Learning to Activate and Scale: Resource Allocation with Semi-Bandit Feedback

Jiawen Yang[†] Xingyu Lu[†,§] Yuxiao Wen[¶] Caihua Chen[†]

Wei Lu[§] Zhengyuan Zhou[‡] Jun Zhou[§]

[†] School of Management and Engineering, Nanjing University, {jwyang, luxingyu.lxy}@smail.nju.edu.cn, chchen@nju.edu.cn

[§] Ant Group, {sing.lxy, xiaobo.lw, jun.zhoujun}@antgroup.com

[¶] Courant Institute of Mathematical Sciences, New York University, yuxiaowen@nyu.edu

[‡] Stern School of Business, New York University, zzhou@stern.nyu.edu

We study how platforms allocate a recurring budget between activating activities and scaling investment when reward means are unknown. Unit minimum commitments, square-root rewards, and one observation per activation make learning compete with current reward for resources. Under a conditional expected per-period budget, randomized activation yields an exact concave formulation. Separate activation and scaling thresholds lead to PriceResolve for computing optimal allocations. Variance-Aware UCB (VA-UCB) uses mean-dependent variance bounds. With fixed tuning coefficients, it achieves near-minimax square-root worst-case expected regret in the stated scarce-budget regime and polylogarithmic regret on fixed positive-price instances without activation ties. To reduce repeated optimization, Noise-Controlled Mirror Descent (NCMD) incrementally updates and rescales allocations. For scarce budgets and fixed tuning, it attains the same polynomial order of worst-case expected regret with linear per-period computation. A horizon-dependent two-instance construction near an activation threshold gives a cube-root expected-regret lower bound matched by VA-UCB up to logarithms. The structure and guarantees extend to heterogeneous minima and concave power rewards, with separate analyses for the activation-only and linear endpoints. Across three Alipay-calibrated users and three budgets, VA-UCB has lower terminal mean regret than NCMD. NCMD computes faster in these settings and at every tested synthetic activity count. Minimum commitments support exploration and can remain optimal with known means. Further investment depends on estimated marginal reward relative to the resource price.

Key words: resource allocation; semi-bandit learning; minimum commitments

1. Introduction

Platforms allocate limited resources by choosing activities and their investment levels. In incentive advertising, they choose tasks such as watching a video or opening an app and set offers between

eligible minima and upper limits. Larger offers can increase an accepted task's completion probability, but additional gains may diminish. Online task experiments report diminishing productivity gains at high incentives (Mourelatos et al. 2025). Another task may then become attractive, but its minimum commitment consumes resources available to scale an existing task. Similar choices arise between assortment variety and inventory per product (Rajaram 2001), or between the number of processes receiving cache and the amount assigned to each (Lattimore et al. 2014). Information search also trades off alternatives examined against investment in each, although additional investment there gathers more evidence (Moreno-Bote et al. 2020).

With unknown reward means, allocation also determines learning. We take the response shape as known and learn each activity's mean from semi-bandit feedback (Chen et al. 2016, Kveton et al. 2015). Each activation yields one observation after allocation. Scaling can increase expected reward without increasing the observation count, so learning about another activity competes for resources. This feedback reveals more than aggregate portfolio reward but requires investing in the observed activities (Slivkins 2019). Inactive activities remain unobserved. Graph feedback can reveal linked activities' outcomes (Wen 2025), while observe-then-decide models reveal a request's reward and requirements before allocation (Li and Ye 2022, Balseiro et al. 2025). Our platform commits resources before observing current outcomes.

The shared budget covers minimum commitments and further investment. It renews each period (Fontaine et al. 2020), whereas bandits with knapsacks consume a finite stock over the horizon (Badanidiyuru et al. 2018, Sankararaman and Slivkins 2018). Activities compete for current resources, and observations inform later allocations. Ranking estimated reward means is a natural starting point. With known means, ranking suffices for additive rewards under a cardinality constraint or divisible linear rewards with equal unit costs. Under diminishing returns, additional resources must be assessed at their current marginal reward. New information may justify reducing or stopping one activity to fund another whose minimum commitment is worthwhile. A correct ranking therefore leaves allocation depth unresolved.

Existing methods use different action spaces or information. Continuous allocation optimizes concave returns (Einbu 1978), while our positive allocations must meet minimum commitments. The mirror descent method of Wen (2025) selects discrete subsets with graph feedback. Price-based observe-then-decide methods use reward information available before allocation (Li and Ye 2022, Balseiro et al. 2025). Existing regret guarantees cover discrete selection (Kveton et al. 2015) and continuous allocation with noisy gradients (Fontaine et al. 2020). Here, minimum commitments couple selection, depth, and observation opportunities. Earlier allocations determine the available observations, and the resulting estimation errors affect future activation and depth choices.

We optimize expected allocations over randomized viable actions under a conditional expected per-period budget, permitting fluctuations in realized use. A common resource price coordinates activation and scaling by comparing reward per unit at activation with marginal reward from further investment. Activation feedback supports two approaches: recomputing optimistic allocations, or incrementally updating and rescaling them within the budget and capacities. These approaches make joint activation and scaling decisions at different computation costs.

1.1. Main Contributions

Minimum commitments create distinct activation and scaling thresholds. The resulting structure separates their estimation errors and guides two learning policies whose guarantees and calibrated comparisons quantify reward and computation costs.

1. **An exact allocation structure and a computation oracle.** Randomized activation converts square-root rewards with unit minimum commitments and a conditional expected budget into an exact concave problem. Its optimal dual price separates activation and scaling thresholds and, away from equalities, four regimes: inactive, minimum-only, interior, and capped. PriceResolve computes an optimal allocation and supporting price. Separate counterexamples show that two ranking rules earn a vanishing fraction of optimal reward as the activity count grows and incur linear regret even as estimates improve.

2. **Learning through optimistic allocation and incremental updates.** Variance-Aware UCB (VA-UCB) calls PriceResolve with optimistic means using variance bounds that shrink at low means. Noise-Controlled Mirror Descent (NCMD) reduces repeated optimization by updating allocations using estimated marginal rewards and rescaling them through a projection threshold. An empirical reward reference, its support-based lower bound, and decreasing steps control estimation noise. Under scarce budgets with fixed tuning coefficients, both attain the same polynomial order of worst-case expected regret. NCMD uses linear per-period unit-cost arithmetic, while VA-UCB sorts price breakpoints. For fixed tuning and a fixed positive-price instance, VA-UCB's high-probability bound separates mistaken activations, scaling errors, and total-allocation errors at activation ties. Their upper bounds grow at most logarithmically, as the square of the logarithm, and at a cube-root rate up to logarithms, respectively. Without activation ties, the last term vanishes. The structure and guarantees extend to heterogeneous minimum commitments and concave power rewards.

3. **Global and local limits on learning performance.** Learning which activities to activate can force square-root worst-case expected regret in the product of activity count, budget, and horizon. Both policies match this order up to logarithms over the integer scarce-budget and horizon range

in Theorem 4 and Corollary 1. Locally, nearby means can imply different optimal activations, while distinguishing them consumes resources. A horizon-dependent pair near an activation boundary forces cube-root expected regret on at least one instance for every policy. VA-UCB matches this local rate up to logarithms on that pair.

4. **Allocation robustness and calibrated numerical evidence.** Within fixed strict regimes at a positive price, sufficiently small estimation errors preserve zero, minimum, and capacity allocations. Unequal proportional errors change relative allocations among activities that remain interior, with slower diminishing returns amplifying the effect. Across three users calibrated from Alipay task-completion predictions and three budgets, VA-UCB has lower terminal mean cumulative regret than NCMD, which computes faster. NCMD is also faster at every tested synthetic activity count. These comparisons document the reward and computation tradeoff between exact optimistic allocation and incremental updates.

1.2. Literature Review

We review resource-allocation learning and semi-continuous allocation to clarify the roles of feedback and minimum commitments. We then examine first-order and dual-price coordination, followed by models of how past exposure changes rewards.

