MULTINOMIAL SUBSET ROUTING WITH SUM-MAX REWARDS AND OPERATIONAL CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online routing to subsets of experts under aggregate bandit feedback and long-run operational constraints. Expert contributions can be complementary: for each task dimension, the best selected expert determines the contribution, and the total reward aggregates these dimension-wise maxima. At the same time, capacity, budget, and fairness requirements impose lower and upper bounds on long-run expert activation frequencies. A fixed deterministic subset cannot generally satisfy such heterogeneous requirements, motivating a stochastic routing policy. We formalize this problem as Multinomial Subset Routing (MSR). The learner maintains a distribution q over K experts, samples an expert independently M times from q, and routes each task to the distinct sampled experts. We propose OMD-Approachability, which combines online mirror descent with Blackwell's approachability to optimize MSR under two-sided operational constraints. We establish $O(T^{-1/2})$ average reward regret and expected constraint violation. We also quantify the approximation gap induced by a linear surrogate, and evaluate the approach on a real-world crowdsourcing dataset.

1 INTRODUCTION

We study online routing to small subsets of experts under aggregate feedback and long-run operational constraints. The selected experts can be complementary: different experts may be best suited to different aspects of a task, so the value of a team depends on the capabilities covered jointly. At the same time, deployment may impose long-term heterogeneous requirements on workload, cost, or fair participation. The learner therefore must optimize subset quality while controlling long-run expert usage from incomplete feedback. These characteristics arise naturally in expert routing, crowdsourcing, and other resource-constrained allocation problems.

A motivating example. Consider a crowdsourcing platform that assigns each task to a small team of workers. A task may involve several skills, and different workers may be particularly effective on different skills. If $V_{k,i} \geq 0$ denotes the ability of worker $i \in [K]$ on skill $k \in [N]$, the quality of a selected subset $A \subseteq [K]$ is naturally represented by the sum-max reward

$$r(A) = \sum_{k=1}^{N} \max_{i \in A} V_{k,i}.$$

The platform observes only the aggregate quality of the selected team and must simultaneously control long-run workload, cost, or participation. This gives a concrete instance of the online routing problem studied here.

We formalize the problem as **Multinomial Subset Routing (MSR)**. The learner maintains a distribution q over K experts, draws an expert independently M times from q, and routes each task to the distinct sampled experts. Rather than choosing a deterministic subset directly, the learner learns the distribution that induces these subsets and their long-run activation frequencies.

Several lines of work address individual components of this problem. Mixture of Experts and LLM-routing methods learn input-dependent expert selection, often through learned gating or performance–cost trade-offs (Jacobs et al., 1991; Shazeer et al., 2017; Fedus et al., 2022; Chen et al., 2020; Shnitzer et al., 2023; Ong et al., 2024), and recent work studies online LLM routing with bandit feedback

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Nguyen et al., 2024; Wei et al., 2025). Sum-max and submodular bandits model complementary subset rewards (Pasteris et al., 2024; Yue & Guestrin, 2011; Yu et al., 2016; Takemori et al., 2020), while constrained online learning and Bandits with Knapsacks control decisions coupled by long-run resource constraints (Badanidiyuru et al., 2018; Agrawal & Devanur, 2016; Slivkins et al., 2023). Our setting combines these ingredients: the learner chooses a routing distribution, observes only aggregate subset feedback, and must satisfy simultaneous lower and upper activation constraints.

This combination creates two coupled challenges. First, expert i is activated with probability $1-(1-q_i)^M$, which is nonlinear in q. Second, the learner observes only the aggregate reward of the sampled subset, so the gradient of the expected sum-max reward must be estimated from subset-level feedback. The key mathematical insight is that the nonlinear activation constraints admit a linear surrogate: for every $q_i \in [0,1]$, $q_i \leq 1-(1-q_i)^M \leq Mq_i$, and the resulting coordinate-wise inequalities yield a distance-transfer guarantee from the surrogate target to the original activation constraints. For reward learning, the concavity of the expected sum-max reward (Pasteris et al., 2024) gives an importance-weighted gradient estimator based only on the aggregate subset reward.

We combine these ideas in **OMD-Approachability**. The primal update uses online mirror descent to maximize the concave reward while remaining in the supporting halfspace of the surrogate constraint set selected by the current dual direction. The dual update uses an OCO-based approachability procedure (Hazan, 2016) to track long-run constraint pressure. Both updates operate on the same routing distribution, allowing reward optimization and operational control to proceed simultaneously.

Our contributions are fourfold. First, we formulate MSR, combining sum-max rewards, aggregate subset feedback, and simultaneous lower and upper long-run activation constraints. Second, we develop OMD-Approachability and establish $O(T^{-1/2})$ average reward regret and expected constraint violation under the benchmark-feasibility condition. Third, we quantify the approximation introduced by the linear surrogate, including a multiplicative $1-e^{-1}$ guarantee under upper constraints. Finally, we evaluate the method in simulated crowdsourcing environments constructed from real data, under both stationary and changing reward conditions. A contextual extension is also provided.

2 Related Work

Expert and model routing. Mixture of Experts methods route inputs to sparse subsets of expert subnetworks through learned gates (Jacobs et al., 1991; Shazeer et al., 2017; Fedus et al., 2022). At the system level, routing among complete models has been studied through learned performance–cost trade-offs and model-response routers (Chen et al., 2022; 2020; Shnitzer et al., 2023; Ong et al., 2024; Chen et al., 2024; Feng et al., 2025), including recent online formulations based on multi-armed or contextual bandits (Nguyen et al., 2024; Wei et al., 2025). These approaches typically select a model or expert subset directly, whereas MSR learns a distribution whose repeated multinomial sampling induces the subset and simultaneously controls long-run activation frequencies.

Constrained online learning. Bandits with knapsacks study sequential reward maximization under resource constraints (Badanidiyuru et al., 2018), with extensions to contextual, linear, combinatorial, and adversarial settings (Agrawal et al., 2016; Agrawal & Devanur, 2016; Slivkins et al., 2023; Sankararaman & Slivkins, 2018; Liu et al., 2022; Immorlica et al., 2022). The concave-reward, convex-constraint formulation of Agrawal & Devanur (2014) is particularly related to our use of Blackwell approachability. In MSR, however, the relevant constraint is the activation probability $1-(1-q_i)^M$, which is nonlinear in the routing distribution. Our analysis therefore introduces a linear surrogate and then transfers the resulting guarantee back to the original activation constraints.

Sum-max, submodular, and combinatorial bandits. Sum-max rewards are monotone submodular and arise naturally in complementary subset selection. Existing submodular and combinatorial bandit methods study direct selection of deterministic super-arms under various feedback models (Yue & Guestrin, 2011; Yu et al., 2016; Takemori et al., 2020; Chen et al., 2017; Kveton et al., 2015; Chen et al., 2016; Wang & Chen, 2018). K-max bandits consider the special case of a single maximum (Gopalan et al., 2014; Fourati et al., 2024; Chen et al., 2025). The closest reward model to ours is the sum-max bandit formulation of Pasteris et al. (2024), whose concavity and gradient characterization we use. Our setting differs in that the learner does not select a subset directly: it

learns a routing distribution, observes only aggregate subset reward, and must satisfy whole-horizon two-sided operational constraints.

3 PROBLEM SETUP

We consider an online routing problem with K available experts, M routing draws per round, and horizon T. Let $[K] := \{1, \dots, K\}$ and $\Delta_K := \{q \in \mathbb{R}_+^K : \sum_i q_i = 1\}$ denote the probability simplex. Each expert $i \in [K]$ has a *known* cost vector $c_i \in [0, \bar{c}]^d$. At round t, nature specifies an *unknown* reward function $r_t : 2^{[K]} \to [0, 1]$ with $r_t(\emptyset) = 0$. The learner chooses $q_t \in \Delta_K$, draws M experts $a_{t,1}, \dots, a_{t,M}$ independently from q_t, and routes to the distinct sampled set $A_t := \{a_{t,1}, \dots, a_{t,M}\}$. It observes only the aggregate reward $r_t(A_t)$ and incurs the realized cost

$$\ell_t := \sum_{i=1}^{K} c_i \, \mathbf{1}\{i \in A_t\}.$$

The reward sequence is fixed before the interaction and is unknown to the learner.

Sum-max reward. We consider the sum-max reward class introduced by Pasteris et al. (2024). Fix $N \in \mathbb{N}$. At each round t, there is a matrix $V_t \in \mathbb{R}^{N \times K}$ such that

$$r_t(A) = \sum_{k=1}^{N} \max_{i \in A} V_{t,k,i}, \qquad \forall A \subseteq [K], \qquad r_t(\emptyset) = 0.$$

For $q \in \Delta_K$, let $B(q)$ be the *distinct* set obtained from M independent draws from q, and define the expected reward

$$\Phi^{r_t}(q) := \mathbb{E}[r_t(B(q))].$$

Operational constraints. The learner must also satisfy long-run operational constraints. Let $S^{\mathrm{ori}} = \prod_{j=1}^{d} [l_j, u_j]$ denote the axis-aligned target set for the time-averaged realized cost. Under q, action i is activated with probability $1 - (1 - q_i)^M$, so the corresponding expected cost is nonlinear in q. We therefore use the linear surrogate $\sum_i q_i c_i$ and the compact target $S = \prod_{j=1}^{d} S^{(j)}$, where

$$S^{(j)} = [0, u_j/M] \text{ for an upper-bound coordinate}, \qquad S^{(j)} = [l_j, \bar{c}] \text{ for a lower-bound coordinate.}$$

Section 4.1 establishes the distance-transfer guarantee connecting the surrogate and original constraints.

