Preference-centric Bandits in Wireless Communications: Theory and Applications

Meltem Tatlı **Ali Tajer**
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
{tatlim, tajera}@rpi.edu

Abstract

Data-driven solutions to resource allocation in wireless communications are becoming increasingly more pervasive to complement legacy model-based solutions. This paper studies resource allocation when the transmitter is oblivious to channel models and channel instantaneous realizations. Decision-making under a finite number of choices is modeled by multi-armed bandits (MABs), which effectively balance the rate of learning the channels (exploration) and their use in the meantime (exploitation). Despite that natural fit, some key metrics of interest (e.g., outage probability) cannot be directly specified by the average-based reward functions that MAB algorithms rely on. This paper adopts a broader notion of reward that subsumes the conventional average-based reward and accommodates other choices that can precisely specify the desired metrics of interest in communications. This leads to different principles for designing bandit algorithms. This framework is presented in a general form, and its specific applications to optimizing outage and latency are investigated.

1 Introduction

Machine learning (ML) can complement legacy model-based resource allocation solutions in wireless communications by leveraging data to capture nuances that models cannot. Model-based methods inevitably adopt simplifying assumptions that, on the one hand, enable analytical tractability at the cost of not fully capturing the true complexity of radio frequency environments. With networks evolving towards massive connectivity, extreme mobility, and heterogeneous services, the gap between stylized models and operational reality is widening. ML-based approaches can potentially overcome such impediments by learning directly from data, bypassing the need for tractable models [11, 16, 26]. Nevertheless, despite such potential, designing only black-box approaches raises questions about their robustness, reliability, and interpretability. Hence, it is imperative to align the learned strategies with performance metrics that matter in practice. This tension signifies the need for a comprehensive perspective that systematically leverages the flexibility of ML while grounding it in principled performance objectives.

The central challenge in resource allocation in wireless systems is that network conditions are rarely known in advance, and often fluctuate unpredictably. Therefore, algorithms that interact with the network and learn the environment are valuable. Online learning methods, and multi-armed bandit (MAB) algorithms in particular, embody this principle by balancing how to explore the new strategies with the information gleaned over time to converge to optimal performance. Such algorithms require *repeated* interaction with the environment, and they are natural choices for decisions that must be made repeatedly under uncertainty. Some examples include channel selection [2, 9, 12], power

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

allocation [15, 24], rate adjustment [18, 17], and beam selection [3, 25, 23, 10, 6]. In contrast to the static optimization approaches that rely on having a fixed and accurate model, MAB-based approaches progressively improve their performance as they gather more, making them suitable for environments with mobility, interference variability, and diverse quality-of-service requirements.

While MAB algorithms provide a natural framework for adaptive decision-making, there is a fundamental disconnect between what bandit theory traditionally optimizes and what communication systems actually require. Specifically, conventional MABs are formalized to identify action with the highest **expected reward**. This aligns with some metrics in communication (e.g., expected throughput/rate); some critical measures in wireless networks go beyond average measures and are highly sensitive to the tail behavior of the statistical models for the metric. For instance, ultra-reliable low-latency communication (URLLC) requires stringent guarantees on rare-event latencies rather than mean delay [5]; outage probability quantifies the likelihood that a user's rate falls below a critical threshold; and fairness objectives ensure that a few strong users do not monopolize resources while others suffer degraded service. In massive machine-type communications (mMTC), reliability of sporadic transmissions is key, while in enhanced mobile broadband (eMBB), extreme-rate events can shape user experience. Such metrics cannot be adequately captured by expectations alone, underscoring the need for a fresh perspective on bandit algorithms that systematically integrate risk sensitivity, preference criteria, and tail behavior.

The limitations of mean-based objectives have led to growing interest in extending bandit models to account for risk-sensitive decision-making. Recent work has explored criteria such as variance, quantiles, and conditional value-at-risk, providing mechanisms to incorporate sensitivity to rare but impactful events [8, 4, 19, 20, 13, 14, 7]. More recently, a framework that integrates diverse notions of risks and preferences was introduced in [21, 22], which moves beyond expectations and designs algorithms tailored to arbitrary performance metrics (PM). This perspective provides the versatility needed to align online learning with requirements on reliability, latency, and fairness. In this paper, we adopt and extend this framework to formalize, and address communication-driven performance metrics.

This shift toward PMs has significant implications in algorithm design. In conventional formulations, the optimal strategy is to identify and converge to a single best arm, corresponding to the action with the highest expected reward. However, when optimizing more general PMs, depending on the geometry of the chosen PM, the optimal solution may no longer be a solitary arm but rather a mixture policy that randomizes across multiple arms. This requires a fresh look at algorithm design to account for the possibility of mixture policies explicitly. On a technical level, it introduces several challenges.

The paper is organized as follows. Section 2 formalizes the PM-centric framework for bandit design. Section 3 applies the framework to outage problem and demonstrates its use across different settings. Section 4 presents an upper confidence bound (UCB)-based bandit algorithm and establishes its performance guarantees. Section 5 concludes the paper. Proofs and empirical evaluations are presented in the supplementary material.

2 PM-centric Bandit Framework

Stochastic bandit model. Consider a K-armed stochastic MAB, where the observations of arm $i \in [K] \triangleq \{1, \cdots, K\}$ are generated according to a distribution with the cumulative distribution function (CDF) $\mathbb{F}_i$. We assume that the distributions $\{F_i : i \in [K]\}$ belong to the parametric family of distributions parameterized by their mean values. We denote the mean under $\mathbb{F}_i$ by θ_i.

At time $t \in \mathbb{N}$, a policy π selects an arm $A_t \in [K]$ and the arm generates a stochastic sample X_t. Denote the sequence of actions and observations that policy π generates up to time $t \in \mathbb{N}$ by $\mathcal{X}_t \triangleq (X_1, \cdots, X_t)$ and $\mathcal{A}_t^\pi \triangleq (A_1, \cdots, A_t)$, respectively. We denote the sequence of independent and identically distributed (i.i.d.) rewards generated by arm $i \in [K]$ up to time $t \in \mathbb{N}$ is denoted by $\mathcal{X}_t(i) \triangleq \{X_t : A_t = i\}$ and define $\tau_t^\pi(i) \triangleq |\mathcal{X}_t(i)|$.

Preference metric. To any given CDF $\mathbb{Q}$ we assign a PM as the *signed Choquet* integral, as follows.

$$U_h(\mathbb{Q}) \triangleq \int_{-\infty}^{0} \Big(h(1 - \mathbb{Q}(x)) - h(1)\Big)\mathrm{d}x + \int_{0}^{\infty} h(1 - \mathbb{Q}(x))\mathrm{d}x \ , \tag{1}$$

where $h : [0,1] \to [0,1]$ is called the *distortion function*, whose role is to distort the tail distribution $1 - \mathbb{Q}(x)$ in any desired form. It can, for instance, emphasize or de-emphasize certain pieces of the tail distribution as desired. The choice of h enables a high degree of generality in selecting the PMs. Corresponding to any given policy π at time t, the overall preference associated with the sequence of arm selections $\mathcal{A}_t^\pi$ by a policy π is specified by

$$U_h\Big(\sum_{s=1}^{t}\sum_{i\in[K]} \frac{1}{t}\mathbb{1}\{A_s = i\}\ \mathbb{F}_i\Big) = U_h\Big(\sum_{i\in[K]} \frac{1}{t}\tau_t^\pi(i)\ \mathbb{F}_i\Big)\ . \tag{2}$$

This is a strict generalization of the notion of rewards in the canonical MAB framework, which can be recovered from the PM induced by setting the distortion function to $h(u) = u$.

Oracle Policy. Next, we specify an oracle that serves as a benchmark for policy performance. Such an oracle accurately identifies the optimal sequence of arm selections $\{A_t : t \in \mathbb{N}\}$. Given the structure in (2), designing an oracle policy is equivalent to determining the optimal mixing of the CDFs, where the mixing coefficients up to time t are $\left\{\frac{1}{t}\tau_t^\pi(i) : i \in [K]\right\}$. For a bandit instance $\boldsymbol{\nu} \triangleq (\mathbb{F}_1, \cdots, \mathbb{F}_K)$, the vector of optimal mixture coefficients is denoted by

$$\boldsymbol{\alpha}^\star \ \in \ \underset{\boldsymbol{\alpha}\in\Delta^{K-1}}{\arg\max}\ U_h\Big(\sum_{i\in[K]} \alpha(i)\ \mathbb{F}_i\Big)\ , \tag{3}$$

where Δ^{K-1} denotes a K-dimensional simplex. When the mixture coefficients vector $\boldsymbol{\alpha}^\star$ is 1-sparse, the optimal policy becomes a solitary (single-arm) policy, and otherwise it is a mixture policy.

Mixture-centric Regret. Depending on the distortion function h, the optimal sampling rule may be a *mixture* of arm distributions. Concrete examples of this behavior are presented in Section 3). This motivates: for a bandit instance $\boldsymbol{\nu}$, oracle policy's $\boldsymbol{\alpha}^\star$ in (3), given policy π and horizon T, define the regret as the gap between the PM of the oracle policy and that of π, i.e.,

$$\mathfrak{R}_{\boldsymbol{\nu}}^\pi(T) \ \triangleq \ U_h\left(\sum_{i\in[K]} \alpha^\star(i)\ \mathbb{F}_i\right) - \mathbb{E}_{\boldsymbol{\nu}}^\pi\left[U_h\Big(\sum_{i\in[K]} \frac{1}{T}\boldsymbol{\tau}_T^\pi\ \mathbb{F}_i\Big)\right]\ . \tag{4}$$

3 Applications to Wireless Communication Metrics

The PM-centric framework provides a natural way to capture certain key performance measures in wireless communications. Classical metrics such as *outage capacity* fit seamlessly into this formulation, as they directly reflect tail behavior rather than averages. The framework also offers a principled foundation for a broader class of metrics for which prior attempts have sought to move past mean-based evaluation but lack a unifying structure. Examples include metrics related to *latency violations* or *fairness across users*. By embedding them within a PM-centric formalism, these metrics can be systematically defined, analyzed, and optimized using learning algorithms, providing both rigor and generality. In this sense, the framework not only subsumes some established communication metrics but also extends them into a cohesive methodology.