1.2.1. Resource allocation and semi-bandit learning. Resource-allocation bandits learn rewards at different investment levels. Lattimore et al. (2014) and Dagan and Crammer (2018) study success probabilities increasing linearly to one at unknown resource thresholds. Fontaine et al. (2020) adapt to the regularity of unknown concave returns using noisy gradients. Zuo and Joe-Wong (2021) use semi-bandit feedback and an offline approximation oracle for discrete or continuous allocations. Our known response shape lets one activation observation inform every feasible level, linking estimation errors across levels to activation frequency.

Upper-confidence-bound (UCB) methods use optimistic means to select arms (Auer et al. 2002, Slivkins 2019), including in contextual recommendation (Li et al. 2010) and structured subset selection (Kveton et al. 2015, Combes et al. 2015). Chen et al. (2016) allow nonlinear rewards and probabilistically triggered arms. Optimism with an allocation oracle is established. Time-uniform Bernstein concentration is also available (Howard et al. 2021). We apply them to bounded observations selected by earlier allocations. The distribution-dependent bound of Chen et al. (2016) uses gaps among finitely many super arms and minimum positive triggering probabilities. Our continuous allocations admit arbitrarily small reward gaps and activation probabilities. Gaps below the endogenous activation threshold, curvature above the minimum, and total-allocation errors at ties separate the cost of observations from depth-adjustment loss. Our conditional expected budget renews each period, while

bandits with knapsacks share resources across periods (Badanidiyuru et al. 2018, Sankararaman and Slivkins 2018).

1.2.2. Semi-continuous resource allocation. Semi-continuous allocations are zero or between a minimum and a capacity (Sun et al. 2013). Einbu (1978) characterizes continuous concave allocation, and Little (1979) reviews concave and S-shaped advertising responses. We combine concave returns with minimum commitments, so scaling one activity must be compared with activating another.

Minimum commitments also determine observation costs. Bhaskaran and Erat (2025) study sequential prototype testing with measurement precision chosen to balance cost and accuracy. Here, meeting the minimum yields one observation. Additional investment can increase expected reward without increasing the observation count, making the minimum both an operating requirement and a cost of learning.

1.2.3. First-order and dual-price online control. Learning can exploit operating-policy or reward structure. Tang et al. (2024) learn demand within a dual-index inventory policy, while Galgana and Golrezaei (2025) use a utility decomposition for repeated multiunit bidding.

In adversarial combinatorial semi-bandits with graph feedback, Wen (2025) uses a shifted importance-weighted estimator in an entropy update. Reference values with importance-weighted residuals also appear in policy evaluation (Dudík et al. 2011), and decreasing steps are standard in stochastic mirror descent (Nemirovski et al. 2009). Our allocations determine observation probabilities that can approach zero. NCMD estimates marginal rewards using the known response slope and controls large positive updates with an empirical reference and a support-based lower bound, permitting incremental updates from activation feedback.

The updates must satisfy the shared budget and heterogeneous capacities. Capping and rescaling, including median-based linear-time computation, have precedents in entropy projection (Warmuth and Kuzmin 2008). CappedScale applies this computation to our feasible allocations.

Dual prices coordinate shared resources. Balseiro et al. (2014) jointly optimize advertising reservation contracts and ad-exchange sales. Online linear programming and regularized allocation use prices when rewards and requirements are known before allocation (Li and Ye 2022, Balseiro et al. 2025). Our decisions precede current reward observations. VA-UCB uses the economic dual price of its optimistic problem, while NCMD uses a projection threshold to enforce the budget and capacities. Balseiro et al. (2025) regularize total consumption for goals such as fairness. Our diminishing returns apply within activities, so reward itself can favor spreading resources.

1.2.4. Repeated interaction and diminishing returns. Our diminishing returns concern investment within a period. Repeated-exposure models let past selections affect current rewards through wear-in and wear-out (Agrawal and Tulabandhula 2020) or advertising spillover and carryover (Chen et al. 2025). Our repeated activations estimate fixed reward means. Expected reward at a fixed allocation does not depend on exposure or cross-channel history.

1.3. Organization

Section 2 defines the model and develops its optimal allocation benchmark. Section 3 presents the online policies. Section 4 establishes their regret guarantees. Section 5 treats heterogeneous minimum commitments, power rewards, and their endpoint models. Section 6 reports numerical experiments calibrated with real operational data. Section 7 concludes. The electronic companion provides proofs, implementation details, and additional numerical results.

2. Model and Optimal Allocation

Section 2.1 defines decisions, feedback, and regret. Section 2.2 derives the exact concave benchmark and its randomized implementation. The allocation structure in Section 2.3 leads to the PriceResolve oracle in Section 2.4.

2.1. Model and Performance Measure

A platform allocates a shared resource among K activities, or arms, over T periods. Arm a has known capacity $n_a \geq 1$ and unknown reward mean $\mu_a \in [0,1]$, collected in vectors $\boldsymbol{n}$ and $\boldsymbol{\mu}$. The per-period budget satisfies $0 < B < \sum_a n_a$, so all activities cannot operate at capacity every period. Implemented allocations satisfy

$$Y_{t,a} \in \{0\} \cup [1, n_a].$$

Zero keeps an activity inactive. Activation requires a unit minimum commitment and permits investment up to capacity.

Write $\boldsymbol{Y}_t = (Y_{t,1}, \ldots, Y_{t,K})$ and $A_{t,a} = \mathbf{1}\{Y_{t,a} > 0\}$ for allocations and activation indicators. The platform first chooses the complete allocation from past decisions and observations and current internal randomization U_t. It then receives one observation $X_{t,a} \in [0,1]$ per activated activity. Multiplying that observation by the known response gives reward. Inactive arms use $A_{t,a}X_{t,a} = 0$ as a history placeholder. Define

$$\mathscr{H}_{t-1} := \left(U_s, \boldsymbol{Y}_s, (A_{s,a}X_{s,a})_{a=1}^{K}\right)_{s<t}.$$

Conditional on this history, expected resource use satisfies

$$\mathbb{E}\left[\sum_{a=1}^{K} Y_{t,a} \mid \mathscr{H}_{t-1}\right] \leq B. \tag{1}$$

This budget controls expected commitments at each decision while permitting realized use above B. It suits operations that absorb such fluctuations, is stronger than an unconditional horizon average, and differs from a hard budget limiting every realized allocation. A policy π is admissible if its choice of $\boldsymbol{Y}_t$ depends only on $(\mathscr{H}_{t-1}, U_t)$ and satisfies (1) in every period. Let $\Pi_K(B, \boldsymbol{n})$ denote the set of admissible policies for K activities with budget B and capacities $\boldsymbol{n}$.

Conditional on this allocation and the preceding history, feedback has mean μ_a:

$$\mathbb{E}[A_{t,a} X_{t,a} \mid \mathscr{H}_{t-1}, U_t, \boldsymbol{Y}_t] = \mu_a A_{t,a}.$$

The reward of a single period t is

$$R_t = \sum_{a=1}^{K} X_{t,a} \sqrt{Y_{t,a}} A_{t,a}.$$

An active arm receiving y units has expected reward $\mu_a \sqrt{y}$. The response is known and μ_a is fixed over time. This canonical model isolates activation and scaling: scaling can increase expected reward, while activations determine observation counts.

For the known-mean benchmark, define the implemented action space $\mathcal{S} := \prod_{a=1}^{K} (\{0\} \cup [1, n_a])$ and let $\Delta(\mathcal{S})$ be its set of probability distributions. The optimal expected reward in one period is

$$\begin{aligned} J^* := \sup_{\nu \in \Delta(\mathcal{S})} \quad & \mathbb{E}_{\boldsymbol{Y} \sim \nu} \left[\sum_{a=1}^{K} \mu_a \sqrt{Y_a}\, \mathbf{1}\{Y_a > 0\} \right] \\ \text{s.t.} \quad & \mathbb{E}_{\boldsymbol{Y} \sim \nu} \left[\sum_{a=1}^{K} Y_a \right] \le B, \end{aligned} \tag{2}$$

where ν is the distribution of the one-period allocation $\boldsymbol{Y} = (Y_1, \ldots, Y_K)$. For any $\pi \in \Pi_K(B, \boldsymbol{n})$, measure loss against this one-period optimum through cumulative conditional expected regret:

$$\mathcal{R}_\pi(T) = \sum_{t=1}^{T} \left[J^* - \mathbb{E}_\pi[R_t \mid \mathscr{H}_{t-1}] \right]. \tag{3}$$

This reward loss depends on the random histories $\mathscr{H}_{t-1}$. Its expectation averages over feedback and policy randomization.