Benchmark and performance criteria. We compare the learner with the best fixed routing distribution in hindsight that is feasible for the surrogate target:

$$\mathrm{OPT}_{\mathrm{sur}} := \max_{q \in \Delta_K} \ \tfrac{1}{T} \sum_{t=1}^{T} \Phi^{r_t}(q) \quad \text{s.t.} \quad \sum_{i=1}^{K} q_i c_i \in S. \tag{1}$$

We refer to $\mathrm{OPT}_{\mathrm{sur}}$ as the *surrogate-feasible benchmark*. For $u \in \mathbb{R}^d$ and a nonempty closed set A, let $\mathrm{dist}(u, A) := \min_{x \in A} \|x - u\|$. We evaluate the average objective regret and realized constraint regret by

$$\mathcal{R}_1(T) := \mathrm{OPT}_{\mathrm{sur}} - \tfrac{1}{T} \sum_{t=1}^{T} \Phi^{r_t}(q_t), \qquad \mathcal{R}_2(T) := \mathrm{dist}\left(\tfrac{1}{T} \sum_{t=1}^{T} \ell_t, \ S^{\mathrm{ori}} \right). \tag{2}$$

Thus $\mathcal{R}_1(T)$ measures the reward gap to the best fixed feasible routing distribution, while $\mathcal{R}_2(T)$ measures violation of the original operational constraints using realized activations.

4 OMD-APPROACHABILITY

We now develop OMD-Approachability. The method combines two structural facts about MSR. The expected sum-max reward is concave in the routing distribution and admits an unbiased importance-weighted gradient estimator, while the nonlinear expected activation cost can be controlled through the linear surrogate introduced above. These two ingredients allow reward optimization and constraint control to be performed simultaneously through a common routing distribution.

4.1 Reward and Constraint Reduction

The first ingredient is the concavity and gradient structure of the expected sum-max reward.

Lemma 1 (Concavity and gradient of Φ^r)**.** *Let r be a sum-max function. Then Φ^r is concave on Δ_K. For every $q \in \Delta_K$,*

$$\partial_i \Phi^r(q) = \mathbb{E}\left[\frac{r(B(q))}{q_i} \sum_{j=1}^{M} \mathbf{1}\{b_j(q) = i\}\right].$$

The concavity statement is Theorem 2.3 of Pasteris et al. (2024), and the gradient expression follows from their Lemma 6.7. This identity motivates the bandit gradient estimator in Algorithm 1.

For the constraints, under routing distribution q_t, action i is activated with probability $1-(1-q_{t,i})^M$. Hence

$$\mathbb{E}[\ell_t] = \sum_{i=1}^{K} \big(1-(1-q_{t,i})^M\big) c_i. \tag{3}$$

Since $p \le 1-(1-p)^M \le Mp$ for $p \in [0,1]$, every coordinate $j \in [d]$ satisfies

$$\left(\sum_{i=1}^{K} q_{t,i} c_i\right)^{(j)} \;\le\; \mathbb{E}[\ell_t]^{(j)} \;\le\; M\left(\sum_{i=1}^{K} q_{t,i} c_i\right)^{(j)}. \tag{4}$$

The rescaling in the definition of S is chosen exactly from these inequalities: upper bounds are divided by M, whereas lower bounds are unchanged.

4.2 OMD-Approachability Algorithm

We next combine online mirror descent for reward learning with Blackwell approachability for constraint control. Algorithm 1 uses the support function $h_S(w) := \max_{x \in S} w^\top x$. Because S is a compact axis-aligned box, h_S is finite and decomposes coordinate-wise:

$$h_{[0,\,u_j/M]}(w_j) = \begin{cases} (u_j/M)\, w_j & w_j \ge 0, \\ 0 & w_j < 0, \end{cases} \qquad h_{[l_j,\,\bar{c}]}(w_j) = \begin{cases} \bar{c}\, w_j & w_j \ge 0, \\ l_j w_j & w_j < 0. \end{cases}$$

By Pasteris et al. (2024), $g_{t,i}$ is an unbiased estimator of the ith coordinate of $\nabla\Phi^{r_t}(q_t)$. Its second moment grows when $q_{t,i}$ approaches zero, so we impose the following lower-bound condition on the benchmark solution.

Assumption 1. *The benchmark problem equation 1 is feasible and admits an optimal solution $q \in \Delta_K$ satisfying $q_i \ge \gamma$ for all $i \in [K]$.*

Remark 1 (feasibility of the QP). Assumption 1 also guarantees feasibility of the quadratic program equation 5. Let $q^\star$ be an optimal solution of equation 1. Its feasibility gives $\sum_{i=1}^{K} q_i^\star c_i \in S$. Lemma 4 then implies

$$w^\top \sum_{i=1}^{K} q_i^\star c_i - h_S(w) \le 0, \qquad \forall\, \|w\| \le 1.$$

The algorithm keeps w_t in the unit Euclidean ball, so the inequality holds at $w = w_t$. Assumption 1 also gives $q_i^\star \ge \gamma$ for every i. Thus $q^\star$ satisfies every constraint in equation 5, proving that the quadratic program is feasible at each round.

At each round, the dual OCO subroutine selects a direction w_t that summarizes the current constraint pressure. The primal QP then updates the routing distribution using the previous reward-gradient estimate while enforcing the supporting halfspace of S induced by w_t. After sampling the subset and observing its aggregate reward, the learner forms the importance-weighted gradient estimate used in the next primal update. The reward and constraint updates therefore share the same routing distribution and use separate feedback quantities.

Algorithm 1: Online Mirror Descent with Approachability Constraints

1. **Input:** rescaled set S, OCO algorithm $\mathcal{A}$, cost vectors $c_i \in [0, \bar{c}]^d$ for $i \in [K]$, floor γ and learning rate η.
2. Set $\mathbb{B} \subset \mathbb{R}^d$ to be the unit Euclidean ball, as decision set for $\mathcal{A}$.
3. Set $q_{0,i} := 1/K$ and initialize $g_{0,i}$ arbitrarily for $i \in [K]$.
4. For $t = 1, \ldots, T$:
   (a) Set $f_{t-1}(w) := w^\top \left(\sum_{i=1}^K q_{t-1,i} c_i \right) - h_S(w)$.
   (b) Query $\mathcal{A}$: $w_t \leftarrow \mathcal{A}(f_1, \ldots, f_{t-1})$
   (c) Set q_t as the solution of:

$$\begin{aligned} \max_{q \in \Delta_K:\, q \geq \gamma} \quad & q^\top g_{t-1} - \tfrac{\eta}{2} \|q - q_{t-1}\|^2 \\ \text{s.t. } & w_t^\top \sum_{i=1}^K q_i \, c_i - h_S(w_t) \leq 0 \end{aligned} \tag{5}$$

   (d) Draw $a_{t,1}, \ldots, a_{t,M}$ independently from q_t;
   (e) Set $A_t \leftarrow \{a_{t,j} : j \in [M]\}$;
   (f) Observe $r_t(A_t)$;
   (g) For each $i \in [K]$, define

$$g_{t,i} \leftarrow \tfrac{r_t(A_t)}{q_{t,i}} \sum_{j=1}^M \mathbb{I}\{a_{t,j} = i\}$$

Lemma 2 (Distance transfer). *For any sequence $q_1, \ldots, q_T \in \Delta_K$,*

$$\operatorname{dist}\left(\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\ell_t],\, S^{\mathrm{ori}} \right) \leq M \, \operatorname{dist}\left(\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^K q_{t,i} c_i,\, S \right).$$

The proof is coordinate-wise and is given in Appendix B.1. In particular, if the average surrogate cost lies in S, then the average expected activation cost lies in S^{ori}. This result is the link between the approachability update below and the original operational constraints.

5 Theoretical Guarantees

The following results characterize the two performance criteria defined in Section 3. The reward bound controls the gap to the surrogate-feasible benchmark, while the constraint bound transfers approachability of the surrogate set to the original realized-cost constraints.

5.1 Reward and constraint guarantees

We first state the reward and constraint guarantees. Complete proofs are given in the appendix.

Theorem 1 (Reward regret). *Let q_t for $t \in [T]$ be the output of Algorithm 1 with proximal parameter $\eta^* = \sqrt{2TKM\big(\frac{1}{\gamma} + (M-1)\big)}$. Then*

$$T \, \mathbb{E}[\mathcal{R}_1(T)] \leq 2\sqrt{2TKM\Big(\tfrac{1}{\gamma} + (M-1)\Big)}.$$

The constraint guarantee follows from the approachability update and the concentration of the realized activation costs.

Theorem 2 (Constraint regret). *Let $\mathcal{A}$ be an OCO algorithm with regret $\mathrm{Regret}_T(\mathcal{A})$. Then for an absolute constant $L > 0$,*

$$\mathbb{E}\left[\mathcal{R}_2(T)\right] \leq M \cdot \tfrac{\mathrm{Regret}_T(\mathcal{A})}{T} + M\bar{c}L\sqrt{\tfrac{d \log(2d)}{T}}.$$

With online gradient descent (OGD) as $\mathcal{A}$, $\mathrm{Regret}_T(\mathcal{A}) = O(\sqrt{T})$. Theorem 1 and Theorem 2 therefore establish $O(T^{-1/2})$ average reward regret and expected constraint violation.

5.2 Benchmark approximation from the Linear Surrogate

The main regret guarantee is relative to the surrogate-feasible benchmark $\mathrm{OPT}_{\mathrm{sur}}$. We next quantify the price of replacing the original nonlinear activation constraint by its linear surrogate.

Our algorithm enforces the linear surrogate constraint $\sum_i q_i c_i \in S$, rather than the original expected activation constraint. To quantify the resulting benchmark gap, define the best fixed routing policy under the latter constraint by

$$\mathrm{OPT}_{\mathrm{ori}} := \max_{q \in \Delta_K} \ \frac{1}{T}\sum_{t=1}^{T} \Phi^{r_t}(q) \quad \text{s.t.} \quad \sum_{i=1}^{K} \big[1-(1-q_i)^M\big] c_i \in S^{\mathrm{ori}}. \tag{6}$$

Since $q_i \le 1-(1-q_i)^M \le M q_i$, surrogate feasibility implies original expected-cost feasibility, and hence $0 \le \mathrm{OPT}_{\mathrm{ori}} - \mathrm{OPT}_{\mathrm{sur}} \le 1$.