We now describe a class of communication problems that can be framed by bandit settings and the PM-centric framework. Consider a communication system, in which the performance under a given decision (e.g., power level, transmission rate, or beamforming vector) is represented by a random variable X. The source of randomness may stem from fading channels, interference, traffic arrivals, or other stochastic radio frequency environments. In these problems, let $[K] \in \{1, \dots\}$ denote a finite set of possible actions, where each action corresponds to a design choice such as a discrete transmission rate, a beam selection, or a power allocation level. Selecting action $A_t \in [K]$ at time t induces a random performance outcome X_t, drawn from an unknown distribution $\mathbb{F}_{A_t}$. Hence, each action A_t can be modeled as an *arm* in a K-armed bandit problem, with the outcome distribution capturing the variability of the wireless system under that choice.

In canonical bandit settings, the performance of an arm is summarized by its expected value, $\mathbb{E}[X_{A_t}]$, and the objective of a bandit algorithm is to identify the arm a^* with the largest expectation. This corresponds to optimizing mean performance, which aligns with conventional metrics such as average rate or average spectral efficiency. However, some metrics cannot be captured by expectation and require a proper PM that accounts for their tail behavior. For instance, in *outage capacity,* the interest is in the probability that the achieved rate falls below a threshold; in *latency guarantees*, the objective is to control the probability that packet latency exceeds a bound, and in *fairness*, some of the widely-used metrics capture the minimum or target quantiles of user-specific random variables.

In the rest of this section, we investigate the outage capacity as a representative and important example that (i) satisfies the PM framework, and (ii) optimizing which requires a mixture policy with a performance that dominates any solitary policy.

3.1 Outage Optimization: A PM-centric Model

Outage capacity is the maximum transmission rate that can be sustained with high reliability in the presence of random channel variations. Given a target reliability level γ, for a single-user Gaussian channel, the outage capacity is specified by

$$C_{\text{out}}(\gamma) = \sup\left\{R : \mathbb{P}\left(C < R\right) \leq \gamma\right\} \quad \text{where} \quad C \triangleq \log(1 + \text{SNR})\ . \tag{5}$$

This represents the largest rate R such that the probability of the instantaneous rate falling below R does not exceed γ. A higher outage capacity means that the system can reliably support higher data rates despite fading or interference. The next lemma shows that the randomness model that optimizes the outage capacity is the same model that optimizes the PM specified in (1) with the choice of $h(u) = \min\left\{\frac{u}{1-\gamma}, 1\right\}$.

Lemma 1 *Consider a single-user Gaussian channel whose capacity C has CDF $\mathbb{Q} \in \mathcal{Q}$, where $\mathcal{Q}$ is any desired set of CDFs. Then*

$$\mathbb{Q}^\star \triangleq \arg\max_{\mathbb{Q}\in\mathcal{Q}} C_{\text{out}}(\gamma) = \arg\max_{\mathbb{Q}\in\mathcal{Q}} U_h(\mathbb{Q})\ , \quad \textit{for} \quad h(u) = \min\left\{\frac{u}{1-\gamma}, 1\right\}\ . \tag{6}$$

Hence, optimizing outage capacity can be expressed as optimizing a proper PM U_h. A key subtlety arises, however, when the set of CDFs $\mathcal{Q}$ is not convex. In such cases, a randomized combination of the distributions in $\mathcal{Q}$ (e.g., via time-sharing) might lead to an outage capacity that is strictly larger than $\max_{\mathbb{Q}\in\mathcal{Q}} C_{\text{out}}(\gamma)$. Operationally, this corresponds to adopting a *mixture policy* that properly randomizes between transmission strategies that induce different distributions for the capacity (e.g., power levels, coding rates, or beamforming choices). This property is formalized in the next lemma.

Lemma 2 *Suppose the CDF set $\mathcal{Q}$ is not convex. When the distortion function is concave (but not affine), there might exists a mixture of CDFs in $\mathcal{Q}$, denoted by $\mathbb{Q}_{\text{mix}}$ for which $U_h(\mathbb{Q}_{\text{mix}}) > \max_{\mathbb{Q}\in\mathcal{Q}} U_h(\mathbb{Q})$.*

An important implication of this property for the MAB problems is that when $\mathcal{Q}$ is the set of K CDFs representing K arms, a mixture of arm selection *might* provide a PM U_h that is strictly larger than that of any single arm. We specialize this to the outage optimization problem in the next subsection through a numerical example and then a more general model.

3.2 Rate Selection for Outage Optimization: Two-state Channels

Consider a two-state single-user channel. The channel is in the *weak* state S_{w} when $\text{SNR} = 3$ and it is in the *strong* state S_{s} when $\text{SNR} = 5$. Accordingly, the channel capacities for these two steps are $C_{\text{w}} = 2$ and $C_{\text{s}} = 3$ bits/sec/Hz, respectively. When the transmission rate is r, the effective rate will be either r (no outage) or 0 (outage). Denote the probability of outage in this channel at rate r by $p_{\text{out}}(r)$. Consider six possible transmission rates $r \in \{1, \frac{3}{2}, 2, \frac{5}{2}, 3, \frac{7}{2}\}$. Define $X \in \{0, r\}$ as the random variable representing the effective rate when we transmit at r. Hence, X has distribution $r \cdot \text{Bern}(1 - p_{\text{out}}(r))$, whose CDF we denote by $\mathbb{F}_r$, with mean parameter is $\theta = r(1 - p_{\text{out}}(r))$. With the choice of h in (6) we have

$$U_h(\mathbb{F}_r) \ = \ r \cdot \min\left\{\frac{1 - p_{\text{out}}(r)}{1-\gamma}, 1\right\}\ . \tag{7}$$

When the channel is strong with probability ζ, we have $p_{\text{out}}(r) = \zeta u(r-2) + (1-\zeta)u(r-3)$, where u is the unit step function. Hence, for the set of possible rates, $\gamma = 0.1$, and $\zeta = 0.6$ we have

$$U_h(\mathbb{F}_1) = 1,\;\; U_h(\mathbb{F}_{\frac{3}{2}}) = \frac{3}{2}\;,\;\; U_h(\mathbb{F}_2) = 2,\;\; U_h(\mathbb{F}_{\frac{5}{2}}) = \frac{5}{3},\;\; U_h(\mathbb{F}_3) = 2,\;\; U_h(\mathbb{F}_{\frac{7}{2}}) = 0\;. \tag{8}$$

Therefore, the best PM (outage capacity) is associated with the rates $r \in \{2,3\}$. When the transmitter is oblivious to all channel states and probability models, it can adopt a bandit framework to identify the best rate by representing each rate with one arm. A canonical bandit algorithm can efficiently identify the arms associated with the rates $r \in \{2,3\}$. However, we can show that a mixture policy can lead to a strictly higher PM (outage capacity). Specifically, consider selecting rates 2 and 3 for %75 and %25 of the time, respectively, i.e., $\boldsymbol{\alpha} = [0,0,0.75,0,0.25,0]$. This mixture has possible rates $r \in \{0,2,3\}$ and has the following mixture CDF.

$$\mathbb{F}_{\text{mix}}(r) = \sum_{i\in[K]} \alpha(i)\mathbb{F}_i = 0.1u(r) + 0.75u(r-2) + 0.15u(r-3)\;. \tag{9}$$

The utility associated with this mixture model is $U_h(\mathbb{F}_{\text{mix}}) \approx 2.16$, which is strictly larger than all the individual utilities in (8).

3.3 Rate Selection for Outage Optimization: Finite-state Channels

In this subsection, we formalize the observations from the previous toy example for a general finite-state channel. Consider a K-state block-fading channel with states $S_t \in [K]$ drawn i.i.d. across episodes with unknown probabilities $(q_1,\ldots,q_K)$, where $\sum_{i=1}^K q_i = 1$. Each state i has an associated SNR γ_i and capacity $c_i \triangleq \log_2(1+\gamma_i)$. Without loss of generality, we assume $c_1 < c_2 < \cdots < c_K$. The action set (arms) is a set of rates $\mathcal{R}$, and we define

$$\mathcal{R} \;\triangleq\; \{r_i = c_i :\; i \in [K]\}\;. \tag{10}$$

The probability of successful transmission at rate r_i in state $S \in [K]$ is given by

$$\bar{p}_{\text{out}}(r_i) \;\triangleq\; \mathbb{P}\{c_s \geq r_i\} \;=\; \sum_{\ell=i}^{K} q_\ell\;, \tag{11}$$

which is strictly decreasing in i, i.e., $1 = \bar{p}_{\text{out}}(r_1) > \bar{p}_{\text{out}}(r_2) > \cdots > \bar{p}_{\text{out}}(r_K) > 0$. The per-episode achieved rate when selecting arm i is

$$X_i \;=\; r_i \cdot \mathbf{1}\{c_i \geq r_i\} \in \{0, r_i\}\;. \tag{12}$$

Let us denote the CDF of X_i by $\mathbb{F}_i$, with mean parameter is $\theta = r_i(\bar{p}_{\text{out}}(r))$. It can be readily verified that for a pure policy that always elects arm i we have

$$U_h(\mathbb{F}_i) \;=\; r_i \cdot \min\left\{\frac{\bar{p}_{\text{out}}(r_i)}{1-\gamma},\, 1\right\}\;. \tag{13}$$

Theorem 1 (Outage optimization – finite-state channel) *For the rate-selection problem specified, let $\alpha(i)$ be the fraction of episodes using rate r_i. Adopt the convention $\bar{p}_{\text{out}}(r_{K+1}) = 0$.*

1. *If $\min_i \bar{p}_{\text{out}}(r_i) \geq 1-\gamma$, the optimal policy is any solitary arm $m^\star \in \arg\max_{i\in[K]} r_i$.*

2. *If $\max_i \bar{p}_{\text{out}}(r_i) < 1-\gamma$, the optimal policy is any solitary arm $m^\star \in \arg\max_{i\in[K]} \bar{p}_{\text{out}}(r_i) r_i$.*

3. *If $\bar{p}_{\text{out}}(r_k) \geq 1-\gamma > \bar{p}_{\text{out}}(r_{k+1})$ for some $k \in \{1,\ldots,K-1\}$, define*

$$m^* \in \arg\max_{i\geq k+1}\{r_i \bar{p}_{\text{out}}(r_i)\}\;, \tag{14}$$

the optimal policy mixes r_k (reliable) and $r_{m^\star}$ (risky) with weights

$$\alpha^\star(k) \;=\; \frac{1-\gamma-\bar{p}_{\text{out}}(r_{m^\star})}{\bar{p}_{\text{out}}(r_k)-\bar{p}_{\text{out}}(r_{m^\star})}, \qquad \alpha^\star(m^\star) \;=\; 1-\alpha^\star(k)\;. \tag{15}$$

4 Param-UCB-M Algorithm

Two horizon-aware algorithms based on the explore-then-commit and UCB principles for subgaussian distributions were introduced in [21, 22]. In this work, within our framework, we adapt the approach to parametric distributions. We refer to this method as Parametric-Upper Confidence Bound-Mixture (Param-UCB-M) algorithm.