For positive integers m, write $[m] = \{1, \ldots, m\}$. The sets $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ contain real, nonnegative, and positive numbers, respectively. Bold symbols denote vectors. Vector inequalities are coordinate-wise and $\mathbf{0}$ is the zero vector. We use $|S|$ or $\#S$ for cardinality, $\lfloor x \rfloor$ for the floor, and $x \wedge y = \min\{x, y\}$, $x \vee y = \max\{x, y\}$ for scalars. Expectation is $\mathbb{E}$ with a subscript for the instance or policy. The symbol $\mathbf{1}\{\cdot\}$ is an event indicator. The operators arg min and arg max give sets of minimizers and maximizers. The orders $O(\cdot)$ and $\Omega(\cdot)$ denote asymptotic upper and lower bounds up to positive constant factors, and $\widetilde{O}(\cdot)$ also suppresses logarithmic factors. Instance-dependent bounds may have constants depending on fixed instance parameters. Universal constants are identified explicitly. Logarithms are natural with base e, so $\log(eK) = 1 + \log K > 0$ even at $K = 1$.

2.2. Exact Concave Reformulation

For expected allocation $x \in [0, n_a]$, the largest attainable expected square-root response is

$$h_a(x) = \begin{cases} x, & 0 \le x \le 1, \\ \sqrt{x}, & 1 \le x \le n_a. \end{cases}$$

On the activation layer $0 \le x \le 1$, allocate one unit with probability x and zero otherwise. On the depth layer $1 < x \le n_a$, allocate x with certainty. Expected allocations belong to

$$\mathcal{X} = \left\{ \boldsymbol{x} \in \mathbb{R}_+^K : 0 \le x_a \le n_a,\ \sum_{a=1}^{K} x_a \le B \right\},$$

and the benchmark has the following concave representation:

$$J^* = \max_{\boldsymbol{x} \in \mathcal{X}} \sum_{a=1}^{K} \mu_a h_a(x_a). \tag{4}$$

PROPOSITION 1 **(Exact randomized benchmark).** *The benchmark in (2) equals the value of (4). For every $\boldsymbol{x} \in \mathcal{X}$, there is a randomized implemented allocation $\boldsymbol{Y}$ such that the following equalities hold for every arm a:*

$$\mathbb{E}[Y_a] = x_a, \qquad \mathbb{E}[\sqrt{Y_a}\,\mathbf{1}\{Y_a > 0\}] = h_a(x_a).$$

Proposition 1 makes (4) attainable. Online, choose $\boldsymbol{x}_t \in \mathcal{X}$ from $\mathscr{H}_{t-1}$ and implement its conditional expectations $x_{t,a} = \mathbb{E}[Y_{t,a} \mid \mathscr{H}_{t-1}]$ by

$$Y_{t,a} = \begin{cases} 0, & x_{t,a} = 0, \\ \text{Bernoulli}(x_{t,a}) \cdot 1, & 0 < x_{t,a} < 1, \\ x_{t,a}, & 1 \le x_{t,a} \le n_a. \end{cases} \tag{5}$$

For example, $x_{t,a} = 0.2$ activates the activity at unit scale with probability 0.2.

Write $x_{t,a}^{\pi} := \mathbb{E}_{\pi}[Y_{t,a} \mid \mathscr{H}_{t-1}]$. Under (5), activity a's conditional expected reward equals $\mu_a h_a(x_{t,a}^{\pi})$, so (3) becomes

$$\mathcal{R}_{\pi}(T) = \sum_{t=1}^{T} \left[J^* - \sum_{a=1}^{K} \mu_a h_a(x_{t,a}^{\pi}) \right].$$

2.3. Optimal Allocation Structure

A common resource price coordinates activation and scaling. To also guide learning, replace $\boldsymbol{\mu}$ in (4) by nonnegative estimated means $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. The optimal expected allocations are

$$\mathcal{X}^*(\boldsymbol{\theta}) = \arg\max \left\{ \sum_{a} \theta_a h_a(x_a) : 0 \le x_a \le n_a,\ \sum_{a} x_a \le B \right\}.$$

The dual minimizes the budget charge plus maximized rewards net of resource costs. Among possibly multiple optimal multipliers, select the largest:

$$\lambda(\boldsymbol{\theta}) := \max \arg\min_{\lambda \geq 0} \left\{ \lambda B + \sum_{a=1}^{K} \max_{0 \leq x \leq n_a} \left(\theta_a h_a(x) - \lambda x \right) \right\}.$$

This is the value function's left derivative in budget, measuring marginal reward lost when resources decrease. At $\lambda > 0$, define

$$\mathcal{D}_a(\theta_a, \lambda) := \arg\max_{0 \leq x \leq n_a} \{\theta_a h_a(x) - \lambda x\},$$

the allocations maximizing activity a's reward net of λx. A common optimal price combines these choices into a feasible portfolio.

PROPOSITION 2 **(Optimal allocation structure).** *For every $\boldsymbol{\theta} \geq \mathbf{0}$, there is $\boldsymbol{x}^* \in \mathcal{X}^*(\boldsymbol{\theta})$ with the following properties.*

(i) At any positive price, the individual optimal allocation sets are

$$\mathcal{D}_a(\theta_a, \lambda) = \begin{cases} \{0\}, & \theta_a < \lambda, \\ \{1\}, & \lambda < \theta_a < 2\lambda, \\ \left\{ \left(\dfrac{\theta_a}{2\lambda} \right)^2 \right\}, & 2\lambda < \theta_a < 2\sqrt{n_a}\lambda, \\ \{n_a\}, & \theta_a > 2\sqrt{n_a}\lambda. \end{cases} \tag{6}$$

At the thresholds, $\mathcal{D}_a(\lambda, \lambda) = [0, 1]$, $\mathcal{D}_a(2\lambda, \lambda) = \{1\}$, and $\mathcal{D}_a(2\sqrt{n_a}\lambda, \lambda) = \{n_a\}$. If $\lambda(\boldsymbol{\theta}) > 0$, then $x_a^* \in \mathcal{D}_a(\theta_a, \lambda(\boldsymbol{\theta}))$ for every arm a, with $\sum_a x_a^* = B$.

(ii) If $\lambda(\boldsymbol{\theta}) = 0$, then $x_a^* = n_a$ whenever $\theta_a > 0$, with $\sum_a x_a^* = B \wedge \sum_a n_a$.

(iii) The optimal dual price is coordinatewise nondecreasing: $\boldsymbol{\theta} \leq \boldsymbol{\theta}'$ implies $\lambda(\boldsymbol{\theta}) \leq \lambda(\boldsymbol{\theta}')$.

The largest optimal price is zero when the budget exceeds the positive-mean activities' total capacity and positive otherwise. Increasing an estimated mean can raise this price and change other allocations. With known means, set $\boldsymbol{\theta} = \boldsymbol{\mu}$ and $\lambda^* = \lambda(\boldsymbol{\mu})$ in Proposition 2, including its threshold equalities. Redistributing a fixed total among activation-tied arms within $[0, 1]$ preserves reward, although the budget can determine a unique allocation to them. When $n_a = 1$, minimum-only and capped allocations coincide.

A unit activation earns μ_a, but its marginal reward just above the minimum is $\mu_a/2$ when capacity permits scaling. Thus $\lambda^* < \mu_a < 2\lambda^*$ makes activation worthwhile while further investment falls below the resource's opportunity cost. An activity can optimally remain at its minimum.