Two-sided constraints. For a vector x, let $x^{(j)}$ denote its j-th coordinate. Let $\mathcal{U}$ and $\mathcal{L}$ index the upper and lower operational constraints. Suppose there exist $q^{\mathrm{ref}} \in \Delta_K$ and $\sigma > 0$ such that

$$M\Big(\sum_i q_i^{\mathrm{ref}} c_i\Big)^{(j)} \le u_j - \sigma \quad (j \in \mathcal{U}), \qquad \Big(\sum_i q_i^{\mathrm{ref}} c_i\Big)^{(j)} \ge l_j + \sigma \quad (j \in \mathcal{L}). \tag{7}$$

For $q \in \Delta_K$, define the surrogate constraint violation

$$\delta(q) := \max\left\{ \max_{j\in\mathcal{U}} \left[M\Big(\sum_i q_i c_i\Big)^{(j)} - u_j \right]_+ , \max_{j\in\mathcal{L}} \left[l_j - \Big(\sum_i q_i c_i\Big)^{(j)} \right]_+ \right\}, \tag{8}$$

where $[x]_+ := \max\{x, 0\}$. Thus, $\delta(q)$ is the maximum violation of the surrogate upper and lower coordinate constraints.

Proposition 1. *Under equation 7, let q^{ori} attain $\mathrm{OPT}_{\mathrm{ori}}$. Then*

$$0 \le \mathrm{OPT}_{\mathrm{ori}} - \mathrm{OPT}_{\mathrm{sur}} \le \frac{\delta(q^{\mathrm{ori}})}{\sigma + \delta(q^{\mathrm{ori}})} \left(\mathrm{OPT}_{\mathrm{ori}} - \frac{1}{T}\sum_{t=1}^{T} \Phi^{r_t}(q^{\mathrm{ref}}) \right) \le \frac{\delta(q^{\mathrm{ori}})}{\sigma + \delta(q^{\mathrm{ori}})}. \tag{9}$$

Indeed, with $\theta = \delta(q^{\mathrm{ori}})/(\sigma + \delta(q^{\mathrm{ori}}))$ and $\widetilde{q} = (1-\theta)q^{\mathrm{ori}} + \theta q^{\mathrm{ref}}$, the distribution $\widetilde{q}$ is surrogate-feasible; concavity of $\frac{1}{T}\sum_t \Phi^{r_t}$ gives equation 9.

Upper constraints only. When only upper constraints are present and a dull action is available, the sum-max structure gives a stronger, multiplicative guarantee:

Corollary 1. *Suppose all constraints are upper bounds and there is an action i_0 with $c_{i_0} = 0$ and $r_t(A \cup \{i_0\}) = r_t(A)$ for all $A \subseteq [K]$ and t. Then, for every $M \ge 2$,*

$$\frac{\mathrm{OPT}_{\mathrm{sur}}}{\mathrm{OPT}_{\mathrm{ori}}} \ge 1 - \left(1 - \frac{1}{M}\right)^M \ge 1 - e^{-1}. \tag{10}$$

The factor $1 - e^{-1}$ is the classical approximation factor that also appears monotone submodular maximization (Calinescu et al., 2011); here it arises from converting activation marginals into an iid multinomial sampling distribution. The factor is sharp; see Appendix F.2.

Together, these results separate the overall performance loss into two components: the $O(T^{-1/2})$ online optimization error relative to $\mathrm{OPT}_{\mathrm{sur}}$, and the approximation gap $\mathrm{OPT}_{\mathrm{ori}} - \mathrm{OPT}_{\mathrm{sur}}$ induced by the surrogate constraint. The latter can be controlled under a strict-feasibility margin and admits the sharper multiplicative guarantee above in the upper-bound case.

5.3 CONTEXTUAL EXTENSION

We next allow the routing policy to depend on an observed context. At round t, the learner observes $x_t \in \mathcal{X}$, where $\mathcal{X}$ is finite, before selecting the routing distribution. The context sequence can be arbitrary. The learner maintains a separate primal OMD state for each context and uses one shared dual iterate w_t so that the operational constraint is enforced on the aggregate sequence. The contextual benchmark chooses the best fixed feasible routing policy separately for each context:

$$\mathrm{OPT}^{\mathrm{ctx}} := \frac{1}{T} \sum_{x \in \mathcal{X}} \max_{\substack{q \in \Delta_K : q_i \geq \gamma \\ \sum_i q_i c_i \in S}} \sum_{t:\, x_t = x} \Phi^{r_t}(q).$$

The contextual objective regret is $\mathcal{R}_1^{\mathrm{ctx}}(T) := \mathrm{OPT}^{\mathrm{ctx}} - \frac{1}{T} \sum_{t=1}^{T} \Phi^{r_t}(q_t^{(x_t)})$. Appendix E gives the complete algorithm.

Corollary 2 (Contextual reward regret)**.** *Run the contextual algorithm (Appendix E) with per-context anytime proximal parameter $\eta_t = a\sqrt{n_{x_t}}$, where $a = 2\sqrt{KM(\frac{1}{\gamma} + M - 1)}$ and n_{x_t} is the visit count for context x_t up to round t. Then, without prior knowledge of T or $\{T^{(x)}\}$,*

$$T\,\mathbb{E}\big[\mathcal{R}_1^{\mathrm{ctx}}(T)\big] \;\leq\; 4\sqrt{|\mathcal{X}|\,TKM\Big(\tfrac{1}{\gamma} + (M-1)\Big)}.$$

Relative to the non-contextual result, the bound incurs a factor of $O(\sqrt{|\mathcal{X}|})$ from learning separate primal policies across contexts.

Remark 2 (Shared dual, decoupled primal). Algorithm 2 maintains one primal iterate $q^{(x)}$ for each $x \in \mathcal{X}$ and a single shared dual iterate w_t. Only $q^{(x_t)}$ is updated at round t, while the dual update uses the played policy $q_t = q^{(x_t)}$. The reward analysis therefore separates across contexts. Constraint control depends on the aggregate surrogate cost $\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{K} q_{t,i} c_i$, which is coupled through the shared dual process. The proof of Theorem 2 uses only this aggregate sequence and the shared dual updates, so the same constraint guarantee applies to Algorithm 2.

6 EXPERIMENTS

We construct a simulated online routing environment from the TREC 2010 Crowdsourcing Track data. The experiments are organized around the two claims established above: convergence of the objective and constraint regret, and the reward–feasibility trade-off induced by operational constraints.

6.1 EXPERIMENTAL SETUP

We use the TREC crowdsourcing data distributed with the SpectralMethodsMeetEM repository of Zhang et al. (2016). For each random seed, we sample $K = 50$ workers, route each task to $M = 5$ workers, and construct the sum-max reward from worker reliability across task groups. The two operational coordinates measure normalized workload and activation of the highest-accuracy workers. We evaluate both a stationary stochastic environment and a regime-switching environment in which worker reliability changes over time. Unless stated otherwise, the horizon is $T = 10{,}000$. Dataset preprocessing, construction of $V^{(t)}$, exact constraint thresholds, and all hyperparameters are reported in Appendix A.

We compare OMD-Approachability with Random and Greedy routing, Combinatorial UCB (CUCB) (Chen et al., 2013), Combinatorial Bandits with Knapsacks (CBwK) (Sankararaman & Slivkins, 2018; Liu et al., 2022), Thompson Sampling (TS) (Wang & Chen, 2018), a constrained TS variant (TS-BwK) (Deb et al., 2025), and sparse Mixture-of-Experts gating (MoE) (Shazeer et al., 2017). Because the dataset provides only aggregate team reward in our online simulation, all baselines are evaluated under the same feedback available to OMD-Approachability. Full implementation details are deferred to Appendix A.

6.2 REGRET CONVERGENCE

We first evaluate empirical counterparts of $\mathcal{R}_1(T)$ and $\mathcal{R}_2(T)$ under $N \in \{3, 5\}$ reward dimensions and tight, medium, and loose operational constraints. OMD-Approachability updates its routing

distribution every $B = 10$ rounds, so Figure 1 indexes the curves by the number of OMD updates, $n = t/B$, and includes an $O(n^{-1/2})$ reference. The objective quantity uses the final learned policy as an empirical proxy for the unavailable hindsight-optimal feasible policy; Appendix A gives the exact Monte Carlo construction.

Across both values of N, the empirical objective regret decreases toward zero for all three constraint levels after a short initial transient. The constraint regret separates the settings more clearly. Under the tight constraint, the running-average realized cost initially violates the target and then decreases to approximately zero as the approachability update adjusts the routing distribution. The medium configuration exhibits only a short initial violation, while the loose configuration remains feasible throughout. The same qualitative behavior is observed for $N = 3$ and $N = 5$, with slower constraint adjustment in the higher-dimensional reward instance.

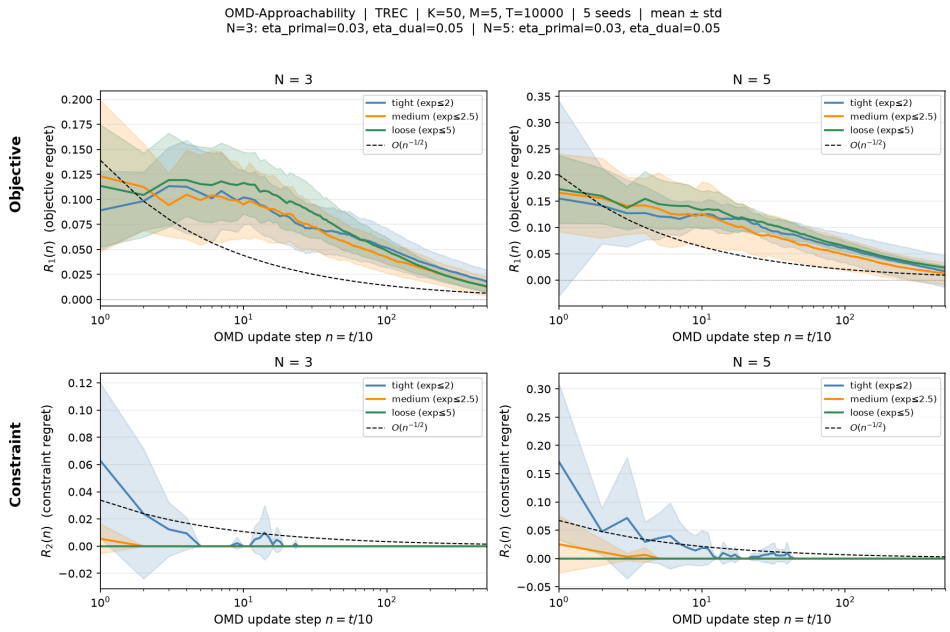

Figure 1: Empirical objective regret (top) and realized constraint regret (bottom) versus OMD update step $n = t/10$, for $N = 3$ (left) and $N = 5$ (right). Curves show the mean $\pm$ one standard deviation over 5 seeds; dashed curves show an $O(n^{-1/2})$ reference.