4.1 Algorithm Details

The overall procedure of Param-UCB-M is summarized in Algorithm 1 and consists of 4 key steps: discretization, forced exploration, parameter estimation, and under-sampling. Let N denote the last time instant of forced exploration, given by

$$N \triangleq K \left\lceil \frac{1}{4}\rho\varepsilon T \right\rceil \ . \tag{16}$$

Algorithm 1 Param-UCB-M

```
Input: Exploration rate $\rho$, discretization level $\varepsilon$, horizon $T$
Sample each arm $\frac{N}{K}$ times where $N \triangleq K \lceil \frac{1}{4}\rho\varepsilon T \rceil$ and obtain observation sequences $\mathcal{X}_N(1), \cdots, \mathcal{X}_N(K)$
Initialize: $\tau_N(i) = \frac{N}{K} \quad \forall i \in [K]$, empirical parameters $\theta_{N(1)}, \cdots, \theta_N(K)$, confidence sets $\mathcal{C}_N(1), \cdots \mathcal{C}_N(K)$
for $t = N+1, \cdots, T$ do
    Compute the optimistic estimate $a_t$ according to (19)
    Select an arm $A_t$ via undersampling and obtain reward $X_t$
    Update the empirical parameter $\theta_t(i)$ according to (17)
    Update the confidence set $\mathcal{C}_t(A_t)$ according to (18)
end for
```

Discretization. We uniformly discretize each coordinate of Δ^{K-1} into length $\varepsilon \in (0, 1]$ intervals and denote this simplex by Δ_ε^{K-1}. Numerical computation has finite precision, therefore, it is impractical to estimate mixing coefficients with finer granularity.

Forced Exploration. For horizon T and exploration parameter $\rho \in (0, 1)$, each arm is sampled uniformly for $\frac{1}{4}\rho\varepsilon T$ times.

Parameter estimation. We define the number of sample of arm i until time t as $\tau_t(i) \triangleq |\mathcal{X}_t(i)|$. We denote the CDF associated with the estimators as $\mathbb{F}(\cdot; \theta_t(i))$ for $\theta_t(i)$ where the estimator is defined as

$$\theta_t(i) = \frac{1}{\tau_t(i)} \sum_{s\in[t]} X_s(i)\mathbb{1}\{A_s = i\} \tag{17}$$

Accordingly, the confidence interval for $\theta(i)$ at time t, denoted $C_t(i)$ is defined as

$$C_t(i) \triangleq \left\{ \eta \in \Theta : |\theta_t(i) - \eta| \le \sqrt{\frac{2\log T}{\tau_t(i)}} \right\} . \tag{18}$$

The confidence bounds are used to determine the mixing coefficient. Formally, the mixing coefficient with the highest UCB value is selected according to

$$\mathbf{a}_t \in \arg\max_{\mathbf{a}\in\Delta_\varepsilon^{K-1}} \max_{\kappa(i)\in\mathcal{C}_t(i), \forall i\in[K]} U_h\left(\sum_{i\in[K]} \alpha(i)\mathbb{F}(\cdot; \kappa(i)) \right) . \tag{19}$$

Undersampling. At time t, after selecting a mixing coefficient, the algorithm decides on which arm to sample based on undersampling, i.e., $A_t = \max_{i\in[K]}\{ta_t(i) - \tau_t(i)\}$ [22]

4.2 Performance Guarantees

In the MAB setting, the performance of an algorithm is typically evaluated through its regret. We first state our technical assumptions, then establish an upper bound, deferring detailed expressions. In these assumptions $\|\cdot\|_W$ denotes the $1-$Wasserstein distance.

Assumption 1 (Hölder continuity of U_h) *There exist $q \in (0, 1]$ and $\mathcal{L}_h$ such that for any $\mathbb{F}, \mathbb{G} \in \mathcal{F}$ $|U_h(\mathbb{F}) - U_h(\mathbb{G})| \leq \mathcal{L}_h \|\mathbb{F} - \mathbb{G}\|_W^q$.*

Assumption 2 (Hölder continuity of CDFs) *There exist $p \in (0, 1]$ and $\mathcal{L}_\theta$ such that for any $\mathbb{F}_i, \mathbb{F}_j \in \mathcal{F}$, $\|\mathbb{F}_i - \mathbb{F}_j\|_W \leq L_\theta|\theta_i - \theta_j|^p$.*

Assumption 3 (Bounded range) *There exists $B_h > 0$ such that for any $\mathbb{F} \in \mathcal{F}$, $|U_h(\mathbb{F})| \leq B_h$.*

We define the best discrete mixing coefficient as

$$\mathbf{a}^{(1)} \in \arg\max_{\mathbf{a}\in\Delta_\varepsilon^{K-1}} U_h(\sum_{i\in[K]} a(i)\mathbb{F}_i) \tag{20}$$

and the sub-optimality gap $\delta_{12}(\varepsilon)$ associated with it as

$$\delta_{12}(\varepsilon) \triangleq \min_{\mathbf{a}\in\Delta_\varepsilon^{K-1}\setminus\{\mathbf{a}^{(1)}\}} \{U_h(\sum_{i\in[K]} a^{(1)}(i)\mathbb{F}_i) - U_h(\sum_{i\in[K]} a(i)\mathbb{F}_i)\}. \tag{21}$$

Based on the sub-optimality gap, define the time instant $T_0(\varepsilon)$ as

$$T_0(\varepsilon) \triangleq \inf\left\{t \in \mathbb{N} : \sqrt{\frac{8\log T}{\varepsilon T}} \leq \left(\frac{\delta_{12}(\varepsilon)}{2K\mathcal{L}_h\mathcal{L}_\theta}\right)^{\frac{1}{pq}} \quad \forall s \geq t\right\} . \tag{22}$$

Theorem 2 *Under Assumptions 1, 2, and 3, for parametrized 1-sub-gaussian distributions, Hölder exponents q and p, any $\frac{4}{\rho T} \leq \varepsilon \leq \frac{1}{K}$, horizon $T > T_0(\varepsilon)$ (see (22)), Param-UCB-M's regret scales as*

$$\mathcal{R}(T) = \mathcal{O}\left(\varepsilon^q, \left(\sqrt{\frac{\log T}{\varepsilon T}}\right)^{pq}\right) . \tag{23}$$

Theorem 2 shows that the regret scaling depends on the discretization level ε and the horizon T. Since the second term in the regret upper bound grows as ε decreases, the discretization level ε cannot be made arbitrarily small. Consequently, the regret cannot be made arbitrarily small. The precise form of the upper bound and the accompanying proof are provided in Appendix D.

5 Conclusion

In this paper, we argued that for modern wireless communications, data-driven methods can be preferable over the model-riven approaches, and average-based metric are inadequate when rare events have catastrophic consequences. To address this, we adopted the PM-centric framework which subsumes mean as a special case, and accounts for distributional tails. This framework provided a principled formulation for resource allocation and even reformulates the existing problems, such as outage. We propose a data-driven algorithm for parametric families of distributions that optimizes PMs, and under mild regularity assumptions on the utility and the model, establish regret guarantees. Empirically, we demonstrated for large enough horizons our algorithm outperforms vanilla UCB. Our analysis assumes stationary. Extending the framework to non-stationary bandits, and using contextual information remain important directions for future work.

References

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

[2] Orly Avner and Shie Mannor. Multi-user lax communications: a multi-armed bandit approach. In *Proc. IEEE International Conference on Computer Communications*, pages 1–9, California, July 2016.

[3] Irmak Aykin, Berk Akgun, Mingjie Feng, and Marwan Krunz. MAMBA: A multi-armed bandit framework for beam tracking in Millimeter-wave systems. In *Proc. IEEE International Conference on Computer Communications*, virtual, July 2020.

[4] Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric Maillard. Optimal thompson sampling strategies for support-aware CVaR bandits. In *Proc. International Conference on Machine Learning*, virtual, July 2021.

[5] Mehdi Bennis, Mérouane Debbah, and H. Vincent Poor. Ultrareliable and low-latency wireless communication: Tail, risk, and scale. *Proceedings of the IEEE*, 106(10):1834–1853, 2018.

[6] Matthew B. Booth, Vinayak Suresh, Nicolò Michelusi, and David J. Love. Multi-armed bandit beam alignment and tracking for mobile millimeter wave communications. *IEEE Communications Letters*, 23(7):1244–1248, 2019.

[7] Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Proc. Conference on Learning Theory*, Stockholm, Sweden, July 2018.

[8] Joel Q. L. Chang and Vincent Y. F. Tan. A unifying theory of thompson sampling for continuous risk-averse bandits. In *Proc. AAAI Conference on Artificial Intelligence*, virtual, February 2022.

[9] Yi Gai and Bhaskar Krishnamachari. Decentralized online learning algorithms for opportunistic spectrum access. In *Proc. IEEE Global Telecommunications Conference*, Texas, December 2011.

[10] Morteza Hashemi, Ashutosh Sabharwal, C. Emre Koksal, and Ness B. Shroff. Efficient beam alignment in mmWave systems using contextual bandits. In *Proc. IEEE International Conference on Computer Communications*, Hawaii, April 2018.

[11] Shuyan Hu, Xiaojing Chen, Wei Ni, Ekram Hossain, and Xin Wang. Distributed machine learning for wireless communication networks: Techniques, architectures, and applications. *IEEE Communications Surveys & Tutorials*, 23(3):1458–1493, 2021.

[12] Sunjung Kang and Changhee Joo. Low-complexity learning for dynamic spectrum access in multi-user multi-channel networks. *IEEE Transactions on Mobile Computing*, 20(11):3267–3281, 2021.

[13] Ravi Kumar Kolla, Prashanth L. A., Aditya Gopalan, Krishna Jagannathan, Michael Fu, and Steve Marcus. Bandit algorithms to emulate human decision making using probabilistic distortions. *arXiv:1611.10283*, 2023.

