{8ML^2\left(\sqrt{\alpha}\left(\rho+1\right)+2\alpha\right)}{\lambda}\sqrt{2T\log(4LT)}+1 \\
&= O\left(\frac{\alpha\rho ML^2}{\lambda}\sqrt{T\log T}\right).
\end{aligned} \tag{44}$$

This completes the proof of fairness regret upper bound of MVFP.

Combining Part 1 and Part 2 of the proof, we complete the proof of Theorem III.1.

□