6.3 Reward and Feasibility under Constraint Tightness

We next compare the eight policies across tight, medium, loose, and unconstrained settings, using $N = 3$ and 20 independent seeds in both stationary and regime-switching environments. Figure 2 reports time-averaged reward together with the fraction of feasible runs, while Figure 3 shows rolling reward under the tight and loose constraints. The exact thresholds defining the four constraint levels are given in Appendix A.

Under the tight constraint, OMD-Approachability is feasible in every seed in both environments while retaining reward close to the unconstrained methods. TS-BwK is feasible in most, but not all, runs, and CBwK is substantially less reliable, especially when worker reliability switches over time. CUCB, TS, and Greedy obtain high reward but are rarely feasible because they do not control cumulative expert usage. As the constraint is relaxed, feasibility differences shrink and the competitive methods approach similar reward levels. The rolling curves show the corresponding adaptation: OMD-Approachability initially sacrifices reward while adjusting expert usage, then approaches the reward range of the unconstrained methods while maintaining feasibility. This pattern persists under regime switching, where the constraint-aware baselines exhibit larger feasibility losses.

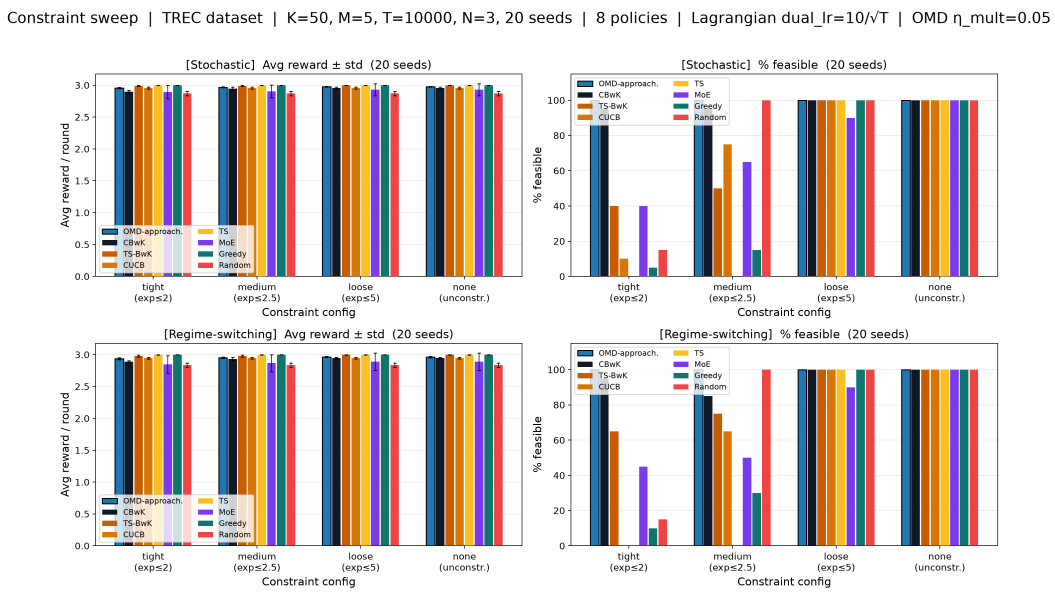

Figure 2: Constraint-tightness sweep with $K = 50$, $M = 5$, $N = 3$, $T = 10{,}000$, and 20 seeds. The left column shows mean $\pm$ standard deviation of time-averaged reward, and the right column shows the percentage of feasible seeds. Rows correspond to the stationary and regime-switching environments.

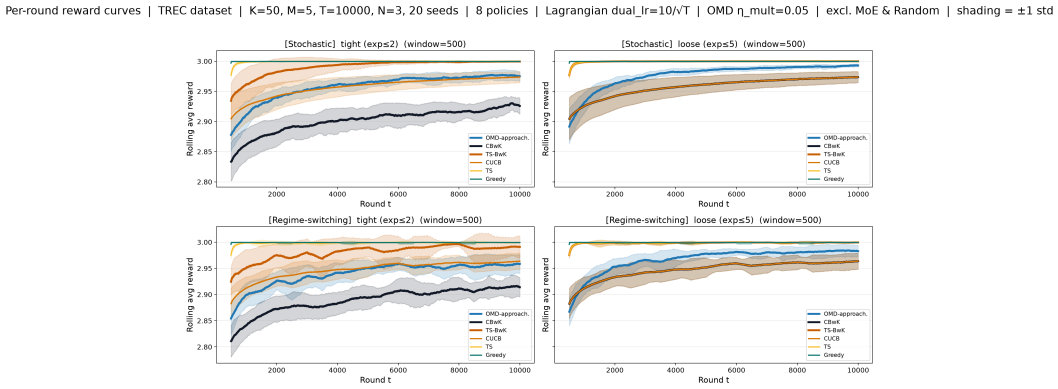

Figure 3: Rolling-average reward (window 5,000) under the tight (left) and loose (right) constraints across 20 seeds. Rows correspond to the stationary and regime-switching environments. MoE and Random are omitted because their higher per-round variance obscures the comparison among the remaining methods.

7 Conclusion

We introduced Multinomial Subset Routing for online subset routing with sum-max rewards, aggregate bandit feedback, and long-run operational constraints. The main difficulty is that repeated multinomial sampling induces nonlinear activation probabilities, while the learner observes only aggregate subset rewards. OMD-Approachability addresses these two challenges by combining online mirror descent for the concave reward with Blackwell approachability for a linear surrogate of the operational constraints. The resulting guarantees give $O(T^{-1/2})$ average objective regret and expected constraint regret relative to the surrogate-feasible benchmark, together with explicit control of the approximation gap to the original benchmark. A contextual extension is also provided, while future work includes richer contextual routing models and alternative subset-generation mechanisms that may provide tighter control of the original nonlinear activation constraints.

AI USE STATEMENT

In this work, we used generative AI tools to assist with experimental implementation, including writing and debugging code. We also used generative AI tools to assist with drafting and editing the manuscript, including parts describing the theoretical development, and to identify relevant literature. The research direction, problem formulation, and main theoretical contributions were developed by the authors. We reviewed and verified all AI-assisted text, code, citations, and experimental results, and take responsibility for the final content of this work.

REFERENCES

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in neural information processing systems*, 29, 2016.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014.

Shipra Agrawal, Nikhil R Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Gruia Calinescu, Chandra Chekuri, Martin Pal, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

Lin Chen, Andreas Krause, and Amin Karbasi. Interactive submodular bandit. *Advances in Neural Information Processing Systems*, 30, 2017.

Lingjiao Chen, Matei Zaharia, and James Y Zou. Frugalml: How to use ml prediction apis more accurately and cheaply. *Advances in neural information processing systems*, 33:10685–10696, 2020.

Lingjiao Chen, Matei Zaharia, and James Zou. Efficient online ml api selection for multi-label classification tasks. In *International conference on machine learning*, pp. 3716–3746. PMLR, 2022.

Shuhao Chen, Weisen Jiang, Baijiong Lin, James Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. *Advances in Neural Information Processing Systems*, 37:66305–66328, 2024.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 151–159, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016. URL `http://jmlr.org/papers/v17/14-298.html`.

Yu Chen, Siwei Wang, Longbo Huang, and Wei Chen. Continuous k-max bandits. *arXiv preprint arXiv:2502.13467*, 2025.

Rohan Deb, Mohammad Ghavamzadeh, and Arindam Banerjee. Thompson sampling for constrained bandits. In *Reinforcement Learning Conference*, 2025.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Tao Feng, Yanzhen Shen, and Jiaxuan You. Graphrouter: A graph-based router for llm selections. In *International Conference on Learning Representations*, volume 2025, pp. 26186–26203, 2025.

Fares Fourati, Christopher John Quinn, Mohamed-Slim Alouini, and Vaneet Aggarwal. Combinatorial stochastic-greedy bandit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12052–12060, 2024.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 100–108, Beijing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/gopalan14.html.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL http://dx.doi.org/10.1561/2400000013.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47, 2022.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pp. 535–543. PMLR, 2015.

Qingsong Liu, Weihang Xu, Siwei Wang, and Zhixuan Fang. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. *Advances in Neural Information Processing Systems*, 35:2997–3010, 2022.

Quang H Nguyen, Thinh Dao, Duy C Hoang, Juliette Decugis, Saurav Manchanda, Nitesh V Chawla, and Khoa D Doan. Metallm: A high-performant and cost-efficient dynamic framework for wrapping llms. *arXiv preprint arXiv:2407.10834*, 2024.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Stephen U. Pasteris, Alberto Rumi, Fabio Vitale, and Nicolò Cesa-Bianchi. Sum-max submodular bandits. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2323–2331. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/pasteris24a.html.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1760–1770. PMLR, 2018.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. Large language model routing with benchmark datasets. *arXiv preprint arXiv:2309.15789*, 2023.

Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4633–4656. PMLR, 2023.