[14] Prashanth L.A. and Sanjay P. Bhat. A Wasserstein distance approach for concentration of empirical risk estimates. *Journal of Machine Learning Research*, 23(238):1–61, 2022.

[15] Nikolaos Nomikos, Mohammad Sadegh Talebi, Themistoklis Charalambous, and Risto Wichman. Bandit-based power control in full-duplex cooperative relay networks with strict-sense stationary and non-stationary wireless communication channels. *IEEE Open Journal of the Communications Society*, 3:366–378, 2022.

[16] Diego Gabriel Soares Pivoto, Felipe A. P. de Figueiredo, Cicek Cavdar, Gustavo Rodrigues de Lima Tejerina, and Luciano Leonel Mendes. A comprehensive survey of machine learning applied to resource allocation in wireless communications. *IEEE Communications Surveys & Tutorials*, 2025.

[17] Saishankar Katri Pulliyakode and Sheetal Kalyani. Reinforcement learning techniques for outer loop link adaptation in 4g/5g systems. *arXiv:1708.00994*, 2017.

[18] Vidit Saxena, Joakim Jaldén, Joseph E. Gonzalez, Mats Bengtsson, Hugo Tullberg, and Ion Stoica. Contextual multi-armed bandits for link adaptation in cellular networks. In *Proc. Workshop on Network Meets AI & ML*, Beijing, China, August 2019.

[19] Alex Tamkin, Ramtin Keramati, Christoph Dann, and Emma Brunskill. Distributionally-aware exploration for CVaR bandits. In *Proc. Advances in Neural Information Processing Systems*, Canada, December 2019.

[20] Chenmien Tan and Paul Weng. CVaR-regret bounds for multi-armed bandits. In *Proc. Asian Conference on Machine Learning*, India, December 2022.

[21] Meltem Tatlı, Arpan Mukherjee, Prashanth L.A., Karthikeyan Shanmugam, and Ali Tajer. Risk-sensitive bandits: Arm mixture optimality and regret-efficient algorithms. In *Proc. International Conference on Artificial Intelligence and Statistics*, Thailand, May 2025.

[22] Meltem Tatlı, Arpan Mukherjee, Prashanth L.A., Karthikeyan Shanmugam, and Ali Tajer. Preference-centric bandits: Optimality of mixtures and regret-efficient algorithms. *arXiv:2504.20877*, 2025.

[23] Wen Wu, Nan Cheng, Ning Zhang, Peng Yang, Weihua Zhuang, and Xuemin Shen. Fast mmWave beam alignment via correlated bandit learning. *IEEE Transactions on Wireless Communications*, 18(12):5894–5908, 2019.

[24] Marie Josepha Youssef, Venugopal V. Veeravalli, Joumana Farah, Charbel Abdel Nour, and Catherine Douillard. Resource allocation in NOMA-based self-organizing networks using stochastic multi-armed bandits. *IEEE Transactions on Communications*, 69(9):6003–6017, 2021.

[25] Jianjun Zhang, Yongming Huang, Yu Zhou, and Xiaohu You. Beam alignment and tracking for millimeter wave communications via bandit learning. *IEEE Transactions on Communications*, 68(9):5519–5533, 2020.

[26] Guangxu Zhu, Dongzhu Liu, Yuqing Du, Changsheng You, Jun Zhang, and Kaibin Huang. Toward an intelligent edge: Wireless communication meets machine learning. *IEEE Communications Magazine*, 58(1):19–25, 2020.

Appendix

Appendix A establishes that optimizing the PM with the distortion function in (6) is equivalent to optimizing γ-outage capacity, as stated in Lemma 1. Appendix B provides the proof of Lemma 2 on mixtures. In Appendix C, we present the proof for Theorem 1, which shows the optimality for different outage parameters. The regret analysis for the Param-UCB-M algorithm and the proof of Theorem 2 are given in Appendix D. Appendix E presents empirical evaluations for the outage problem, and Appendix F details the computational setup used in the empirical evaluations.

A Proof of Lemma 1

In this section, we show that optimizing $\gamma-$outage capacity is equivalent to optimizing PM with the distortion function defined (6). Considering capacity is non-negative $C \geq 0$, for PM, we have

$$U_h(\mathbb{Q}) = \int_0^\infty h(1-\mathbb{Q}(x))dx = \int_0^\infty \min\Big\{\frac{1-\mathbb{Q}(x)}{1-\gamma}, 1\Big\}dx \ . \tag{24}$$

Additionally, from the definition of $\gamma-$outage capacity, we have

$$1-\mathbb{Q}(x) \leq 1-\gamma \Rightarrow \mathbb{Q}(C_{\text{out}}(\gamma)) = \gamma \Rightarrow C_{\text{out}}(\gamma) := \mathbb{Q}^{-1}(\gamma) \ . \tag{25}$$

Based on this equivalence, we can decompose the utility into two parts as

$$\begin{aligned} U_h(\mathbb{Q}) &\overset{(25)}{=} \int_0^{C_{\text{out}}(\gamma)} 1dx + \int_{C_{\text{out}}(\gamma)}^\infty \frac{1-\mathbb{Q}(x)}{1-\gamma}dx && (26)\\ &= C_{\text{out}}(\gamma) + \frac{1}{1-\gamma}\int_{C_{\text{out}}(\gamma)}^\infty 1-\mathbb{Q}(x)dx && (27)\\ &\overset{(25)}{=} C_{\text{out}}(\gamma) + \frac{1}{1-\mathbb{Q}(C_{\text{out}}(\gamma))}\int_{C_{\text{out}}(\gamma)}^\infty 1-\mathbb{Q}(x)dx && (28)\end{aligned}$$

Now, we denote $C_{\text{out}}(\gamma)$ as $x^\star$, $N(x) = 1-\mathbb{Q}(x)$ and $A(x^\star) = \int_{x^\star}^\infty 1-\mathbb{Q}(x)dx$ and use them in (28) as

$$U_h(\mathbb{Q}) = x^\star + \frac{A(x^\star)}{N(x^\star)} \ . \tag{29}$$

Now, let us take the derivative with respect to $x^\star$

$$\begin{aligned} \frac{\partial U_h(\mathbb{Q})}{\partial x^\star} &= 1 + \frac{A'(x^\star)N(x^\star) - A(x^\star)N'(x^\star)}{N^2(x^\star)} && (30)\\ &= 1 + \frac{A'(x^\star)N(x^\star) + q(x^\star)A(x^\star)}{N^2(x^\star)} && (31)\end{aligned}$$

where $q(x)$ is the probability density function of C. From the definition, we have

$$A'(x^\star) = -N(x^\star) \tag{32}$$

Now, we use this equality to provide a lower-bound on the derivative

$$\begin{aligned} \frac{\partial U_h(\mathbb{Q})}{\partial x^\star} &= 1 + \frac{-N(x^\star)N(x^\star) + q(x^\star)A(x^\star)}{N^2(x^\star)} && (33)\\ &= 1 + \frac{-N(x^\star)N(x^\star)}{N^2(x^\star)} + \frac{q(x^\star)A(x^\star)}{N^2(x^\star)} && (34)\\ &= 1 - \frac{N^2(x^\star)}{N^2(x^\star)} + \frac{q(x^\star)A(x^\star)}{N^2(x^\star)} && (35)\\ &\geq \frac{q(x^\star)A(x^\star)}{N^2(x^\star)} > 0 && (36)\end{aligned}$$

Therefore, we can conclude that $U_h(\mathbb{Q})$ is monotonic in $C_{\text{out}}(\gamma)$, which means optimizing $\gamma-$outage capacity is equivalent to optimizing this PM.

B Proof of Lemma 2

Since h is concave but not affine, there exist $u, v \in (0,1)$, $u \neq v$, and $\lambda \in (0,1)$ such that

$$h\big(\lambda u + (1-\lambda)v\big) \ > \ \lambda h(u) + (1-\lambda)h(v) \ . \tag{37}$$

Let us start with constructing two CDFs as follows. For $s \in (0,1)$ and $L > 0$, define $Q^{(s,L)}$ as the two-point distribution

$$\mathbb{P}(X = L) = s \ , \qquad \text{and} \qquad \mathbb{P}(X = 0) = 1 - s \ . \tag{38}$$

Therefore,

$$U_h(Q^{(s,L)}) = \int_0^L h(s)\, dx = L\, h(s) \ . \tag{39}$$

Next, equalize the U_h-values by setting $L_u = 1/h(u)$ and $L_v = 1/h(v)$, which is possible since $h(u), h(v) > 0$. Define

$$Q_1 \triangleq Q^{(u,L_u)} \ , \qquad \text{and} \qquad Q_2 \triangleq Q^{(v,L_v)} \ . \tag{40}$$

By (39), $U_h(Q_1) = U_h(Q_2) = 1$. Now, let $\mathcal{Q} = \{Q_1, Q_2\}$. Clearly $\mathcal{Q}$ is non-convex, and

$$\max_{Q \in \mathcal{Q}} U_h(Q) = 1 \ . \tag{41}$$

Finally, we create a mixture distribution as follows. Consider $Q_{\text{mix}} = \lambda Q_1 + (1-\lambda)Q_2$. On $[0, \min\{L_u, L_v\})$ the survival values are constant at u and v respectively, so

$$1 - Q_{\text{mix}}(x) = \lambda u + (1-\lambda)v \ . \tag{42}$$

Thus,

$$U_h(Q_{\text{mix}}) = \int_0^{\min\{L_u,L_v\}} h\big(\lambda u + (1-\lambda)v\big)\, dx \ . \tag{43}$$

By (37) and strict concavity of h, this integral strictly exceeds

$$\lambda U_h(Q_1) + (1-\lambda)U_h(Q_2) = 1 \ . \tag{44}$$

Hence $U_h(Q_{\text{mix}}) > 1 = \max_{Q \in \mathcal{Q}} U_h(Q)$.

C Proof of Theorem 1

In this section, we prove the Theorem 1 on the rate selection for outage optimization for finite channels. We consider three cases:

1. $\min_i \bar{p}_{\text{out}}(r_i) \geq 1 - \gamma$,
2. $\max_i \bar{p}_{\text{out}}(r_i) < 1 - \gamma$,
3. there exists $k \in [K-1]$ such that $\bar{p}_{\text{out}}(r_k) \geq 1 - \gamma > \bar{p}_{\text{out}}(r_{k+1})$.