The distinct thresholds expose two limits of ranking. Let $U_{t,a}$ be an arm's upper-confidence estimate from the rule's own activation history, and break ties by fixed activity order. UCB Rank-and-Fill scans decreasing estimates and assigns the smaller of capacity and remaining budget whenever at least one

unit remains. UCB Rank-and-Activate first reserves one unit for each activity in that order while the budget permits. On the resulting nonempty set S, it maximizes $\sum_{a\in S} U_{t,a}\sqrt{x_a}$ subject to $1 \le x_a \le n_a$ and $\sum_{a\in S} x_a \le B$, setting other coordinates to zero. It optimizes depth after fixing the active set. If $B < 1$, it assigns expected allocation B to the first-ranked activity through randomized activation. Rank-and-Fill stops when less than one unit remains. Write $J_t^r := \mathbb{E}_r[R_t \mid \mathscr{H}_{t-1}]$, with indices URF and URA for Rank-and-Fill and Rank-and-Activate and the same regret subscripts in (3).

PROPOSITION 3 **(Reward loss from ranking-based allocation).** *For each rule* $r \in \{\mathrm{URF}, \mathrm{URA}\}$*, there is a family containing one instance for each* $K \ge 4$ *such that the following bounds hold for every history:*

$$\frac{J_t^r}{J^*} = O(K^{-1/2}), \qquad \mathcal{R}_r(T) = \Omega(T). \tag{7}$$

The reward bound holds in every period, uniformly over the horizon. For each fixed instance, the regret bound holds for all $T \ge 1$*.*

The bounds also hold in expectation. Rank-and-Fill concentrates resources despite declining marginal rewards. Rank-and-Activate can spend the budget on minima, leaving nothing to scale the most rewarding activity even after optimizing depth within its selected set. These restrictions persist as reward estimates improve.

2.4. PRICERESOLVE Oracle

PRICERESOLVE takes estimated means $\boldsymbol{\theta} \ge \mathbf{0}$, capacities $\boldsymbol{n}$, and budget B, returning an optimal $\boldsymbol{x} \in \mathcal{X}^*(\boldsymbol{\theta})$ and supporting price λ_{out}. Algorithm 1 searches breakpoints where allocation formulas change and the intervening intervals. At $\lambda > 0$, the budget is feasible within the individual optimal sets if their smallest allocations sum to at most B and their largest to at least B. Appendix B gives the breakpoints and interval roots.

At a positive returned price, start from each set's smallest allocation and distribute the remaining budget within activation-tied intervals in fixed order $\prec$. At zero price, use that order to allocate any remainder among zero-mean activities. The decreasing sweep selects the largest multiplier. Binary search over sorted breakpoints has the same complexity and may return another optimal multiplier. Every optimal allocation is supported by the largest multiplier $\lambda(\boldsymbol{\theta})$ used in the analysis.

Sorting the breakpoints determines the running time. Under real arithmetic that treats square roots as unit-cost operations, the oracle has the following guarantee.

Algorithm 1 PRICERESOLVE for allocation using estimated reward means

Require: Nonnegative estimated reward means $\boldsymbol{\theta}$, capacities $\boldsymbol{n}$, and budget $B>0$; use the fixed order $\prec$.

Set $x_a^0 \leftarrow n_a$ if $\theta_a > 0$, and $x_a^0 \leftarrow 0$ otherwise.
if $\sum_a x_a^0 < B$ **then**
    Increase coordinates with $\theta_a = 0$ in order $\prec$, up to capacity and the remaining budget; **return** $(\boldsymbol{x}^0, 0)$.
end if
Construct and sort the $O(K)$ positive price breakpoints.
Sweep prices downward; test interval roots and breakpoint clearing conditions.
Take the first price λ_{out} that clears the budget; set $x_a \leftarrow \min \mathcal{D}_a(\theta_a, \lambda_{\text{out}})$.
Distribute $B - \sum_a x_a$ within the allocation intervals tied at λ_{out}, in order $\prec$.
return $(\boldsymbol{x}, \lambda_{\text{out}})$.

PROPOSITION 4 **(Correctness and complexity of** PRICERESOLVE**).** *For* $\boldsymbol{\theta} \geq \mathbf{0}$ *and* $B > 0$, PRICERESOLVE *returns an allocation* $\boldsymbol{x} \in \mathcal{X}^*(\boldsymbol{\theta})$ *and a price* λ *satisfying the scalar optimality and budget complementarity conditions, with* $\sum_a x_a = B \wedge \sum_a n_a$. *Its exact unit-cost arithmetic running time is* $O(K \log(eK))$ *and its storage is* $O(K)$. *The decreasing sweep selects* $\lambda = \lambda(\boldsymbol{\theta})$.

3. Learning Policies for Activation and Scaling

The platform learns reward means from activities it activates. Section 3.1 combines variance-aware upper confidence estimates with PRICERESOLVE in VA-UCB. Section 3.2 develops NCMD's cheaper incremental updates. Both implement expected allocations through (5).

3.1. VARIANCE-AWARE UCB (VA-UCB)

An empirical mean can undervalue an activity after a few unfavorable observations, reducing its future feedback. VA-UCB instead optimizes upper-confidence estimates of reward means. Activities compete using means compatible with their observations, and the allocation structure jointly determines activation probabilities and depth.

For activated feedback $X \in [0, r]$ with mean μ, condition all moments on the past, chosen allocation, and activation. Since $X^2 \leq rX$,

$$\operatorname{Var}(X) = \mathbb{E}[X^2] - \mu^2 \leq r\mu - \mu^2 = \mu(r-\mu) \leq r^2/4.$$

Low means permit smaller variance bounds than the range-based maximum. Retaining this dependence limits excessive optimism, which can divert resources from better activities while further observations still require minimum commitments.

Before period t, define arm a's activation count and empirical feedback mean by

$$N_{t,a} = \sum_{s<t} A_{s,a}, \qquad \widehat{\mu}_{t,a} = \frac{\sum_{s<t} A_{s,a} X_{s,a}}{N_{t,a} \vee 1}.$$

The logarithmic factor L_c protects all arms at the adaptively reached observation counts through the horizon. Fix its confidence coefficient $b>0$ before learning. Appendix C.1 specifies L_c and the closed-form endpoint. For sample mean m, count $N>0$, and $l=L_c/N$, Appendix Lemma EC.3 gives the largest compatible candidate mean:

$$F(m,l,r)=\max\left\{u\in[0,r]:u-m\le\sqrt{\tfrac{9}{4}u(r-u)l}+ul/3\right\}.$$

The candidate variance bound $u(r-u)$ can narrow the range of plausible means when feedback is low. The true mean remains admissible on the confidence event, preserving optimism.

In the canonical model, both feedback and means are bounded by one. Use this public bound before any observation, retaining an activity's eligibility, and the empirical endpoint afterward:

$$U_{t,a}=\begin{cases}1, & N_{t,a}=0,\\ \min\{1,F(\widehat{\mu}_{t,a},L_c/N_{t,a},1)\}, & N_{t,a}>0.\end{cases}\tag{8}$$

With few observations, all means in $[0,1]$ may remain admissible. Each activated observation, including zero, increases the next decision's count. PriceResolve uses these upper endpoints to jointly choose activation and depth under the expected budget.

Algorithm 2 VA-UCB

Require: $T,B,\boldsymbol{n},b>0$; fixed arm order $\prec$.

1: Initialize $N_a=S_a=0$ for every arm and compute L_c.
2: **for** $t=1,\dots,T$ **do**
3: Form $U_{t,a}$ by (8), with $\widehat{\mu}_{t,a}=S_a/(N_a\vee 1)$.
4: Set $(\boldsymbol{x}_t,\lambda_{\mathrm{out},t})\leftarrow\textsc{PriceResolve}(\boldsymbol{U}_t;\boldsymbol{n},B)$, resolving ties in order $\prec$.
5: Implement $\boldsymbol{x}_t$ by (5) and observe activated arms.
6: For each activated arm set $N_a\leftarrow N_a+1$, $S_a\leftarrow S_a+X_{t,a}$.
7: **end for**

The oracle applies Proposition 2 at $\boldsymbol{\theta}=\boldsymbol{U}_t$, including ties and zero price. Write $\varepsilon_{t,a}=U_{t,a}-\mu_a$ and $p_{t,a}=x_{t,a}\wedge 1$ for estimation error and activation probability. Optimism and exact optimization give the following loss control before the optimal active set is identified.