Sho Takemori, Masahiro Sato, Takashi Sonoda, Janmajay Singh, and Tomoko Ohkuma. Submodular bandit problem under multiple constraints. In *Conference on Uncertainty in Artificial Intelligence*, pp. 191–200. PMLR, 2020.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 5114–5122. PMLR, 2018.

Wang Wei, Tiankai Yang, Hongjie Chen, Yue Zhao, Franck Dernoncourt, Ryan A Rossi, and Hoda Eldardiry. Learning to route llms from bandit feedback: One policy, many trade-offs. *arXiv preprint arXiv:2510.07429*, 2025.

Baosheng Yu, Meng Fang, and Dacheng Tao. Linear submodular bandits with a knapsack constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Yisong Yue and Carlos Guestrin. Linear submodular bandits and their application to diversified retrieval. *Advances in neural information processing systems*, 24, 2011.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral methods meet em: a provably optimal algorithm for crowdsourcing. *J. Mach. Learn. Res.*, 17(1):3537–3580, January 2016. ISSN 1532-4435.

APPENDIX CONTENTS

- Appendix A: Experimental Details 12
- Appendix B: Auxiliary Lemmas 13
- Appendix C: Proof of Theorem 1 15
- Appendix D: Proof of Theorem 2 16
- Appendix E: Contextual Algorithm and Proof of Corollary 2 18

A EXPERIMENTAL DETAILS

A.1 DATASET AND CONSTRUCTION

We use the TREC 2010 Crowdsourcing Track data distributed with the SpectralMethodsMeetEM repository of Zhang et al. (2016). The file `trec_crowd.txt` contains each worker's binary response to a task, and `trec_truth.txt` contains the corresponding gold label. Joining the files by task ID and retaining tasks with a gold label yields 677 workers, 2,275 tasks, and 12,863 annotations. For each annotation, we define *correct*$= \mathbf{1}\{\text{response} = \text{gold label}\}$ and compute each worker's overall accuracy. Removing workers with fewer than 5 annotations leaves 270 workers.

For each seed, we sample $K = 50$ workers uniformly without replacement. All skill and cost statistics are recomputed within this pool. We partition the 2,275 tasks into N groups by sorted task ID and set $V_{\text{base}}[n, i]$ to worker i's accuracy in group n, using the worker's overall accuracy when that group has no annotation. Each worker has two cost coordinates. The *workload* coordinate is its annotation count divided by the maximum count in the sampled pool. The *expert* coordinate equals 1 for the 10 highest-accuracy workers and 0 otherwise.

The base matrix is fixed within a seed, while the reward uses a fresh $V^{(t)}$ each round. In the *stochastic* environment, $V^{(t)}[n, i] \sim \text{Beta}(\kappa V_{\text{base}}[n, i], \kappa(1 - V_{\text{base}}[n, i]))$ independently across rounds with $\kappa = 20$. In the *regime-switching* environment, every $T/10$ rounds we sample a new set of 5 workers and multiply their mean accuracies by 0.2 for that block.

A.2 CONSTRAINT CONFIGURATIONS AND OPTIMIZATION DETAILS

Table 1 varies the expert-cap bound and the margin added to the baseline workload budget $M \cdot \overline{C_0}$. The first three configurations are used in the regret-convergence experiment, and all four are used in the policy comparison.

For the regret-convergence experiment, we set $K = 50$, $M = 5$, $T = 10{,}000$, $N \in \{3, 5\}$, and average over 5 seeds. OMD-Approachability holds q fixed for batches of $B = 10$ rounds. At each batch endpoint $t = nB$, the empirical objective quantity is

$$R_1(n) = \frac{1}{t} \sum_{s=1}^{t} \Big(\mathbb{E}_{A \sim q_T}[r(A, V^{(s)})] - r_s \Big),$$

Table 1: Constraint configurations used in the experiments.

Config	Budget margin	Expert cap α_{expert}
tight	$+0.0$	2.0
medium	$+0.1$	2.5
loose	$+0.5$	5.0
none	$-$	∞

where the expectation is estimated by 200 Monte Carlo draws for each stored $V^{(s)}$. The realized constraint quantity is

$$R_2(n) = \text{dist}\left(\frac{1}{t}\sum_{s=1}^{t}\ell_s, S^{\text{ori}}\right).$$

We use $\eta = \eta_{\text{mult}}\sqrt{2TKM(1/\gamma + M - 1)}$ with $\eta_{\text{mult}} = 0.03$, dual OGD step size $\eta_{\text{ogd}} = 0.05$, and exploration floor $\gamma = 1/(10K) = 0.002$.

For the policy comparison, we set $N = 3$ and use 20 independent seeds. OMD-Approachability again updates q every 10 rounds and uses $\eta_{\text{mult}} = 0.05$, $\eta_{\text{ogd}} = 0.05$, and $\gamma = 0.002$.

A.3 Baselines and implementation details

All UCB- and TS-based baselines use a 10-round round-robin warm-up. Since the simulation exposes only the aggregate team reward, these baselines assign the full observed reward r_t to every selected arm as a proxy for its contribution.

- **Random** selects M workers uniformly without replacement and does not use the constraints.
- **Greedy** selects the M workers with the largest empirical shared-reward means after warm-up.
- **CUCB** follows the combinatorial UCB principle of Chen et al. (2013), selecting the top-M arms according to $\hat{\mu}_i + 2\sqrt{2\log t/n_i}$.
- **CBwK** augments CUCB with a nonnegative Lagrangian multiplier for the operational constraints, following the constrained combinatorial-bandit perspective of Sankararaman & Slivkins (2018); Liu et al. (2022). The dual step size is $10/\sqrt{T}$.
- **TS** uses independent Beta posteriors and selects the top-M sampled values, following combinatorial Thompson sampling (Wang & Chen, 2018). Selected arms receive updates $\alpha_i \mathrel{+}= r_t/N$ and $\beta_i \mathrel{+}= 1 - r_t/N$.
- **TS-BwK** adds the same Lagrangian constraint adjustment to TS; see also constrained Thompson sampling (Deb et al., 2025). The dual step size is $10/\sqrt{T}$.
- **MoE** adapts sparse Mixture-of-Experts gating (Shazeer et al., 2017). A logit vector parameterizes the gate, Gumbel-Top-M produces the selected workers, and REINFORCE updates the logits from the Lagrangian-adjusted aggregate reward. We use learning rate 0.01 and dual step size 4×10^{-4}.
- **OMD-Approachability** is Algorithm 1; its routing distribution is updated by the constrained QP every 10 rounds using the importance-weighted aggregate-reward gradient, and the dual iterate is updated by OGD on the unit ball.

B Auxiliary Lemmas

This section collects the auxiliary results used in the proofs.

B.1 Proof of Lemma 2

For any sequence $q_1, \dots, q_T \in \Delta_K$, define

$$\phi := \tfrac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ell_t], \qquad \varphi := \tfrac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K} q_{t,i} c_i.$$

Because both target sets are axis-aligned, it suffices to work coordinate-wise. For an upper-bound coordinate, averaging equation 4 gives $0 \le \phi^{(j)} \le M\varphi^{(j)}$, and therefore

$$\operatorname{dist}\big(\phi^{(j)}, (-\infty, u_j]\big) = \max(\phi^{(j)} - u_j, 0) \le M \max(\varphi^{(j)} - u_j/M, 0) = M \operatorname{dist}\big(\varphi^{(j)}, [0, u_j/M]\big).$$

For a lower-bound coordinate, $\phi^{(j)} \ge \varphi^{(j)}$ and $\varphi^{(j)} \le \bar{c}$, so

$$\operatorname{dist}\big(\phi^{(j)}, [l_j, \infty)\big) = \max(l_j - \phi^{(j)}, 0) \le \max(l_j - \varphi^{(j)}, 0) = \operatorname{dist}\big(\varphi^{(j)}, [l_j, \bar{c}]\big).$$

Combining the coordinate-wise inequalities and taking Euclidean norms gives $\operatorname{dist}(\phi, S^{\mathrm{ori}}) \le M \operatorname{dist}(\varphi, S)$.

B.2 Other Lemmas

Let w_t denote the dual iterate in Algorithm 1. We use the following feasible sets:

$$\mathcal{F}_\circ := \{q \in \Delta_K : q_i \ge \gamma;\ \textstyle\sum_i q_i c_i \in S\}$$

$$\mathcal{F}_\tau := \{q \in \Delta_K : q_i \ge \gamma;\ w_t^\top \textstyle\sum_i q_i c_i - h_S(w_t) \le 0,\ \forall t \in [\tau]\},$$

The per-context sets $\mathcal{F}_\circ^{(x)}$ and $\mathcal{F}_\tau^{(x)}$ are defined analogously, with the condition over $t \in [\tau]$ restricted to rounds satisfying $x_t = x$. The sets $\mathcal{F}_\tau$ and $\mathcal{F}_\tau^{(x)}$ are random because they depend on the algorithmic iterates w_t.

For the conditional-expectation arguments, define the history available immediately before sampling A_t by

$$\mathcal{H}_{t-1} = \sigma(w_1, \dots, w_t,\ q_1, \dots, q_t,\ A_1, r_1(A_1), \dots, A_{t-1}, r_{t-1}(A_{t-1}))\,.$$

Under Algorithm 1, q_t and w_t are $\mathcal{H}_{t-1}$-measurable, while g_t is $\mathcal{H}_t$-measurable.

Lemma 3 (Three-point identity)**.** *For any $q, q_t, q_{t+1} \in \mathbb{R}^K$,*

$$\langle q_{t+1} - q_t,\ q - q_{t+1}\rangle = \tfrac{1}{2}\|q - q_t\|^2 - \tfrac{1}{2}\|q - q_{t+1}\|^2 - \tfrac{1}{2}\|q_{t+1} - q_t\|^2.$$

Lemma 4. *Let S be a convex and compact set,* $\operatorname{dist}(u, S) = \max_{\|w\|\le 1}\left\{w^\top u - h_S(w)\right\}$.