For any $\alpha \in \Delta^{K-1}$, we have

$$U_h\Big(\sum_{i \in [K]} \alpha(i)\mathbb{F}_i\Big) = \int_0^\infty \min\left\{\frac{1 - \sum_{i\in[K]} \alpha(i)\mathbb{F}_i(x)}{1-\gamma}, 1\right\} \mathrm{d}x \tag{45}$$

$$= \sum_{j \in [K]} (r_j - r_{j-1}) \min\left\{\frac{\sum_{i=j}^K \alpha(i)\bar{p}_{\text{out}}(r_i)}{1-\gamma}, 1\right\} \ . \tag{46}$$

with the convention $r_0 = 0$.

Case 1. If $\min_i \bar{p}_{\text{out}}(r_i) \geq 1-\gamma$, then, for all $i \in [K]$, we have the following upper-bound

$$U_h\bigg(\sum_{i\in[K]} \alpha(i)\mathbb{F}_i\bigg) \leq \sum_{j\in[K]} (r_j - r_{j-1}) \min\left\{\frac{\sum_{i=j}^K \alpha(i)\bar{p}_{\text{out}}(r_i)}{1-\gamma}, 1\right\} \leq \sum_{j\in[K]} (r_j - r_{j-1}) = r_K \ . \tag{47}$$

For arm K, we have

$$U_h(\mathbb{F}_K) = r_K \min\left\{\frac{\bar{p}_{\text{out}}(r_K)}{1-\gamma}, 1\right\} = r_K \tag{48}$$

which means

$$U_h(\mathbb{F}_K) \geq \max_{\boldsymbol{\alpha}\in\Delta^{K-1}} U_h\bigg(\sum_{i\in[K]} \alpha(i)\mathbb{F}_i\bigg) \tag{49}$$

Hence, the optimal solution is a solitary arm with the highest transmission rate

$$m^\star \in \arg\max_{i\in[K]} r_i \tag{50}$$

and in this problem, we have $i^\star = K$.

Case 2. If $\max_i \bar{p}_{\text{out}}(r_i) < 1-\gamma$, then, for any $\boldsymbol{\alpha} \in \Delta^{K-1}$, we have $\sum_{i\in[K]} \alpha(i)\bar{p}_{\text{out}}(r_i) < 1-\gamma$. Therefore,

$$\begin{aligned} U_h\bigg(\sum_{i\in[K]} \alpha(i)\mathbb{F}_i\bigg) &= \sum_{j\in[K]} (r_j - r_{j-1}) \frac{\sum_{i=j}^K \alpha(i)\bar{p}_{\text{out}}(r_i)}{1-\gamma} &(51)\\ &= \frac{1}{1-\gamma} \sum_{i\in[K]} \alpha(i)\, \bar{p}_{\text{out}}(r_i) \bigg(\sum_{j\in[i]} (r_j - r_{j-1})\bigg) \\ &= \frac{1}{1-\gamma} \sum_{i\in[K]} r_i \alpha(i)\bar{p}_{\text{out}}(r_i) &(52)\\ &\leq \frac{1}{1-\gamma} \max_{i\in[K]} r_i\bar{p}_{\text{out}}(r_i) &(53)\\ &= \max_{i\in[K]} U_h(\mathbb{F}_i) \ . &(54)\end{aligned}$$

Therefore, the optimal solution is the arm such that

$$m^\star \in \arg\max_{i\in[K]} r_i\bar{p}_{\text{out}}(r_i) \ . \tag{55}$$

Case 3. Now, let us consider the case when $\bar{p}_{\text{out}}(r_k) \geq 1-\gamma > \bar{p}_{\text{out}}(r_{k+1})$.

For every $j \geq k+1$,

$$\sum_{i=j}^K \alpha(i)\bar{p}_{\text{out}}(r_i) < 1-\gamma \ . \tag{56}$$

Therefore, we have

$$\begin{aligned} U_h\bigg(\sum_{i\in[K]} \alpha(i)\mathbb{F}_i\bigg) &= \int_0^\infty \min\left\{\frac{1-\sum_{i\in[K]}\alpha(i)\mathbb{F}_i(x)}{1-\gamma}, 1\right\} \mathrm{d}x &(57)\\ &= \sum_{j\in[K]} (r_j - r_{j-1}) \min\left\{\frac{\sum_{i=j}^K \alpha(i)\bar{p}_{\text{out}}(r_i)}{1-\gamma}, 1\right\} &(58)\\ &= \sum_{j\in[k]} (r_j - r_{j-1}) \min\left\{\frac{\sum_{i=j}^K \alpha(i)\bar{p}_{\text{out}}(r_i)}{1-\gamma}, 1\right\} &(59)\\ &+ \sum_{j=k+1}^K (r_j - r_{j-1}) \frac{\sum_{i=j}^K \alpha(i)\bar{p}_{\text{out}}(r_i)}{1-\gamma} &(60)\end{aligned}$$

The first term is maximized when $\sum_{i=j}^{K} \alpha(i)\bar{p}_{\text{out}}(r_i) \geq 1-\gamma$ for all $j \leq k$, resulting in value r_k. Since $\bar{p}_{\text{out}}(r_i) \geq 1-\gamma$ for $i \leq k$, we can set $\alpha(i)$ for $i \in [k-1]$ for simplification. Thus,

$$U_h\bigg(\sum_{i\in[K]} \alpha(i)\mathbb{F}_i\bigg) = r_k + \sum_{j=k+1}^{K} (r_j - r_{j-1}) \frac{\sum_{i=j}^{K} \alpha(i)\bar{p}_{\text{out}}(r_i)}{1-\gamma} \tag{61}$$

$$= r_k + \frac{1}{1-\gamma} \sum_{i=k+1}^{K} \alpha(i) r_i \bar{p}_{\text{out}}(r_i) \tag{62}$$

The second term is linear in $\alpha(i)$ for $i \geq k+1$ and as in Case 2 maximized by

$$m^\star \in \underset{i\geq k+1}{\arg\max}\, r_i \bar{p}_{\text{out}}(r_i) \tag{63}$$

Considering $\sum_{i=j}^{K} \alpha(i)\bar{p}_{\text{out}}(r_i) \geq 1-\gamma$, our simplification for $\alpha(i)$ for $i \in [k-1]$ and (63), we have

$$\alpha(k) \;=\; \frac{1-\gamma-\bar{p}_{\text{out}}(r_{m^\star})}{\bar{p}_{\text{out}}(r_k)-\bar{p}_{\text{out}}(r_{m^\star})}, \qquad \alpha(m^\star) \;=\; 1-\alpha(k)\,. \tag{64}$$

Therefore, an optimal solution is

$$\alpha^\star(i) = \begin{cases} \alpha(k) & \text{if } i = k \\ \alpha(m^\star) & \text{if } i = m^\star \\ 0 & \text{otherwise} \end{cases} \tag{65}$$

and the maximum value is

$$U_h\left(\sum_{i\in[K]} \alpha^\star(i)\mathbb{F}_i\right) = r_k + \frac{1}{1-\gamma}\alpha(m^\star) r_{m^\star} \bar{p}_{\text{out}}(r_{m^\star}) \tag{66}$$

$$= r_k \;+\; \frac{1}{1-\gamma}\; \frac{\bar{p}_{\text{out}}(r_k)-(1-\gamma)}{\bar{p}_{\text{out}}(r_k)-\bar{p}_{\text{out}}(r_{m^\star})}\; r_{m^\star}\, \bar{p}_{\text{out}}(r_{m^\star}). \tag{67}$$

This completes the proof.

D Regret Analysis

In this section, we provide an analysis on the regret of Param-UCB-M and prove Theorem 2. Let K arms have CDFs $\{\mathbb{F}_i\}_{i=1}^{K}$ with parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]$. For any mixing coefficient $\boldsymbol{\alpha} \in \Delta^{K-1}$, we write the mixture distribution and define the utility $V(\boldsymbol{\alpha}, \boldsymbol{\theta}) \triangleq U_h(\sum_{i\in[K]} \alpha(i)\mathbb{F}_i)$.

Define

$$W \;\triangleq\; \sup_{\boldsymbol{\alpha}\neq\boldsymbol{\beta}\,\in\,\Delta^{K-1}} \frac{\big\|\sum_{i=1}^{K}\alpha_i\mathbb{F}_i - \sum_{i=1}^{K}\beta_i\mathbb{F}_i\big\|_{\mathrm{W}}}{\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|_1} \;<\; \infty, \tag{68}$$

which has the following upper bound $W \leq \sqrt{2\pi}$ [22]. Consequently,

$$\big|V(\boldsymbol{\alpha}, \boldsymbol{\theta}) - V(\boldsymbol{\beta}, \boldsymbol{\theta})\big| \;\leq\; L_h\, W^q\, \|\boldsymbol{\alpha}-\boldsymbol{\beta}\|_1^q. \tag{69}$$

Regret decomposition. We decompose the regret into two parts

$$\mathcal{R}(T) = V(\boldsymbol{\alpha}^\star, \boldsymbol{\theta}) - \mathbb{E}\Big[V\left(\frac{\boldsymbol{\tau}_T}{T}, \boldsymbol{\theta}\right)\Big] \tag{70}$$

$$= \underbrace{V(\boldsymbol{\alpha}^\star, \boldsymbol{\theta}) - V(\mathbf{a}^{(1)}, \boldsymbol{\theta})}_{\text{estimation error } \delta_{01}(\varepsilon)} + \underbrace{\mathbb{E}\Big[V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) - V\left(\frac{\boldsymbol{\tau}_T}{T}, \boldsymbol{\theta}\right)\Big]}_{\text{arm-selection regret } \bar{\mathcal{R}}(T)}. \tag{71}$$

The estimation regret $\delta_{01}(T)$ arises because the optimal solution may involve continuous mixing coefficients, whereas in practice we approximate them using discretization granularity ε. In the following lemma, we show how it scales with discretization granularity ε.