Proposition 5 **(Loss control for** VA-UCB**).** *In any period of Algorithm 2 with $U_{t,a}\ge\mu_a$ for every a,*

$$J^*-\sum_a\mu_a h_a(x_{t,a})\le\sum_a\varepsilon_{t,a}h_a(x_{t,a})\le\sqrt{B\sum_a p_{t,a}\varepsilon_{t,a}^2}.$$

Squared errors are weighted by the activation probabilities governing observation counts. Section 4.1 further separates mistaken activations from depth errors through threshold separation.

The policy stores one activation count and one feedback sum per arm. The oracle determines the per-period computation cost of $O(K\log(eK))$ unit-cost operations.

3.2. Noise-Controlled Mirror Descent (NCMD)

NCMD incrementally updates allocations using estimated marginal rewards, with the same worst-case expected regret order as VA-UCB for $B \le K$ up to logarithms (Section 4.2). Updates change observation probabilities below the minimum and depth above it. Apply mirror ascent (mirror descent on negative reward) to

$$f_{\boldsymbol{\mu}}(\boldsymbol{x}) := \sum_{a=1}^{K} \mu_a h_a(x_a).$$

Since every h_a is nondecreasing and $B < \sum_a n_a$, a full-budget optimizer exists. Restrict updates to

$$\mathcal{X}_B = \left\{ \boldsymbol{x} \in \mathbb{R}_+^K : 0 \le x_a \le n_a,\ \sum_a x_a = B \right\}.$$

Negative entropy gives multiplicative resource changes and weights noise by current allocations, controlling noise when activation is rare. Positive allocations preserve observation opportunities. Its Bregman divergence defines projection onto $\mathcal{X}_B$. For $\boldsymbol{u} \ge \mathbf{0}$ and $\boldsymbol{x} > \mathbf{0}$, define

$$\Phi(\boldsymbol{x}) = \sum_{a=1}^{K} x_a(\log x_a - 1), \qquad D_\Phi(\boldsymbol{u}, \boldsymbol{x}) = \sum_{a=1}^{K} \left[u_a \log \frac{u_a}{x_a} - u_a + x_a \right].$$

The regularizer uses $0 \log 0 := 0$. In the divergence, the continuous extension applies to $u_a = 0$ while $x_a > 0$. Initialize at the strictly positive negative-entropy minimizer

$$\boldsymbol{x}_1 \in \arg\min_{\boldsymbol{x} \in \mathcal{X}_B} \Phi(\boldsymbol{x}).$$

The estimator combines the known response slope with the observation probability. For $x > 0$, let $d_a(x)$ be the selected envelope supergradient and $p_a(x)$ the activation probability:

$$d_a(x) = \begin{cases} 1, & 0 < x \le 1, \\ (2\sqrt{x})^{-1}, & 1 < x \le n_a, \end{cases} \qquad p_a(x) = x \wedge 1.$$

A supergradient is a slope whose line through the current response lies above the response curve. At $x = 1$, we use the left slope $d_a(1) = 1$.

Inverse-probability weighting corrects observation frequency but can amplify rare feedback. Use the residual from a historical reference $\beta_{t,a}$. With $p_{t,a} = p_a(x_{t,a})$, form

$$\widehat{g}_{t,a} = d_a(x_{t,a}) \left[\beta_{t,a} + \frac{A_{t,a}(X_{t,a} - \beta_{t,a})}{p_{t,a}} \right]. \tag{9}$$

For an inactive activity, the estimate is $d_a(x_{t,a})\beta_{t,a}$. Conditional on $\mathscr{H}_{t-1}$, the reference is fixed and the inverse-probability residual has mean $\mu_a - \beta_{t,a}$. Adding the reference and multiplying by the slope gives mean $\mu_a d_a(x_{t,a})$.

At a fixed history $\mathscr{H}_{t-1}$, write $Z = \widehat{g}_{t,a}/d_a(x_{t,a})$, $p = p_{t,a}$, and $\beta = \beta_{t,a}$. Let σ^2 be feedback variance conditional on that history and activation. Then

$$\operatorname{Var}(Z \mid \mathscr{H}_{t-1}) = \frac{\sigma^2}{p} + \frac{1-p}{p}(\beta - \mu_a)^2.$$

The first term is feedback variability when active. The second is activation noise, which decreases as the reference approaches the mean and vanishes there. For $0 < x_{t,a} \le 1$, $p = x_{t,a}$ and $d_a(x_{t,a}) = 1$, so

$$x_{t,a} \operatorname{Var}(\widehat{g}_{t,a} \mid \mathscr{H}_{t-1}) = \sigma^2 + (1 - x_{t,a})(\beta - \mu_a)^2.$$

In the entropy analysis, the resource weight $x_{t,a}$ therefore cancels the inverse-activation factor in this noise contribution.

The reference uses the historical empirical mean, with a floor controlling large positive updates from rare high observations. Write $S_{t,a} = \sum_{v<t} A_{v,a} X_{v,a}$ for the observation sum and use the count $N_{t,a}$ above. At step size η_t, raise the reference as needed so that

$$\eta_t \widehat{g}_{t,a} \le 1.$$

The floor is the smallest nonnegative reference ensuring this inequality for every observation in $[0,1]$ at the current step and activation probability. It reduces the positive residual correction without changing activation probabilities or excluding negative estimates, preserving unbiasedness.

Decreasing steps balance initial divergence against accumulated noise. The horizon T and coefficient $c > 0$ are fixed before learning. Appendix D gives the exact schedule and reference, including the canonical specialization used here.

Coordinate updates may exceed the budget or capacities. Project them onto $\mathcal{X}_B$ using entropy:

$$w_{t+1,a} = x_{t,a} \exp(\eta_t \widehat{g}_{t,a}), \qquad \boldsymbol{x}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}_B} D_\Phi(\boldsymbol{x}, \boldsymbol{w}_{t+1}).$$

The projection uses a common positive scaling factor z_t:

$$x_{t+1,a} = \min\{n_a, z_t w_{t+1,a}\}, \qquad \sum_a \min\{n_a, z_t w_{t+1,a}\} = B. \tag{10}$$

For $\boldsymbol{w} \in \mathbb{R}^K_{++}$ and $0 < B < \sum_a n_a$, the budget equation uniquely determines $z > 0$. CappedScale$(\boldsymbol{w}; B, \boldsymbol{n})$ returns the resulting unique allocation using deterministic median selection, fixed-order ties, and $O(K)$ operations (Appendix D).

The projection conditions express budget rescaling through a common threshold for estimated marginal rewards, with separate multipliers for binding caps.

Algorithm 3 NCMD

Require: $T, B, \boldsymbol{n}, c > 0$.