Proof. The result follows from the support-function representation in Lemma 13.5 of (Hazan, 2016). Specifically,

$$\begin{aligned}
\max_{\|w\|\le 1}\left\{w^\top u - h_S(w)\right\} &= \max_{\|w\|\le 1}\left\{w^\top u - \max_{\mathbf{x}\in S} w^\top \mathbf{x}\right\} \\
&= \max_{\|w\|\le 1}\min_{\mathbf{x}\in S}\left\{w^\top u - w^\top \mathbf{x}\right\} \\
&= \min_{\mathbf{x}\in S}\max_{\|w\|\le 1}\left\{w^\top u - w^\top \mathbf{x}\right\} \quad \text{(minimax theorem)} \\
&= \min_{\mathbf{x}\in S}\|\mathbf{x} - u\| \\
&= \operatorname{dist}(u, S).
\end{aligned}$$

□

C Proof of Theorem 1

Proof of Theorem 1. Let $q^\star \in \arg\max_{q\in\mathcal{F}_\circ} \sum_{t=1}^{T} \Phi^{r_t}(q)$. By the definition of $\mathrm{OPT}_{\mathrm{sur}}$ in equation 1 and Assumption 1,

$$\sum_{t=1}^{T} \Phi^{r_t}(q^\star) \geq T \cdot \mathrm{OPT}_{\mathrm{sur}}. \tag{11}$$

Therefore

$$T \cdot \mathrm{OPT}_{\mathrm{sur}} - \sum_{t=1}^{T} \Phi^{r_t}(q_t) \leq \sum_{t=1}^{T} \big(\Phi^{r_t}(q^\star) - \Phi^{r_t}(q_t)\big). \tag{12}$$

Step 1: one-step mirror-descent inequality. Lemma 4 and $\|w_t\| \leq 1$ imply $\mathcal{F}_\circ \subseteq \mathcal{F}_t$ for every realization. Hence $q^\star \in \mathcal{F}_t$ almost surely. Because q_{t+1} solves the proximal problem equation 5, its first-order optimality condition gives

$$\langle g_t - \eta(q_{t+1} - q_t),\, q - q_{t+1}\rangle \leq 0, \qquad \forall q \in \mathcal{F}_t.$$

Set $q = q^\star$ and apply the three-point identity in Lemma 3:

$$\langle g_t, q^\star - q_t\rangle \leq \tfrac{\eta}{2}\left(\|q^\star - q_t\|^2 - \|q^\star - q_{t+1}\|^2 - \|q_{t+1} - q_t\|^2\right) + \langle g_t, q_{t+1} - q_t\rangle.$$

Concavity of Φ^{r_t} gives

$$\Phi^{r_t}(q^\star) - \Phi^{r_t}(q_t) \leq \langle \nabla\Phi^{r_t}(q_t),\, q^\star - q_t\rangle.$$

Insert the stochastic gradient estimator g_t to obtain

$$\begin{aligned}\Phi^{r_t}(q^\star) - \Phi^{r_t}(q_t) &\leq \langle g_t, q^\star - q_t\rangle + \langle \nabla\Phi^{r_t}(q_t) - g_t,\, q^\star - q_t\rangle \\ &\leq \tfrac{\eta}{2}\left(\|q^\star - q_t\|^2 - \|q^\star - q_{t+1}\|^2 - \|q_{t+1} - q_t\|^2\right) \\ &\quad + \langle g_t, q_{t+1} - q_t\rangle + \langle \nabla\Phi^{r_t}(q_t) - g_t,\, q^\star - q_t\rangle.\end{aligned}$$

Summing over $t = 1, \ldots, T$ and telescoping the squared-distance terms yields

$$\begin{aligned}\sum_{t=1}^{T}\big(\Phi^{r_t}(q^\star) - \Phi^{r_t}(q_t)\big) \leq \tfrac{\eta}{2}\|q^\star - q_1\|^2 &+ \sum_{t=1}^{T}\langle g_t, q_{t+1} - q_t\rangle \\ &- \tfrac{\eta}{2}\sum_{t=1}^{T}\|q_{t+1} - q_t\|^2 \\ &+ \sum_{t=1}^{T}\langle \nabla\Phi^{r_t}(q_t) - g_t,\, q^\star - q_t\rangle.\end{aligned} \tag{13}$$

Step 2: stochastic-gradient error. For the final term in equation 13, Lemma 6.6 of Pasteris et al. (2024) gives

$$\mathbb{E}[g_t \mid \mathcal{H}_{t-1}] = \nabla\Phi^{r_t}(q_t).$$

Since both $q^\star$ and q_t are $\mathcal{H}_{t-1}$-measurable,

$$\mathbb{E}[\langle \nabla\Phi^{r_t}(q_t) - g_t,\, q^\star - q_t\rangle \mid \mathcal{H}_{t-1}] = 0.$$

Taking expectations and applying the tower property therefore gives

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla\Phi^{r_t}(q_t) - g_t,\, q^\star - q_t\rangle\right] = 0.$$

Step 3: second moment of the bandit gradient. Young's inequality bounds the remaining inner-product term in equation 13 as

$$\langle g_t, q_{t+1} - q_t\rangle \leq \tfrac{2}{\eta}\|g_t\|^2 + \tfrac{\eta}{2}\|q_{t+1} - q_t\|^2.$$

After summing over t, the second term on the right cancels the negative squared-step term in equation 13. Conditioning on $\mathcal{H}_{t-1}$, Lemma 6.14 of Pasteris et al. (2024) gives

$$\mathbb{E}[g_{t,i}^2 \mid \mathcal{H}_{t-1}] \le \tfrac{M}{q_{t,i}} + M(M-1).$$

Since $q_{t,i} \ge \gamma$ almost surely,

$$\mathbb{E}[g_{t,i}^2 \mid \mathcal{H}_{t-1}] \le \tfrac{M}{\gamma} + M(M-1).$$

Taking expectations and summing over i,

$$\mathbb{E}[\|g_t\|^2] = \sum_{i=1}^{K} \mathbb{E}[g_{t,i}^2] \le KM\left(\frac{1}{\gamma} + M - 1\right).$$

Summing this bound over the horizon gives

$$\sum_{t=1}^{T} \mathbb{E}[\|g_t\|^2] \le TKM\left(\frac{1}{\gamma} + M - 1\right).$$

Step 4: aggregate bound. Take expectations in equation 12 and combine the preceding bounds with equation 13. Since the squared Euclidean diameter of the simplex is at most 2, $\|q^\star - q_1\|^2 \le 2$, and we obtain

$$\begin{aligned}\mathbb{E}\left[T \cdot \mathrm{OPT}_{\mathrm{sur}} - \sum_{t=1}^{T} \Phi^{r_t}(q_t)\right] &\le \eta + \tfrac{2}{\eta} \sum_{t=1}^{T} \mathbb{E}[\|g_t\|^2] \\ &\le \eta + \tfrac{2}{\eta} TKM\left(\frac{1}{\gamma} + M - 1\right).\end{aligned}$$

Step 5: choice of proximal parameter. Minimizing the upper bound over η gives

$$\eta^\star = \sqrt{2TKM\left(\frac{1}{\gamma} + M - 1\right)}.$$

Substituting $\eta^\star$ yields

$$\mathbb{E}\left[T \cdot \mathrm{OPT}_{\mathrm{sur}} - \sum_{t=1}^{T} \Phi^{r_t}(q_t)\right] \le 2\sqrt{2TKM\left(\frac{1}{\gamma} + M - 1\right)},$$

which proves the theorem. □

D Proof of Theorem 2

We decompose the constraint regret into a sampling term and an approachability term:

$$\begin{aligned}\mathcal{R}_2(T) &= \mathrm{dist}\left(\tfrac{1}{T}\sum_{t=1}^{T} \ell_t,\ S^{\mathrm{ori}}\right) = \mathrm{dist}\left(\tfrac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K} \mathbf{1}[i \in A_t]\, c_i,\ S^{\mathrm{ori}}\right) \\ &\le \mathrm{dist}\left(\tfrac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K} \mathbf{1}[i \in A_t]\, c_i,\ \tfrac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ell_t]\right) + \mathrm{dist}\left(\tfrac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ell_t],\ S^{\mathrm{ori}}\right) \\ &\le \underbrace{\mathrm{dist}\left(\tfrac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K} \mathbf{1}[i \in A_t]\, c_i,\ \tfrac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ell_t]\right)}_{\text{sampling error}} + M \cdot \underbrace{\mathrm{dist}\left(\tfrac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K} q_{t,i}\, c_i,\ S\right)}_{\text{approachability error}},\end{aligned} \tag{14}$$

The first inequality is the triangle inequality, and the second applies Lemma 2. We bound the two terms separately.