Lemma 3 *For the PM with the Hölder parameters q and L_h, for estimating regret, we have*

$$\delta_{01}(\varepsilon) \leq \mathcal{L} K^q \varepsilon^q W^q \tag{72}$$

Proof: We first define $\bar{\mathbf{a}}$ as the discrete mixing coefficient that is closest to the optimal coefficient $\boldsymbol{\alpha}^\star$, i.e.,

$$\bar{\mathbf{a}} \ \in \ \underset{\mathbf{a}\in\Delta_\varepsilon^{K-1}}{\arg\min} \left\|\boldsymbol{\alpha}^\star - \mathbf{a}\right\|_1 \ . \tag{73}$$

Accordingly, we have

$$\begin{aligned}
\delta_{01}(\varepsilon) &= V(\boldsymbol{\alpha}^\star, \boldsymbol{\theta}) - V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) && (74)\\
&\leq V(\boldsymbol{\alpha}^\star, \boldsymbol{\theta}) - V(\bar{\mathbf{a}}, \boldsymbol{\theta}) && (75)\\
&\leq \mathcal{L} \left\| \sum_{i\in[K]} \Big(\alpha^\star(i) - \bar{a}(i)\Big)\mathbb{F}_i \right\|_{\mathrm{W}}^q && (76)\\
&\leq \mathcal{L} \left\|\boldsymbol{\alpha}^\star - \bar{\mathbf{a}}\right\|_1^q W^q && (77)\\
&\leq \mathcal{L} K^q \varepsilon^q W^q , && (78)
\end{aligned}$$

where, (76) follows from Definition 1, (77) follows from the definition of W in (68), and, (78) follows from the fact that $\bar{\boldsymbol{\alpha}}$ may lie at most $\frac{\varepsilon}{2}$ away from the optimal coefficient $\boldsymbol{\alpha}^\star$ along each coordinate. ■

From Lemma 3, we have $\delta_{01}(\varepsilon) = \mathcal{O}(\varepsilon^q)$, therefore, now, we analyze the arm-selection regret. We condition on the event that all parameters being contained in their respective confidence sets at horizon T, i.e., $i \in [K]$ we have $\theta(i) \in C_T(i)$.

$$\begin{aligned}
\bar{\mathfrak{R}}(T) &= V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) - \mathbb{E}\left[V\left(\frac{\boldsymbol{\tau}_T}{T}, \boldsymbol{\theta}\right)\right] && (79)\\
&= \mathbb{E}\left[V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) - V(\mathbf{a}_T, \boldsymbol{\theta}) + V(\mathbf{a}_T, \boldsymbol{\theta}) - V\left(\frac{\boldsymbol{\tau}_T}{T}, \boldsymbol{\theta}\right)\right] && (80)\\
&\leq \underbrace{\sum_{\mathcal{D}\subseteq[K]:\mathcal{D}\neq\emptyset} \mathbb{E}\left[\left(V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) - V\left(\mathbf{a}_T, \boldsymbol{\theta}\right)\right)\mathbb{1}\{\theta(i) \notin \mathcal{C}_T(i) : i \in \mathcal{D}\}\right]}_{\triangleq B_1(T)} && (81)\\
&\quad + \underbrace{\mathbb{E}\left[\left(V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) - V\left(\mathbf{a}_T, \boldsymbol{\theta}\right)\right)\mathbb{1}\{\theta(i) \in \mathcal{C}_T \ \forall i \in [K]\}\right]}_{\triangleq B_2(T)} && (82)\\
&\quad + \underbrace{\mathbb{E}\left[\left(V(\mathbf{a}_T, \boldsymbol{\theta}) - V\left(\frac{\boldsymbol{\tau}_T}{T}, \boldsymbol{\theta}\right)\right]}_{\triangleq B_3(T)} \ . && (83)
\end{aligned}$$

Expanding $B_1(T)$, we have

$$\begin{aligned}
B_1(T) &= \sum_{\mathcal{D}\subseteq[K]:\mathcal{S}\neq\emptyset} \mathbb{P}\Big(\theta(i) \notin \mathcal{C}_T(i) : i \in \mathcal{S}\Big)\\
&\qquad \times \ \ \mathbb{E}\left[V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) - V\left(\mathbf{a}_T, \boldsymbol{\theta}\right) \,\middle|\, \theta(i) \notin \mathcal{C}_T(i) : i \in \mathcal{S}\right] && (84)\\
&\leq B_h \sum_{\mathcal{D}\subseteq[K]:\mathcal{S}\neq\emptyset} \mathbb{P}\Big(\theta(i) \notin \mathcal{C}_T(i) : i \in \mathcal{S}\Big) && (85)
\end{aligned}$$

where the last line follows from Assumption (3). Leveraging the fact that

$$\sum_{i=1}^{K} \binom{K}{i} x^i = (x+1)^K - 1 \ , \tag{86}$$

along with the sub-gaussian concentration bound for its mean $\mathbb{P}(\theta(i) \notin \mathcal{C}_T(i)) \leq \frac{1}{T^2}$ for every $i \in [K]$, we have

$$\sum_{\mathcal{S}\subseteq[K]:\mathcal{S}\neq\emptyset} \mathbb{P}\Big(\theta(i) \notin \mathcal{C}_T(i) : i \in \mathcal{S}\Big) = \left(\frac{1}{T^2}+1\right)^K - 1\,. \tag{87}$$

Therefore, as an upper-bound on $B_1(T)$ we have

$$B_1(T) \leq B_h\Big(\left(\frac{1}{T^2}+1\right)^K - 1\Big) \tag{88}$$

We now move on to bounding $B_2(T)$. We first define the value of the UCB value for a_t as

$$\mathrm{UCB_t}(\mathbf{a}) = \max_{\mathbf{a}\in\Delta_\varepsilon^{\mathrm{K}-1}} \max_{\kappa\in\mathrm{C_t(i)},\forall\mathrm{i}\in[\mathrm{K}]} \mathrm{V}(\alpha,\kappa)\,. \tag{89}$$

and we define the optimistic parameter estimates which maximize the upper confidence bound for every arm $i \in [K]$ as

$$\widetilde{\theta}_t(i) \ \triangleq \ \arg\max_{\kappa(i)\in\mathcal{C}_t(i)} \max_{\kappa(j)\in\mathcal{C}_t(j):j\neq i} \mathrm{UCB}_t(\mathbf{a}_t)\,. \tag{90}$$

and for their CDF, we define

$$\widetilde{\mathbb{F}}_{i,t} \triangleq \mathbb{F}(;\widetilde{\theta}_t(i))\,. \tag{91}$$

Expanding $B_2(T)$, we have

$$\begin{aligned}
B_2(T) &= \mathbb{E}\left[V(\mathbf{a}^{(1)},\boldsymbol{\theta}) - V\left(\mathbf{a}_T,\boldsymbol{\theta}\right)\Big|\,\theta(i)\in\mathcal{C}_T(i)\ \forall i\in[K]\right] && (92)\\
&\leq \mathbb{E}\left[\mathrm{UCB}_T(\mathbf{a}^{(1)}) - V\left(\mathbf{a}_T,\boldsymbol{\theta}\right)\Big|\,\theta(i)\in\mathcal{C}_T(i)\ \forall i\in[K]\right] && (93)\\
&\leq \mathbb{E}\left[\mathrm{UCB}_T(\mathbf{a}_T) - V\left(\mathbf{a}_T,\boldsymbol{\theta}\right)\middle|\,\theta(i)\in\mathcal{C}_T(i)\ \forall i\in[K]\right] && (94)\\
&= \mathbb{E}\left[\mathrm{UCB}_T(\mathbf{a}_T) - V(\mathbf{a}_T,\boldsymbol{\theta})\middle|\,\theta(i)\in\mathcal{C}_T(i)\ \forall i\in[K]\right] && (95)\\
&= \mathbb{E}\left[V(\mathbf{a}_T,\widetilde{\theta}_t(i)) - V(\mathbf{a}_T,\boldsymbol{\theta})\Big|\,\theta(i)\in\mathcal{C}_T(i)\ \forall i\in[K]\right] && (96)\\
&\leq \mathbb{E}\left[\mathcal{L}\sum_{i\in[K]}\left(a_T(i)\right)^q\|\mathbb{F}(\cdot;\widetilde{\theta}_T(i)) - \mathbb{F}_i\|_{\mathrm{W}}^q\middle|\,\theta(i)\in\mathcal{C}_T(i)\ \forall i\in[K]\right] && (97)\\
&\leq \mathcal{L}_h\mathcal{L}_\theta\mathbb{E}\left[\sum_{i\in[K]}|\widetilde{\theta}_T(i)-\theta(i)|^{pq}\right] && (98)\\
&\leq \mathcal{L}_h\mathcal{L}_\theta\mathbb{E}\left[\sum_{i\in[K]}\sqrt{2\frac{\log T}{\tau_T(i)}}\right] && (99)\\
&\leq \mathcal{L}_h\mathcal{L}_\theta K\left(\sqrt{8\frac{\log T}{\rho\varepsilon T}}\right)^{pq} && (100)
\end{aligned}$$

where (98) follows from Assumption 2, (99) follows from the definition of confidence radius, and (100) follows from forced exploration.

Now, we focus on the $B_3(T)$ which measures the error between the chosen mixing coefficient and the ratios of arm samples.

$$\begin{aligned}
B_3(T) &\leq \mathcal{L}_h\mathbb{E}\left[\|\sum_{i\in[K]}\left(a_T(i) - \frac{\tau_T(i)}{T}\right)\mathrm{F}_i\|_{\mathrm{W}}^q\right] && (101)\\
&\leq\ \mathcal{L}_h W^q\mathbb{E}\left[|\mathbf{a}_T - \frac{1}{T}\boldsymbol{\tau}_T|^q\right]\,. && (102)
\end{aligned}$$

First, let us define the event

$$\mathcal{E}_{0,t} \triangleq \Big\{\theta(i) \in \mathcal{C}_t(i) \quad \forall i \in [K]\Big\} . \tag{103}$$

Let us denote the *set* of discrete optimal mixtures of the PM U_h by

$$\mathrm{OPT}_\varepsilon \ \triangleq \ \arg \max_{a \in \Delta_\varepsilon^{K-1}} \ V\left(\mathbf{a}, \boldsymbol{\theta}\right) \ , \tag{104}$$

and the *set* of optimistic mixtures at each instant t by

$$\widetilde{\mathrm{OPT}}_{\varepsilon,t} \ \triangleq \ \arg \max_{\mathbf{a} \in \Delta_\varepsilon^{K-1}} \ \max_{\kappa(i) \in \mathcal{C}_t(i) \ \forall i \in [K]} \ V\left(\mathbf{a}, \kappa\right) \ . \tag{105}$$