Initialize observation counts and sums $N_a = S_a = 0$.
Set $\boldsymbol{x}_1 \leftarrow \arg\min_{\boldsymbol{x} \in \mathcal{X}_B} \Phi(\boldsymbol{x})$ by CappedScale applied to the all-one vector.
for $t = 1, \ldots, T$ **do**
    Set the prescribed step size η_t, compute $p_{t,a}$ from $x_{t,a}$, and form $\beta_{t,a}$ using the previous counts and sums and the canonical construction in Appendix D.
    Implement $\boldsymbol{x}_t$ by (5) and observe the activated coordinates.
    Form $\widehat{\boldsymbol{g}}_t$ by (9).
    Set $w_{t+1,a} \leftarrow x_{t,a} \exp(\eta_t \widehat{g}_{t,a})$ for every a.
    Set $\boldsymbol{x}_{t+1} \leftarrow \textsc{CappedScale}(\boldsymbol{w}_{t+1}; B, \boldsymbol{n})$ using (10).
    For each activated arm set $N_a \leftarrow N_a + 1$, $S_a \leftarrow S_a + X_{t,a}$.
end for

Proposition 6 **(Projection conditions for** NCMD**).** *In every* NCMD *period, there are a common scalar* $\lambda_t^{\mathrm{M}} \in \mathbb{R}$ *and cap multipliers* $\xi_{t,a} \geq 0$ *such that*

$$\log \frac{x_{t+1,a}}{x_{t,a}} = \eta_t \big(\widehat{g}_{t,a} - \lambda_t^{\mathrm{M}} - \xi_{t,a}\big), \qquad \xi_{t,a}(n_a - x_{t+1,a}) = 0, \qquad \lambda_t^{\mathrm{M}} = -\frac{\log z_t}{\eta_t}.$$

An uncapped allocation grows, contracts, or stays fixed as its estimated marginal reward exceeds, falls below, or equals λ_t^{M}. This threshold enforces the budget and may have either sign. PriceResolve's optimal dual price instead measures the marginal value of budget for given reward means. The next result gives feasibility, unbiasedness, and per-period cost.

Proposition 7 **(Feasibility, unbiasedness, and linear-time updates).** *For every period of Algorithm 3:*

(i) $\boldsymbol{x}_t$ *is determined by* $\mathscr{H}_{t-1}$*, is strictly positive, and belongs to* $\mathcal{X}_B$*. Its randomized implementation satisfies* (1) *with equality.*

(ii) Conditional on $\mathscr{H}_{t-1}$*, the marginal estimate is unbiased:*

$$\mathbb{E}[\widehat{g}_{t,a} \mid \mathscr{H}_{t-1}] = \mu_a d_a(x_{t,a}).$$

(iii) Under real arithmetic that treats exponentiation as a unit-cost operation, CappedScale *and the complete period update use* $O(K)$ *operations and* $O(K)$ *memory.*

4. Regret Guarantees and Policy Comparison

Section 4.1 gives instance-dependent and uniform VA-UCB bounds, and Section 4.2 gives NCMD's expected bound. Section 4.3 establishes a global lower bound and near-minimax comparison. Section 4.4 isolates the learning cost near an activation boundary.

Use regret (3) over $\Pi_K(B,\boldsymbol{n})$, with subscripts U and M for Algorithms 2 and 3. For fixed $B,\boldsymbol{n}$, let $\mathfrak{I}_K(B,\boldsymbol{n})$ contain the instances of Section 2.1 with activation observations in $[0,1]$ and fixed conditional means μ_a. Let $\mathfrak{I}_K^+(B,\boldsymbol{n})$ be its positive-price subclass. The lower bounds use independent Bernoulli instances in that subclass. VA-UCB uses known horizon T, fixed $b>0$, $\alpha_T = b/(T+b)$, exact PriceResolve calls, and deterministic ties. Write $L_T = \log(eKT/\alpha_T)$. NCMD uses the same conditional-mean model and the canonical decreasing schedule in Appendix D.

4.1. Regret of VA-UCB

Relative to the optimal activation threshold, define inactive activities $\mathcal{I}$, tied activities $\mathcal{G}$, and their count G:

$$\mathcal{I} = \{a : \mu_a < \lambda^*\}, \qquad \mathcal{G} = \{a : \mu_a = \lambda^*\}, \qquad G = |\mathcal{G}|.$$

The positive instance-dependent coefficient α_G relates reward loss to squared deviation of tied activities' aggregate activation allocation from its optimum. Appendix C.4 defines it and the convention for a vanishing tie term. The analysis permits arbitrarily small allocation reward gaps and activation probabilities.

Theorem 1 **(Instance-dependent regret bound for VA-UCB).** *Under the standing* VA-UCB *conditions above and $\lambda^* > 0$, the following bound holds with probability at least $1-\alpha_T$:*

$$\mathcal{R}_{\mathsf{U}}(T) \le C\left(L_T \sum_{a\in\mathcal{I}} \frac{1}{\lambda^* - \mu_a} + \frac{KL_T^2}{\lambda^*} + \alpha_G^{-1/3} G^{2/3} T^{1/3} L_T^{4/3}\right), \tag{11}$$

where C is universal and an empty sum is zero.

For a fixed positive-price instance and fixed b, mistaken-activation loss in (11) is $O(\log T)$. Smaller gaps $\lambda^* - \mu_a$ require more observations to identify activities the known-mean allocation leaves inactive. Scaling loss is $O(\log^2 T)$: curvature reduces it to squared confidence errors, and allocations above the minimum yield observations with certainty. Their distinct counts contribute a harmonic sum. One logarithm comes from squared widths and the other from this sum.

At an activation tie, $\mu_a = \lambda^*$, so the unit commitment's expected reward equals its resource opportunity cost. Even one activity can form a tie. When nonzero, its aggregate term is $O(T^{1/3}\log^{4/3} T) = \widetilde{O}(T^{1/3})$. Redistributing a fixed total among tied activation layers preserves reward, but changing that total diverts resources from other activities. If all means are strictly above or below the price, the term vanishes and the bound is polylogarithmic in T.

We next bound regret uniformly over the instance class, covering activities arbitrarily close to activation thresholds and instances with zero optimal price.

THEOREM 2 **(Uniform regret bound for** VA-UCB**).** *There is a universal constant C such that the following bound holds for every $\mathsf{P} \in \mathfrak{I}_K(B, \boldsymbol{n})$ with probability at least $1 - \alpha_T$ under P:*

$$\mathcal{R}_{\mathsf{U}}(T) \le C\sqrt{KBT}L_T.$$

Its expectation is at most

$$C\sqrt{KBT}L_T + b\min\{B, \sqrt{KB}\}.$$

For fixed b, this is $\widetilde{O}(\sqrt{KBT})$, uniformly over the instance class. The expectation of the applicable instance-dependent bound adds the same failure-event term.

On the joint confidence and counting event, $\sum_{t,a} p_{t,a}\varepsilon_{t,a}^2 \le CKL_T^2$. Summing Proposition 5 and applying Cauchy–Schwarz across periods gives $\mathcal{R}_{\mathsf{U}}(T) \le \sqrt{BT\sum_{t,a} p_{t,a}\varepsilon_{t,a}^2}$, yielding the uniform order. The same activation probabilities weight learning loss and govern observation counts. Instance-dependent bounds can be sharper when activation decisions are easier to resolve.

4.2. Regret of NCMD

NCMD's incremental updates retain VA-UCB's polynomial order of worst-case expected regret for $B \le K$ against the same benchmark. Initial divergence is at most $D_B = B\log(eK)$. The positive constants A_c, C_c depend only on fixed c and are defined in Appendix D.3. For the prescribed schedule, that appendix gives $\mathbb{E}\mathcal{R}_{\mathsf{M}}(T) \le A_c D_B/\eta_T + (9K/4)\sum_{t=1}^{T}\eta_t$. Small steps increase the reciprocal-step term, while large steps increase accumulated noise. Balancing them yields square-root dependence on $D_B KT$, with an additional term proportional to D_B from the step cap.

THEOREM 3 **(Uniform regret bound for** NCMD**).** *For fixed $c > 0$, Algorithm 3 with the canonical decreasing schedule specified in Appendix D satisfies*

$$\sup_{\mathsf{P}\in\mathfrak{I}_K(B,\boldsymbol{n})} \mathbb{E}_{\mathsf{P}}\mathcal{R}_{\mathsf{M}}(T) \le A_c D_B + \frac{3}{2}C_c\sqrt{D_B KT}.$$

When $B \le K$, it also satisfies

$$\sup_{\mathsf{P}\in\mathfrak{I}_K(B,\boldsymbol{n})} \mathbb{E}_{\mathsf{P}}\mathcal{R}_{\mathsf{M}}(T) \le (\sqrt{A_c} + \tfrac{3}{2}C_c)\sqrt{BKT\log(eK)}.$$

For fixed K and $B \le K$, average expected reward loss decreases at rate $T^{-1/2}$ across the instance class, including optimally inactive activities. Under the stated unit-cost model, NCMD reduces per-period work from $O(K\log(eK))$ to $O(K)$ with the same polynomial regret order. The logarithmic factors differ, and VA-UCB also has the instance-dependent guarantee in Theorem 1.