Lemma 5 (Approachability error). *Let q_t, $t \in [T]$, be generated by Algorithm 1 with OCO subroutine $\mathcal{A}$. Then*

$$\operatorname{dist}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K} q_{t,i}\, c_i,\ S\right) \ \le\ \frac{\operatorname{Regret}_T(\mathcal{A})}{T}.$$

Proof. The argument follows Theorem 13.7 of (Hazan, 2016). Set $u_t := \sum_{i=1}^{K} q_{t,i} c_i$. Since q_t satisfies the constraint in equation 5, the dual iterate w_t obeys

$$f_t(w_t) = w_t^\top u_t - h_S(w_t) \le 0. \tag{15}$$

Let $\bar{u}_T := \frac{1}{T}\sum_{t=1}^{T} u_t$. Lemma 4 then yields

$$\begin{aligned}
\operatorname{dist}(\bar{u}_T, S) &= \max_{\|w\|\le 1} \left\{ w^\top \bar{u}_T - h_S(w) \right\} \\
&= \max_{w^* \in \mathbb{B}} \frac{1}{T}\sum_{t=1}^{T} f_t(w^*) \quad \text{(definition of } f_t\text{)} \\
&\le \frac{1}{T}\sum_{t=1}^{T} f_t(w_t) + \frac{\operatorname{Regret}_T(\mathcal{A})}{T} \quad \text{(OCO guarantee of } \mathcal{A}\text{)} \\
&\le \frac{\operatorname{Regret}_T(\mathcal{A})}{T} \quad \text{(Equation equation 15).}
\end{aligned}$$

□

Lemma 6 (Sampling error). *With probability at least $1-\delta$, for each coordinate $j \in [d]$,*

$$\left| \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K} (\mathbf{1}[i \in A_t] - \mathbb{E}[\mathbf{1}[i \in A_t]]) c_i^{(j)} \right| \le M\bar{c}\sqrt{\tfrac{2\log(2d/\delta)}{T}}.$$

Therefore, for an absolute constant $L > 0$,

$$\mathbb{E}\operatorname{dist}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K} \mathbf{1}[i \in A_t]\, c_i,\ \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ell_t]\right) \le M\bar{c}L\sqrt{\tfrac{d\log(2d)}{T}}.$$

Proof. **Step 1: high-probability bound.** Define $Y_t := \sum_{i=1}^{K} (\mathbf{1}[i \in A_t] - \mathbb{E}[\mathbf{1}[i \in A_t] \mid \mathcal{H}_{t-1}]) c_i$. For each $j \in [d]$, $\mathbb{E}[Y_t^{(j)} \mid \mathcal{H}_{t-1}] = 0$, so $(Y_t^{(j)}, \mathcal{H}_t)$ is a martingale difference sequence. Moreover, $Y_t^{(j)} \in [-M\bar{c}, M\bar{c}]$ because $c_i^{(j)} \in [0, \bar{c}]$ and $|A_t| \le M$. Let $\bar{Y} := \frac{1}{T}\sum_{t=1}^{T} Y_t$. Azuma–Hoeffding gives

$$\Pr\left(\left|\bar{Y}^{(j)}\right| \ge \epsilon\right) \le 2\exp\left(-\tfrac{T\epsilon^2}{2M^2\bar{c}^2}\right).$$

Applying a union bound over $j \in [d]$ and setting $\epsilon = M\bar{c}\sqrt{2\log(2d/\delta)/T}$ gives, with probability at least $1-\delta$,

$$\left|\bar{Y}^{(j)}\right| \le M\bar{c}\sqrt{\tfrac{2\log(2d/\delta)}{T}} \quad \text{for all } j \in [d].$$

Denote this event by $\mathcal{E}(\delta)$.

Step 2: Euclidean norm. On $\mathcal{E}(\delta)$,

$$\|\bar{Y}\|_2 = \sqrt{\sum_{j=1}^{d} (\bar{Y}^{(j)})^2} \le \sqrt{d \cdot \tfrac{2M^2\bar{c}^2\log(2d/\delta)}{T}} = \sqrt{\tfrac{2dM^2\bar{c}^2\log(2d/\delta)}{T}}.$$

Step 3: expectation.

$$\mathbb{E}\|\bar{Y}\|_2 = \int_0^\infty \mathbb{P}\big(\|\bar{Y}\|_2 > u\big)\, du.$$

Choose $\delta = \delta(u)$ so that the high-probability threshold equals u:

$$\sqrt{\tfrac{2dM^2\bar{c}^2\log(2d/\delta)}{T}} = u \implies \delta(u) = 2d\exp\Big(-\tfrac{Tu^2}{2dM^2\bar{c}^2}\Big).$$

On $\mathcal{E}(\delta)$, $\|\bar{Y}\|_2 \le u$, hence $\mathbb{P}(\|\bar{Y}\|_2 > u) \le \delta(u)$. Therefore

$$\mathbb{E}\|\bar{Y}\|_2 \le \int_0^\infty \min\Big(1,\ 2d\exp\Big(-\tfrac{Tu^2}{2dM^2\bar{c}^2}\Big)\Big)\,du.$$

Let $u_0 = \sqrt{2dM^2\bar{c}^2\log(2d)/T}$, for which $2d\exp(-Tu_0^2/(2dM^2\bar{c}^2)) = 1$. Splitting the integral at u_0 gives

$$\mathbb{E}\|\bar{Y}\|_2 \ \le\ \underbrace{\int_0^{u_0} 1\,du}_{=\,u_0} \ +\ \int_{u_0}^\infty 2d\exp\Big(-\tfrac{Tu^2}{2dM^2\bar{c}^2}\Big)\,du.$$

Apply the Gaussian tail bound $\int_\nu^\infty e^{-t^2/a^2}dt \le \frac{a^2}{2\nu}e^{-\nu^2/a^2}$ with $a^2 = 2dM^2\bar{c}^2/T$:

$$\int_{u_0}^\infty 2d\,e^{-Tu^2/(2dM^2\bar{c}^2)}\,du \le \tfrac{2d^2M^2\bar{c}^2}{Tu_0}\cdot e^{-Tu_0^2/(2dM^2\bar{c}^2)} = \tfrac{2d^2M^2\bar{c}^2}{Tu_0}\cdot\tfrac{1}{2d} = \sqrt{\tfrac{dM^2\bar{c}^2}{2T\log(2d)}}.$$

Combining the two terms,

$$\mathbb{E}\|\bar{Y}\|_2 \ \le\ \sqrt{\tfrac{2dM^2\bar{c}^2\log(2d)}{T}} + \sqrt{\tfrac{dM^2\bar{c}^2}{2T\log(2d)}} \ \le\ M\bar{c}L\sqrt{\tfrac{d\log(2d)}{T}}$$

for an absolute constant $L > 0$. □

Proof of Theorem 2. Taking expectations in equation 14 and applying Lemmas 5 and 6 proves the theorem. □

E Contextual Algorithm and Proof of Corollary 2

Algorithm 2: Contextual OMD with Approachability Constraints

1. **Input:** rescaled set S, OCO algorithm $\mathcal{A}$, context set $\mathcal{X}$, step-size constant $a = 2\sqrt{KM(\frac{1}{\gamma} + M - 1)}$
2. Set $\mathbb{B} \subset \mathbb{R}^d$ to the unit Euclidean ball, used as the decision set for $\mathcal{A}$
3. For each $x \in \mathcal{X}$: set $q_i^{(x)} := 1/K$ for $i \in [K]$, $g^{(x)} := \mathbf{0}$, and visit count $n^{(x)} := 0$
4. For $t = 1, \dots, T$:
   (a) Observe the context $x \leftarrow x_t$ and increment $n^{(x)} \leftarrow n^{(x)} + 1$
   (b) Set $f_{t-1}(w) = w^\top u_{t-1} - h_S(w)$, where $u_{t-1} = \sum_{i=1}^K q_{t-1,i}^{(x)}\, c_i$
   (c) Query $\mathcal{A}$: $w_t \leftarrow \mathcal{A}(f_1, \dots, f_{t-1})$
   (d) Set the per-context anytime proximal parameter $\eta \leftarrow a\sqrt{n^{(x)}}$
   (e) Update only the current context's iterate, using its own stored gradient $g^{(x)}$:

$$q^{(x)} \leftarrow \arg\max_{q \in \Delta_K:\, q \ge \gamma} \quad q^\top g^{(x)} - \tfrac{\eta}{2}\big\|q - q^{(x)}\big\|^2 \qquad \text{s.t. } w_t^\top \sum_{i=1}^K q_i\, c_i - h_S(w_t) \le 0 \tag{16}$$

   (f) Draw $a_{t,1}, \dots, a_{t,M}$ independently from $q^{(x)}$ and set $A_t \leftarrow \{a_{t,j} : j \in [M]\}$
   (g) Observe $r_t(A_t)$
   (h) Refresh the current context's gradient: for each $i \in [K]$, $g_i^{(x)} \leftarrow \frac{r_t(A_t)}{q_i^{(x)}} \sum_{j=1}^M \mathbb{I}\{a_{t,j} = i\}$

Proof of Corollary 2. Let $G^2 := KM(1/\gamma + M - 1)$, so $a = 2G$. Fix $x \in \mathcal{X}$, and let $T^{(x)}$ be the number of rounds for which $x_t = x$. Index these visits by $s = 1, \dots, T^{(x)}$. Restricted to

this subsequence, the iterates $\{q_s^{(x)}\}$ follow the same primal update as Algorithm 1, with proximal parameter $\eta_s^{(x)} = a\sqrt{s}$.

Apply the one-step inequality from Step 1 of the proof of Theorem 1 to the subsequence for context x, using the corresponding filtration and conditional expectation. Summing over $s = 1, \dots, T^{(x)}$ gives, for any comparator $q_\star^{(x)} \in \mathcal{F}_\circ^{(x)}$,

$$\mathbb{E}\left[\sum_{s=1}^{T^{(x)}} \left(\Phi^{r_s}(q_\star^{(x)}) - \Phi^{r_s}(q_s^{(x)})\right)\right] \le \sum_{s=1}^{T^{(x)}} \frac{\eta_s^{(x)}}{2}\left(\|q_\star^{(x)} - q_s^{(x)}\|^2 - \|q_\star^{(x)} - q_{s+1}^{(x)}\|^2\right) + \sum_{s=1}^{T^{(x)}} \frac{2G^2}{\eta_s^{(x)}}.$$

The stochastic inner-product terms have zero expectation by the same argument used in Theorem 1.

For the first sum, Abel summation and $\|q_\star^{(x)} - q_s^{(x)}\|^2 \le 2$ give

$$\sum_{s=1}^{T^{(x)}} \frac{\eta_s^{(x)}}{2}\left(\left\|q_\star^{(x)} - q_s^{(x)}\right\|^2 - \left\|q_\star^{(x)} - q_{s+1}^{(x)}\right\|^2\right) \;\le\; \eta_1^{(x)} + \sum_{s=2}^{T^{(x)}} \left(\eta_s^{(x)} - \eta_{s-1}^{(x)}\right) \;=\; \eta_{T^{(x)}}^{(x)} \;=\; a\sqrt{T^{(x)}}.$$

For the second sum, we use $\sum_{s=1}^{n} s^{-1/2} \le 2\sqrt{n}$:

$$\sum_{s=1}^{T^{(x)}} \frac{1}{\eta_s^{(x)}} = \frac{1}{a}\sum_{s=1}^{T^{(x)}} s^{-1/2} \le \tfrac{2}{a}\sqrt{T^{(x)}}.$$

Thus the expected regret for context x satisfies

$$\mathbb{E}\left[\sum_{s=1}^{T^{(x)}} \left(\Phi^{r_s}(q_\star^{(x)}) - \Phi^{r_s}(q_s^{(x)})\right)\right] \le \left(a + \tfrac{4G^2}{a}\right)\sqrt{T^{(x)}} = 4G\sqrt{T^{(x)}},$$

where the last equality uses $a = 2G$.