For any $\mathbf{a}^{(1)} \in \mathrm{OPT}_\varepsilon$ and $\widetilde{\boldsymbol{\alpha}}_t \in \widetilde{\mathrm{OPT}}_{\varepsilon,t}$, let us define the events

$$\mathcal{E}_{1,t}(x) \triangleq \Big\{\Big|V(\mathbf{a}^{(1)}, \widetilde{\boldsymbol{\theta}}_t) - V(\mathbf{a}^{(1)}, \boldsymbol{\theta})\Big| < x\Big\} \ , \tag{106}$$

$$\mathcal{E}_{2,t}(x) \triangleq \Big\{\Big|V(\mathbf{a}_t, \widetilde{\boldsymbol{\theta}}_t) - V(\mathbf{a}_t, \boldsymbol{\theta})\Big| < x\Big\} \ , \tag{107}$$

$$\text{and,} \quad \mathcal{E}_t(x) \triangleq \mathcal{E}_{1,t}(x) \bigcap \mathcal{E}_{2,t}(x) \ . \tag{108}$$

Note that

$$\mathbb{P}\Big(\widetilde{\mathrm{OPT}}_{\varepsilon,t} \neq \mathrm{OPT}_\varepsilon\Big) \ = \ \mathbb{P}\Big(\exists \mathbf{a} \in \widetilde{\mathrm{OPT}}_{\varepsilon,t} : \mathbf{a} \notin \mathrm{OPT}_\varepsilon\Big) \ . \tag{109}$$

Then, the error probability at time t is

$$\begin{aligned}
\mathbb{P}\Big(\mathbf{a}_t \notin \mathrm{OPT}_\varepsilon\Big) &= \mathbb{P}\Big(\mathbf{a}_t \notin \mathrm{OPT}_\varepsilon \mid \bar{\mathcal{E}}_{0,t}\Big)\mathbb{P}(\bar{\mathcal{E}}_{0,t}) + \mathbb{P}\Big(\mathbf{a}_t \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big)\mathbb{P}(\mathcal{E}_{0,t}) && (110)\\
&\leq \mathbb{P}(\bar{\mathcal{E}}_{0,t}) + \mathbb{P}\Big(\mathbf{a}_t \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big) && (111)\\
&\leq \left(\frac{1}{T^2} + 1\right)^K - 1 + \mathbb{P}\Big(a_t \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big) \ . && (112)
\end{aligned}$$

Next, note that

$$\mathbb{P}\Big(\mathbf{a}_t \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big)$$

$$= \mathbb{P}\left(V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) > V(\mathbf{a}_t, \boldsymbol{\theta}) + \delta_{12}(\varepsilon) \middle| \mathcal{E}_{0,t}\right) \tag{113}$$

$$= \mathbb{P}\left(V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) > V(\mathbf{a}_t, \boldsymbol{\theta}) + \delta_{12}(\varepsilon) \middle| \mathcal{E}_{0,t},\ \mathcal{E}_t\left(\frac{1}{2}\delta_{12}(\varepsilon)\right)\right)$$

$$\times \mathbb{P}\left(\mathcal{E}_t\left(\frac{1}{2}\delta_{12}(\varepsilon)\right) \middle| \mathcal{E}_{0,t}\right) \tag{114}$$

$$+ \mathbb{P}\left(V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) > V(\mathbf{a}_t, \boldsymbol{\theta}) + \delta_{12}(\varepsilon) \middle| \mathcal{E}_{0,t},\ \bar{\mathcal{E}}_t\left(\frac{1}{2}\delta_{12}(\varepsilon)\right)\right)$$

$$\times \mathbb{P}\left(\bar{\mathcal{E}}_t\left(\frac{1}{2}\delta_{12}(\varepsilon)\right) \middle| \mathcal{E}_{0,t}\right) \tag{115}$$

$$\leq \mathbb{P}\left(V(\mathbf{a}^{(1)}, \boldsymbol{\theta}) > V(\mathbf{a}_t, \boldsymbol{\theta}) + \delta_{12}(\varepsilon) \middle| \mathcal{E}_{0,t},\ \mathcal{E}_t\left(\frac{1}{2}\delta_{12}(\varepsilon)\right)\right)$$

$$+ \mathbb{P}\left(\bar{\mathcal{E}}_t\left(\frac{1}{2}\delta_{12}(\varepsilon)\right) \middle| \mathcal{E}_{0,t}\right) \tag{116}$$

$$\leq \mathbb{P}\left(V(\mathbf{a}^{(1)}, \widetilde{\boldsymbol{\theta}}_t) > V(\mathbf{a}_t, \widetilde{\boldsymbol{\theta}}_t) \middle| \mathcal{E}_{0,t},\ \bar{\mathcal{E}}_t\left(\frac{1}{2}\delta_{12}(\varepsilon)\right)\right)$$

$$+ \mathbb{P}\left(\bar{\mathcal{E}}_t\left(\frac{1}{2}\delta_{12}(\varepsilon)\right) \middle| \mathcal{E}_{0,t}\right) \tag{117}$$

$$= \mathbb{P}\left(\bar{\mathcal{E}}_t\left(\frac{1}{2}\delta_{12}(\varepsilon)\right) \middle| \mathcal{E}_{0,t}\right) \tag{118}$$

$$\leq \mathbb{P}\left(\bar{\mathcal{E}}_{1,t}\left(\frac{1}{2}\delta_{12}(\varepsilon)\right) \middle| \mathcal{E}_{0,t}\right) + \mathbb{P}\left(\bar{\mathcal{E}}_{2,t}\left(\frac{1}{2}\delta_{12}(\varepsilon)\right) \middle| \mathcal{E}_{0,t}\right)\ , \tag{119}$$

Next, we will find upper bounds on the two probability terms in (119). Note that for any $t > N$ (16), i.e., after the forced exploration phase, we have

$$\mathbb{P}\left(\bar{\mathcal{E}}_{1,t}\left(\frac{1}{2}\delta_{12}(\varepsilon)\right) \middle| \mathcal{E}_{0,t}\right)$$

$$= \mathbb{P}\left(\left|V(\mathbf{a}^{(1)}, \tilde{\boldsymbol{\theta}}_t(i)) - V\left(\mathbf{a}^{(1)}, \boldsymbol{\theta}\right)\right| \geq \frac{1}{2}\delta_{12}(\varepsilon) \middle| \mathcal{E}_{0,t}\right) \tag{120}$$

$$\leq \mathbb{P}\left(\sum_{i\in[K]} (a^{(1)}(i))^q \left\|\mathbb{F}(\cdot; \widetilde{\theta}_t(i)) - \mathbb{F}(\cdot; \theta(i))\right\|_W^q \geq \frac{1}{2\mathcal{L}_h}\delta_{12}(\varepsilon) \middle| \mathcal{E}_{0,t}\right) \tag{121}$$

$$\leq \mathbb{P}\left(\sum_{i\in[K]} (a^{(1)}(i))^q \left\|\widetilde{\theta}_t(i) - \theta(i)\right\|_W^{pq} \geq \frac{1}{2\mathcal{L}_h\mathcal{L}_\theta}\delta_{12}(\varepsilon) \middle| \mathcal{E}_{0,t}\right) \tag{122}$$

$$\leq \mathbb{P}\left(\sum_{i\in[K]} \underbrace{\left(a^{(1)}(i)\right)^q}_{\leq 1} \left(\sqrt{\frac{2\log T}{\tau_t(i)}}\right)^{pq} > \frac{1}{2\mathcal{L}_h\mathcal{L}_\theta}\delta_{12}(\varepsilon)\right) \tag{123}$$

$$\leq \mathbb{P}\left(\left(\sqrt{\frac{2\log T}{\frac{1}{4}\rho\varepsilon T}}\right)^{pq} > \frac{1}{2K\mathcal{L}_h\mathcal{L}_\theta}\delta_{12}(\varepsilon) \middle| \mathcal{E}_{0,t}\right) \tag{124}$$

$$\leq \mathbb{P}\left(\left(\sqrt{\frac{8\log T}{\varepsilon T}}\right)^{pq} > \frac{1}{2K\mathcal{L}_h\mathcal{L}_\theta}\delta_{12}(\varepsilon)\right) \tag{125}$$

Let us define

$$T_0(\varepsilon) \;\triangleq\; \inf\left\{t \in \mathbb{N} : \sqrt{\frac{8\log T}{\varepsilon T}} \le \left(\frac{\delta_{12}(\varepsilon)}{2K\mathcal{L}_h\mathcal{L}_\theta}\right)^{\frac{1}{pq}} \quad \forall s \ge t\right\} . \tag{126}$$

Hence, $T \ge T_0(\varepsilon)$, for all $t \ge N$, we have

$$\mathbb{P}\left(\bar{\mathcal{E}}_{1,t}\left(\frac{1}{2}\delta_{12}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right) \;=\; 0\,. \tag{127}$$

following similar arguments we have

$$\mathbb{P}\left(\bar{\mathcal{E}}_{2,t}\left(\frac{1}{2}\delta_{12}(\varepsilon)\right)\middle|\mathcal{E}_{0,t}\right) \;=\; 0\,. \tag{128}$$

From (127) and (128), we have

$$\mathbb{P}\Big(\widetilde{\mathrm{OPT}}_{\varepsilon,t} \ne \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big) \;=\; 0 \tag{129}$$

and

$$\mathbb{P}\Big(\mathbf{a}_t \notin \mathrm{OPT}_\varepsilon\Big) \le \left(\frac{1}{T^2}+1\right)^K - 1 + \mathbb{P}\Big(a_t \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,t}\Big) \overset{(129)}{=} \left(\frac{1}{T^2}+1\right)^K - 1 \tag{130}$$

We now define the low probability event $\mathcal{E}_{3,T}$. Let us define

$$\mathcal{E}_{3,T} \triangleq \Big\{\exists t \in [N,T] : \widetilde{\mathrm{OPT}}_{\varepsilon,t} \ne \mathrm{OPT}_\varepsilon\Big\}\,. \tag{131}$$

We show that the probability of $\mathcal{E}_{3,T}$ happening is bounded from above by