4.3. Global Lower Bound and Near-Minimax Comparison

To compare full-class upper bounds with a positive-price Bernoulli lower bound, define minimax regret by

$$\mathfrak{R}_T^*(K,B,\boldsymbol{n}) := \inf_{\pi\in\Pi_K(B,\boldsymbol{n})} \sup_{\mathsf{P}\in\mathfrak{I}_K(B,\boldsymbol{n})} \mathbb{E}_{\mathsf{P}}[\mathcal{R}_\pi(T)].$$

The following lower bound shows that learning which activities to activate already imposes a square-root cost on every admissible policy.

THEOREM 4 **(Scarce-budget minimax lower bound)**. *Suppose $B=m$ is an integer, $1\le m\le K/2$, and $T\ge K/B$. For every finite capacity vector,*

$$\inf_{\pi\in\Pi_K(B,\boldsymbol{n})} \sup_{\mathsf{P}\in\mathfrak{I}_K^+(B,\boldsymbol{n})} \mathbb{E}_{\mathsf{P}}[\mathcal{R}_\pi(T)] \ge c\sqrt{KBT},$$

where $c>0$ is universal.

Similar means make valuable activations difficult to identify, and minimum commitments limit observations. On this family, replacing depth by minimum-scale activations preserves expected allocation while weakly increasing reward and feedback. The activation-learning cost therefore persists across depth choices. Combining this subclass lower bound with NCMD's full-class upper bound gives the following comparison.

COROLLARY 1 **(Near-minimax guarantee)**. *Fix the* NCMD *coefficient $c=1$. Let $B=m\in\{1,\dots,\lfloor K/2\rfloor\}$ and $T\ge K/B$. There are universal constants $c_0,C>0$ such that*

$$c_0\sqrt{KBT} \le \mathfrak{R}_T^*(K,B,\boldsymbol{n}) \le \sup_{\mathsf{P}\in\mathfrak{I}_K(B,\boldsymbol{n})} \mathbb{E}_{\mathsf{P}}[\mathcal{R}_{\mathsf{M}}(T)] \le C\sqrt{KBT\log(eK)}.$$

On the stated integer scarce-budget range, NCMD is near-minimax within $\sqrt{\log(eK)}$, and Theorem 2 gives a near-minimax expected rate for VA-UCB up to logarithmic factors.

4.4. Local Activation-Boundary Difficulty

To examine activation-tie difficulty, consider independent Bernoulli feedback with $B=4$, capacities $(n_1,n_2)=(4,1)$, and $\mu_1=1$. The second mean is $1/4$ in ν_0 and $1/4+\Delta$ in ν_Δ. Instance ν_0 allocates everything to arm 1, while a small increase in arm 2's mean makes a strictly fractional activation probability optimal. The following result measures the cost of distinguishing these decisions.

THEOREM 5 **(Local activation-boundary lower bound)**. *There exist universal constants $c,\eta_0>0$ and $T_0\ge 1$ such that, for every $T\ge T_0$, the two instances above with $\Delta=\eta_0 T^{-1/3}$ satisfy*

$$\inf_{\pi\in\Pi_2(4,(4,1))} \max_{\nu\in\{\nu_0,\nu_\Delta\}} \mathbb{E}_\nu[\mathcal{R}_\pi(T)] \ge cT^{1/3}.$$

As Δ decreases with T, the instances become harder to distinguish. In ν_0, a small arm-2 activation probability reduces resources for arm 1 and causes loss bounded below by a positive constant times its square. Observations arrive in proportion to that probability, balancing experimentation cost against activating a valuable activity too rarely. In both instances, arm 2's mean equals the endogenous optimal price. This price and α_G stay uniformly positive. With common fixed b, Proposition EC.1 makes Theorem 1's expected bound uniform over the pair and horizon: $C[T^{1/3}L_T^{4/3} + L_T^2] + b\sqrt{8}$. It matches the lower bound up to logarithms on the horizon-dependent pair.

5. Model Extensions and Operating Implications

Section 5.1 extends the unit-minimum, square-root model to heterogeneous commitments and power rewards, including independent activation-only and linear endpoints. Section 5.2 examines estimation errors in interior allocations and the protection afforded by boundary regimes.

5.1. General Minimum Scales and Power Rewards

Activity-specific minima change observation costs, and the power exponent changes returns from additional investment. Let $0 < \ell_a \le n_a < \infty$ and let known $q \in (0,1)$ be common to all activities. Feedback lies in $[0, r_a]$, with known support bound r_a and public mean bound $0 \le \mu_a \le M_a \le r_a$. Conditional means and decision timing are unchanged, and $M_a = r_a$ is always valid. The factor $s > 0$ converts normalized into physical reward: active allocation y earns expected reward $s\mu_a y^q$. For $0 < B < \sum_a n_a$, the exact envelope and implementation are

$$h_a(x) = \begin{cases} \ell_a^{q-1} x, & 0 \le x \le \ell_a, \\ x^q, & \ell_a < x \le n_a, \end{cases} \qquad p_a(x) = \min\{1, x/\ell_a\}, \quad Y_a = A_a \max\{x_a, \ell_a\}. \tag{12}$$

The normalized benchmark is $J^* = \max_{x \in \mathcal{X}} \sum_a \mu_a h_a(x_a)$. Under this implementation, the regret of either policy is $\mathcal{R}^{(q)}(T) = s \sum_t [J^* - \sum_a \mu_a h_a(x_{t,a})]$. Prices below refer to normalized reward. At the largest optimal price $\lambda_q^* > 0$, activation and scaling thresholds are

$$\tau_a^0 = \lambda_q^* \ell_a^{1-q}, \qquad \tau_a^H = (\lambda_q^*/q)\ell_a^{1-q}. \tag{13}$$

An activity with $\tau_a^0 < \mu_a < \tau_a^H$ operates at its minimum. The scaling threshold is $1/q$ times the activation threshold.

Generalized VA-UCB uses $U_{t,a} = M_a$ at zero count and otherwise $U_{t,a} = \min\{M_a, F(\widehat{\mu}_{t,a}, L_c/N_{t,a}, r_a)\}$, then calls the generalized oracle. Clipping can place this endpoint below a sample mean exceeding M_a. Generalized NCMD uses slope ℓ_a^{q-1} below and at the minimum and qx^{q-1} above it, with the same residual estimate and multiplicative update. It clips the historical mean at M_a before applying the floor. The final reference lies in $[0, r_a]$ and may exceed M_a. Appendix D gives its reference and decreasing step.

The bounds account for support-weighted slopes and noise. $M_r = \bar{d}_r = \max_a r_a \ell_a^{q-1}$ is the largest support-weighted slope at a minimum. D_r bounds the squared increase in support-weighted response per unit of above-minimum allocation, and V bounds the allocation-weighted second moment of marginal estimates. Appendices C.6 and D give their precise definitions.

THEOREM 6 **(Regret bounds for the generalized model).** *For known T and fixed $b, c > 0$, generalized* VA-UCB *with exact oracle calls satisfies, on every instance including zero-price instances, with probability at least $1 - \alpha_T$,*

$$\mathcal{R}_{\mathsf{U}}^{(q)}(T) \le Cs \left[M_r \sqrt{BTL_T \sum_a \ell_a} + \sqrt{D_r K B T L_T \log(eT)} \right].$$

Its expected bound adds at most bsH_r, where $H_r = \max_{x \in \mathcal{X}} \sum_a r_a h_a(x_a)$. Generalized NCMD *with the reference and decreasing step specified in Appendix D satisfies*

$$\mathbb{E}\mathcal{R}_{\mathsf{M}}^{(q)}(T) \le s[A_c D_B \bar{d}_r + C_c \sqrt{D_B V T}].$$

If every $r_a = 0$, use a fixed feasible action to obtain zero regret. At $B = 0$ or $B = \sum_a n_a$, use the corresponding fixed zero or capacity allocation, which also has zero regret.