Finally, summing over contexts and applying Cauchy–Schwarz, $\sum_{x\in\mathcal{X}} \sqrt{T^{(x)}} \le \sqrt{|\mathcal{X}|T}$, gives

$$\begin{aligned} T\,\mathbb{E}\big[R_1^{\mathrm{ctx}}(T)\big] &\le 4G \sum_{x\in\mathcal{X}} \sqrt{T^{(x)}} \\ &\le 4G\sqrt{|\mathcal{X}|\,T} \\ &= 4\sqrt{|\mathcal{X}|\,TKM\left(\frac{1}{\gamma} + M - 1\right)}. \end{aligned}$$

This proves the corollary. □

F Suboptimality of the Surrogate-Restricted Benchmark: Proofs and Additional Bounds

For $q \in \Delta_K$, recall that $B(q)$ denotes the random subset obtained by drawing M actions independently from q, and write

$$g_M(p) := 1 - (1-p)^M. \tag{17}$$

Thus $g_M(q_i) = \Pr(i \in B(q))$. All results in this appendix concern the benchmarks in equation 6. Since the surrogate feasible set is nonempty, both benchmark optima exist: the relevant feasible sets are closed subsets of the compact simplex, and their objectives are continuous. The case $M = 1$ is exact, because the surrogate and original expected-cost constraints coincide.

F.1 Proof of Proposition 1

Proof. Write $\delta = \delta(q^{\mathrm{ori}})$, $\theta = \delta/(\sigma + \delta)$, and define

$$\widetilde{q} = (1-\theta)q^{\mathrm{ori}} + \theta q^{\mathrm{ref}}.$$

For an upper constraint, affinity of the surrogate cost in q implies

$$M\big(\sum_i \widetilde{q}_i c_i\big)^{(j)} - u_j = (1-\theta)\big(M\big(\sum_i q_i^{\mathrm{ori}} c_i\big)^{(j)} - u_j\big) + \theta\big(M\big(\sum_i q_i^{\mathrm{ref}} c_i\big)^{(j)} - u_j\big) \le (1-\theta)\delta - \theta\sigma = 0.$$

For a lower constraint, the same calculation gives

$$l_j - \big(\sum_i \widetilde{q}_i c_i\big)^{(j)} \le (1-\theta)\delta - \theta\sigma = 0.$$

Thus $\sum_i \widetilde{q}_i c_i \in S$. By Lemma 1, $\frac{1}{T}\sum_{t=1}^{T} \Phi^{r_t}$ is concave, and therefore

$$\mathrm{OPT}_{\mathrm{sur}} \ge \frac{1}{T}\sum_{t=1}^{T} \Phi^{r_t}(\widetilde{q}) \ge (1-\theta)\mathrm{OPT}_{\mathrm{ori}} + \theta\frac{1}{T}\sum_{t=1}^{T} \Phi^{r_t}(q^{\mathrm{ref}}).$$

Rearranging proves the first upper bound in equation 9. Because q^{ref} is also original-feasible and rewards lie in $[0,1]$, $0 \le \mathrm{OPT}_{\mathrm{ori}} - \frac{1}{T}\sum_{t=1}^{T} \Phi^{r_t}(q^{\mathrm{ref}}) \le 1$, which proves the second bound. If $\delta = 0$, then $\theta = 0$ and the same argument gives $\mathrm{OPT}_{\mathrm{ori}} - \mathrm{OPT}_{\mathrm{sur}} = 0$. □

F.2 Proof and sharpness of Corollary 1

Proof. **Step 1.** Let q^{ori} be an original optimizer and let $p_i := g_M(q_i^{\mathrm{ori}}) = \Pr(i \in B(q^{\mathrm{ori}}))$ be its activation marginals, i.e., the probability that action i appears at least once among the M iid draws. Let $I_i := \mathbf{1}\{i \in B(q^{\mathrm{ori}})\}$. Since at most M distinct actions are selected, we have:

$$\sum_i p_i = \sum_i \mathbb{E}[I_i] = \mathbb{E}\left[\sum_i I_i\right] = \mathbb{E}[|B(q^{\mathrm{ori}})|] \le M.$$

Let i_0 be the dull action with $c_{i_0} = \mathbf{0}$ and $r_t(A \cup \{i_0\}) = r_t(A)$ for every $A \subseteq [K]$. Let $\mathbf{e}_{i_0}$ denote the point mass at i_0, and construct a new probability distribution

$$\widetilde{q} := \frac{p}{M} + \left(1 - \frac{1}{M}\sum_i p_i\right)\mathbf{e}_{i_0}. \tag{18}$$

Because the dull action induces zero cost $c_{i_0} = \mathbf{0}$,

$$M\sum_i \widetilde{q}_i c_i = \sum_i p_i c_i = \sum_i g_M(q_i^{\mathrm{ori}}) c_i. \tag{19}$$

Since q^{ori} is the optimizer of $\mathrm{OPT}_{\mathrm{ori}}$, we have $\big(\sum_i g_M(q_i^{\mathrm{ori}})c_i\big)^{(j)} \le u_j$ for all upper original constraints $j \in \mathcal{U}$. So $\widetilde{q}$ satisfies every upper surrogate constraint.

Step 2. We next prove the reward comparison. For simplicity, we first assume that the sum-max representation is nonnegative, i.e., $V_{t,k,i} \ge 0$ for all t, k, i, and remove this assumption afterward. Then, $r_t(\{i_0\}) = 0$ implies $V_{t,k,i_0} = 0$ for all t, k.

Fix t, k, and a threshold $z > 0$, and define $D_z := \{i \in [K] : V_{t,k,i} \ge z\}$. This set excludes i_0. Write $\rho_z := \Pr(B(q^{\mathrm{ori}}) \cap D_z \ne \varnothing)$. The union bound gives $\rho_z \le \sum_{i \in D_z} p_i$. Then, we have

$$\Pr(B(\widetilde{q}) \cap D_z \ne \varnothing) = g_M\left(\frac{1}{M}\sum_{i \in D_z} p_i\right) \ge g_M\left(\frac{\rho_z}{M}\right) \ge \left(1 - \left(1 - \frac{1}{M}\right)^M\right)\rho_z. \tag{20}$$

The equality uses equation 18, the definition of $\widetilde{q}$ and $i_0 \notin D_z$. The middle inequality uses $\rho_z \le \sum_{i \in D_z} p_i$ and that the map $x \mapsto g_M(x)$ is increasing on $[0,1]$. The last inequality follows from the concavity of $x \mapsto g_M(x/M)$ on $[0,1]$ and the endpoint values $g_M(0) = 0$ and $g_M(1/M) = 1 - (1-1/M)^M$, which imply

$$g_M(x/M) \ge x g_M(1/M) + (1-x)g_M(0/M) = \left(1 - \left(1 - \frac{1}{M}\right)^M\right)x.$$

Step 3. For $X_q := \max_{i \in B(q)} V_{t,k,i} \geq 0$, the layer-cake representation and Tonelli's theorem give

$$\mathbb{E}[X_q] = \int_0^\infty \Pr(B(q) \cap D_z \neq \varnothing)\, dz.$$

Therefore, by equation 20,

$$\mathbb{E}\left[\max_{i \in B(\widetilde{q})} V_{t,k,i}\right] \geq \left(1 - \left(1 - \frac{1}{M}\right)^M\right) \int_0^\infty \rho_z\, dz = \left(1 - \left(1 - \frac{1}{M}\right)^M\right) \mathbb{E}\left[\max_{i \in B(q^{\mathrm{ori}})} V_{t,k,i}\right].$$

Summing over k and averaging over t yields

$$\frac{1}{T}\sum_{t=1}^{T} \Phi^{r_t}(\widetilde{q}) \geq \left(1 - \left(1 - \frac{1}{M}\right)^M\right) \frac{1}{T}\sum_{t=1}^{T} \Phi^{r_t}(q^{\mathrm{ori}}).$$

Since the same $\widetilde{q}$ satisfies this inequality for every t, this proves $equation$ 10.

Step 4. The inequality $(1 - 1/M)^M \leq e^{-1}$ follows from $\log(1 - x) \leq -x$ for $0 < x < 1$; the case $M = 1$ is immediate. □

Lifting the nonnegative assumption. Even without assuming $V_{t,k,i} \geq 0$, the presence of the dull action i_0 allows an equivalent nonnegative sum-max representation. Since $r_t(\{i_0\}) = 0$, we have $\sum_k V_{t,k,i_0} = 0$. Also,

$$0 = r_t(\{i, i_0\}) - r_t(\{i\}) = \sum_k [V_{t,k,i_0} - V_{t,k,i}]_+,$$

so $V_{t,k,i} \geq V_{t,k,i_0}$. Hence

$$W_{t,k,i} := V_{t,k,i} - V_{t,k,i_0} \geq 0, \qquad W_{t,k,i_0} = 0,$$

and W gives the same reward representation.

Sharpness. Take $K = 2$, let action 2 be the dull action, and use $c_1 = 1$, $c_2 = 0$, $u = 1$, and $r_t(A) = \mathbf{1}\{1 \in A\}$. The original optimum is 1, attained by $q = (1, 0)$. Surrogate feasibility requires $q_1 \leq 1/M$, so $\mathrm{OPT}_{\mathrm{sur}} = g_M(1/M)$. Thus the factor cannot be improved uniformly under the corollary's assumptions. For $M \geq 2$, this example also satisfies the existing Assumption 1 whenever $0 < \gamma \leq 1/M$.