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_{3,T}\right) &= \mathbb{P}\left(\exists t > N : \widetilde{\mathrm{OPT}}_{\varepsilon,t} \ne \mathrm{OPT}_{\varepsilon,\mathrm{t}}\right) &(132)\\
&\le \sum_{t=N}^{T} \mathbb{P}\left(\widetilde{\mathrm{OPT}}_{\varepsilon,t} \ne \mathrm{OPT}_{\varepsilon,\mathrm{t}}\right) &(133)\\
&\le \sum_{t=N}^{T} (\frac{1}{T^2}+1)^K - 1) + \mathbb{P}(\boldsymbol{\alpha}_t \notin \mathrm{OPT}_\varepsilon \mid \mathcal{E}_{0,\mathrm{t}}) &(134)\\
&\le \sum_{t=N}^{T} \left(\frac{1}{T^2}+1\right)^K - 1 &(135)\\
&\le T\left(\frac{1}{T^2}+1)^K - 1\right)\,. &(136)
\end{aligned}$$

Therefore, for $B_3(T)$, we have

$$\begin{aligned}
B_3(T) \le{}& \mathcal{L}_h W^q\,\mathbb{E}\left[\sum_{i\in[K]} \left\|a_T(i) - \tfrac{1}{T}\tau_T(i)\right\|_1^q \,\middle|\, \mathcal{E}_{3,T}\right]\mathbb{P}(\mathcal{E}_{3,T}) &(137)\\
&+ \mathcal{L}_h W^q\,\mathbb{E}\left[\sum_{i\in[K]} \left\|a_T(i) - \tfrac{1}{T}\tau_T(i)\right\|_1^q \,\middle|\, \bar{\mathcal{E}}_{3,T}\right]\mathbb{P}(\bar{\mathcal{E}}_{3,T})\\
\le{}& \mathcal{L}_h W^q K\,\mathbb{P}(\mathcal{E}_{3,T}) + \mathcal{L}_h W^q\,\mathbb{E}\left[\sum_{i\in[K]} \left\|a_T(i) - \tfrac{1}{T}\tau_T(i)\right\|_1^q \,\middle|\, \bar{\mathcal{E}}_{3,T}\right]. &(138)
\end{aligned}$$

Lemma 4 *Under the event $\bar{\mathcal{E}}_{3,T}$, $\mathbf{a}^{(1)} \in \mathrm{OPT}_\varepsilon$ and $\frac{4}{\rho T} \le \varepsilon \le \frac{1}{K}$ we have*

$$\left\|a^{(1)}(i) - \tfrac{1}{T}\tau_t(i)\right\|_1 \le \frac{K}{T} \tag{139}$$

Proof: Proof follows from the under-sampling proof [22]. ■

Therefore, we have

$$B_3(T) \leq \mathcal{L}_h W^q K\, \mathbb{P}(\mathcal{E}_{3,T}) + \mathcal{L}_h W^q\, \mathbb{E}\left[\sum_{i\in[K]} \left\| a_T(i) - \tfrac{1}{T}\tau_T(i) \right\|_1^q \,\middle|\, \bar{\mathcal{E}}_{3,T} \right] \tag{140}$$

$$\leq \mathcal{L}_h W^q K \mathbb{P}(\mathcal{E}_{3,T}) + \mathcal{L}_h W^q \frac{K^{q+1}}{T^q} \tag{141}$$

$$\leq \mathcal{L}_h W^q K \Big(\mathbb{P}(\mathcal{E}_{3,T}) + \frac{K^q}{T^q} \Big) \tag{142}$$

$$\leq \mathcal{L}_h W^q K \left(\left(T(\frac{1}{T^2} + 1)^K - 1 \right) + \frac{K^q}{T^q} \right) \tag{143}$$

Therefore, for the arm selection regret, we have

$$\bar{\mathcal{R}}_T = B_1(T) + B_2(T) + B_3(T) \tag{144}$$

$$\leq B_h \Bigg(\left(\frac{1}{T^2} + 1 \right)^K - 1 \Bigg) + \mathcal{L}_h \mathcal{L}_\theta K \left(\sqrt{\frac{8\log T}{\varepsilon T}} \right)^{pq} \tag{145}$$

$$+ \mathcal{L}_\langle W^q K \left(\left(T(\frac{1}{T^2} + 1)^K - 1 \right) + \frac{K^q}{T^q} \right) . \tag{146}$$

Therefore, we can bound the arm selection regret from above using (88), (100), and (143),

$$\bar{\mathcal{R}}_T = B_1(T) + B_2(T) + B_3(T) \tag{147}$$

$$\leq B_h \Bigg(\left(\frac{1}{T^2} + 1 \right)^K - 1 \Bigg) + \mathcal{L}_h \mathcal{L}_\theta K \left(\sqrt{\frac{8\log T}{\varepsilon T}} \right)^{pq} \tag{148}$$

$$+ \mathcal{L}_\langle W^q K \left(\left(T(\frac{1}{T^2} + 1)^K - 1 \right) + \frac{K^q}{T^q} \right) . \tag{149}$$

Now, from (147), and Lemma 3, we have

$$\mathcal{R}(T) \leq \delta_{01}(\varepsilon) + \bar{\mathcal{R}}(T) \tag{150}$$

$$\leq \delta_{01} + B_h \Bigg(\left(\frac{1}{T^2} + 1 \right)^K - 1 \Bigg) + \mathcal{L}_h \mathcal{L}_\theta K \left(\sqrt{\frac{8\log T}{\varepsilon T}} \right)^{pq} \tag{151}$$

$$+ \mathcal{L}_h W^q \left(T \left(\frac{1}{T^2} + 1)^K - 1 \right) + \frac{K^{q+1}}{T^q} \right) \tag{152}$$

$$\leq \mathcal{L}_h K W^q \varepsilon^q + B_h \Bigg(\left(\frac{1}{T^2} + 1 \right)^K - 1 \Bigg) + \mathcal{L}_h \mathcal{L}_\theta K \left(\sqrt{\frac{8\log T}{\varepsilon T}} \right)^{pq} \tag{153}$$

$$+ \mathcal{L}_h W^q \left(T \left(\frac{1}{T^2} + 1 \right)^K - 1 \right) + \frac{K^{q+1}}{T^q} \Bigg) \tag{154}$$

$$\leq \mathcal{L}_h K W^q \varepsilon^q + 3\max\{B_h, L_h L_\theta, L_h W^q\} K \left(\sqrt{\frac{8\log T}{\varepsilon T}} \right)^{pq} \tag{155}$$

$$\leq 4K \max\{B_h, L_h L_\theta, L_h W^q\} \max\left\{ \varepsilon^q, \left(\sqrt{\frac{8\log T}{\varepsilon T}} \right)^{pq} \right\} . \tag{156}$$

. Therefore, regret scales as

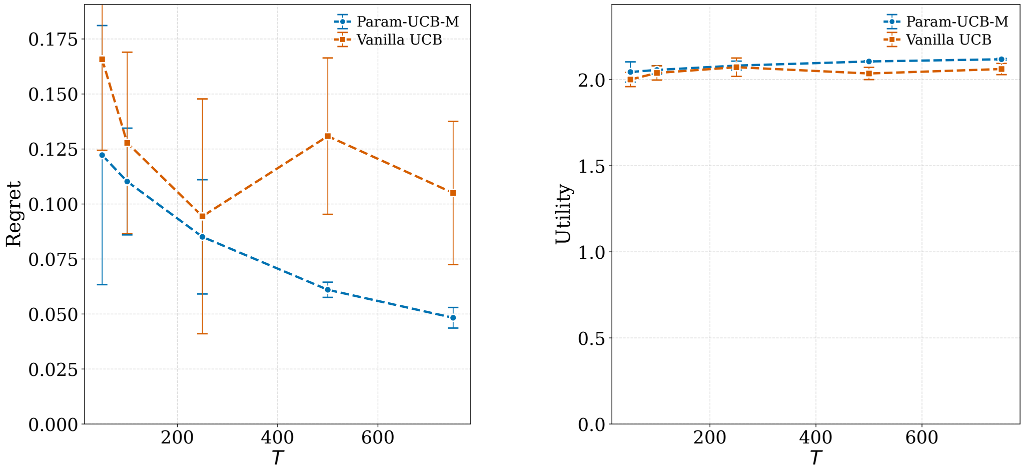
$$\mathcal{R}(T) = \mathcal{O}\left(\varepsilon^q, \left(\sqrt{\frac{\log T}{\varepsilon T}} \right)^{pq} \right) \tag{157}$$

where p is the Hölder parameter of the CDF and q is the Hölder parameter of the utility.

E Empirical Evaluations

In section 3.2, we introduced an outage example where the optimal solution is a mixture of two distributions. We now provide empirical evaluations for this setting. Since the relevant arms, i.e., arms with non-zero weights are transmission rates, are 2 and 3, we include them and add a third arm with rate 4 to show convergence to the mixture solution. We evaluate ther performance of Param-UCB-M and report average regret across horizons under the PM with distortion function defined in (6) for $\gamma = 0.1$. Each result is averaged over 10 independent trials. As a baseline, we compare against vanilla UCB[1] algorithm.

Results. We compare the regret of Param-UCB-M against the vanilla UCB algorithm [1] over horizons $50, 100, 250, 500, 750$. For Param-UCB-M, we fix $\rho = 0.9$ and set $\varepsilon = \mathcal{O}\left(\left(\frac{\log T}{T}\right)^{\frac{1}{3}}\right)$. Figure 1a show a clear separation: vanilla UCB consistently incurs higher regret than Param-UCB-M across all horizons. This occurs because vanilla UCB converges to a single arm that maximizes the mean reward, which is suboptimal in settings where the optimal solution is a mixture. Its regret stays slightly above 2 since we evaluate performance under a PM rather than the mean. Early exploration by vanilla UCB temporarily creates a mixture, but it ultimately concentrates on a solitary arm, leading to persistent sub-optimality. By contrast, Param-UCB-M is explicitly designed to converge to the optimal mixture, resulting in consistently lower regret that decreases sub-linearly in Figure 1a and approaches the true optimal value of 2.16667 in Figure 1b .



(a) Average regret for Param-UCB-M and Vanilla UCB across different horizons.

(b) Average utility value for Param-UCB-M and vanilla UCB across different horizons

Figure 1: Comparison of performance for Param-UCB-M and vanilla UCB algorithms.

F Computational Resources

All experiments were executed on a dedicated Ubuntu 24.04.2 LTS server with dual-socket AMD EPYC 9124 CPUs (2×16 cores; 32 cores total) and 377 GiB RAM. Although the server includes GPUs, all runs in this paper used CPUs only (no GPU acceleration).