Observation costs and the reward consequences of allocation errors can differ at the same portfolio size. Unit minima, square-root rewards, $r_a = 1$, and $s = 1$ recover the canonical model. Its VA-UCB and NCMD bounds follow from Appendix C.6's canonical refinement and Appendix D's specialization. Appendix C.5 gives the general instance-dependent bound.

We treat the activation-only and linear endpoints as separate models with the same activation and feedback rules. At $q = 0$, expected reward is constant beyond the minimum, so the benchmark chooses activation probabilities.

PROPOSITION 8 **(Optimal allocation at the activation endpoint).** *At $q = 0$, let y_a be activity a's activation probability. The normalized benchmark is*

$$J_0^* = \max \left\{ \sum_a \mu_a y_a : 0 \le y_a \le 1,\ \sum_a \ell_a y_a \le B \right\}.$$

There is an extreme optimizer $\boldsymbol{y}^*$ with

$$\#\{a : 0 < y_a^* < 1\} \le 1, \qquad \frac{\mu_i}{\ell_i} > \frac{\mu_j}{\ell_j} \implies y_i^* = 1 \text{ or } y_j^* = 0.$$

This fractional-knapsack oracle activates activities in decreasing reward per unit of minimum commitment, with at most one randomized activation in an extreme optimum. At $q = 1$, additional allocation has constant marginal reward, so the benchmark sorts reward means.

PROPOSITION 9 **(Optimal allocation at the linear endpoint).** *At $q=1$, the normalized benchmark is*

$$J_1^* = \max\left\{\sum_a \mu_a x_a : 0 \le x_a \le n_a,\ \sum_a x_a \le B\right\}.$$

There is an extreme optimizer $\boldsymbol{x}^*$ with

$$\#\{a : 0 < x_a^* < n_a\} \le 1, \qquad \mu_i > \mu_j \implies x_i^* = n_i \text{ or } x_j^* = 0.$$

The linear oracle fills capacities in decreasing reward-mean order, with at most one partially filled activity in an extreme optimum. Ranking therefore suffices at these endpoints, where reward is constant beyond the minimum or marginal reward is constant. Both piecewise-linear envelopes retain the PRICERESOLVE interface. Theorem EC.6 analyzes the two learning policies directly with endpoint actions and slopes. These are independent results because Theorem 6 assumes $0 < q < 1$.

5.2. Estimation Accuracy and Allocation Robustness

To examine which allocations respond to estimation error, fix $q \in (0,1)$ and common $(B, \boldsymbol{\ell}, \boldsymbol{n}, q)$. Let $\boldsymbol{x}^*$ and $\widehat{\boldsymbol{x}}$ optimize positive means $\boldsymbol{\mu}$ and positive estimates $\widehat{\boldsymbol{\mu}}$. An interior activity equates marginal reward to the common price:

$$q\mu_i (x_i^*)^{q-1} = \lambda_q^*, \qquad \ell_i < x_i^* < n_i.$$

Taking ratios between two interior activities cancels the price in each problem. Comparing those ratios gives the following relation.

PROPOSITION 10 **(Relative estimation errors and allocation ratios).** *If activities a, b satisfy $\ell_i < x_i^* < n_i$ and $\ell_i < \widehat{x}_i < n_i$ for $i = a, b$, then*

$$\frac{\widehat{x}_a/\widehat{x}_b}{x_a^*/x_b^*} = \left(\frac{\widehat{\mu}_a/\widehat{\mu}_b}{\mu_a/\mu_b}\right)^{1/(1-q)}. \tag{14}$$

A common proportional error leaves the optimal allocation set unchanged. For activities remaining interior, unequal proportional errors change relative allocations by the log-scale factor $1/(1-q)$. Accuracy therefore matters more with slowly diminishing returns. Across q, hold minimum reward $\mu_a(q)\ell_a^q$ fixed to isolate curvature. Proportional mean error then equals proportional minimum-reward error. A 5% overestimate of μ_a/μ_b raises the allocation ratio by approximately 10.3% at $q = 1/2$ and 27.6% at $q = 0.8$, provided both activities remain interior.

Minimum or capacity allocations can remain unchanged while estimates stay on the same side of their operating thresholds. To account for the endogenous price $\lambda_q^* > 0$, augment τ_a^0, τ_a^H in (13) by $\tau_a^C = (\lambda_q^*/q) n_a^{1-q}$. Set $\mathcal{T}_a = \{\tau_a^0, \tau_a^H, \tau_a^C\}$ if $n_a > \ell_a$ and $\mathcal{T}_a = \{\tau_a^0\}$ if $n_a = \ell_a$, when only activation is chosen. The smallest log distance to these thresholds is

$$\gamma_{\text{reg}} = \min_{a \in [K]} \min_{\tau \in \mathcal{T}_a} \left|\log \frac{\mu_a}{\tau}\right|.$$

PROPOSITION 11 **(Allocation stability away from thresholds)**. *Suppose the relative estimation error satisfies* $\max_a |\log(\widehat{\mu}_a/\mu_a)| \le \varepsilon$ *for some* $0 \le \varepsilon < \gamma_{\mathrm{reg}}/2$. *Then every activity retains its strict operating regime under the estimated reward means. In particular,* $\widehat{x}_a = x_a^*$ *whenever* $x_a^* \in \{0, \ell_a, n_a\}$.

Each log mean changes by at most ε. Price monotonicity and homogeneity give the same bound for the log price, so each log mean-to-threshold ratio changes by at most $2\varepsilon < \gamma_{\mathrm{reg}}$. Only interior allocations can change. The minimum-only interval's log width $\log(\tau_a^H/\tau_a^0) = \log(1/q)$ shrinks as q increases. Slowly diminishing returns thus amplify interior errors and narrow the relative reward interval supporting minimum-only operation.

Budget determines the resource exposed to remaining errors. On a budget interval where true and estimated allocations share fixed strict regime sets, let $\mathcal{M}, \mathcal{H}, \mathcal{C}$ denote minimum-only, interior, and capped activities, assigning active $n_a = \ell_a$ activities to $\mathcal{C}$. If $\mathcal{H} \neq \varnothing$, its resource is $S = B - \sum_{a\in\mathcal{M}} \ell_a - \sum_{a\in\mathcal{C}} n_a$. For fixed true and estimated means,

$$\widehat{x}_a - x_a^* = S\left(\frac{\widehat{\mu}_a^{1/(1-q)}}{\sum_{i\in\mathcal{H}} \widehat{\mu}_i^{1/(1-q)}} - \frac{\mu_a^{1/(1-q)}}{\sum_{i\in\mathcal{H}} \mu_i^{1/(1-q)}}\right), \qquad a \in \mathcal{H}. \tag{15}$$

The magnitude of each nonzero allocation error grows in proportion to S, while the error in the activity's share of this interior budget is unchanged. A budget change that crosses a threshold requires reassessing the regimes and their margins.

6. Numerical Study with Real-Data Calibration

We compare seven learning policies and the computation costs of VA-UCB and NCMD for three Alipay-calibrated users at three budgets. A synthetic experiment varies activity count. Section 6.1 describes calibration, Section 6.2 defines the policies, Section 6.3 specifies the design, and Section 6.4 presents results.

6.1. Real-Data Calibration

Alipay's incentive-advertising platform selects tasks and cash offers at user visits. A user may accept a task and receives payment after completion. Activity a combines business entry and task type, so the same task type in different entries defines distinct activities. For a given user, the model allocates among K candidate activities. Offer $Y_{t,a} = 0$ leaves an activity inactive, and positive offers lie in $[\ell_a, n_a]$. Budget B limits the conditional expected total offer per period. Task selection determines activation and offers determine scale, both before current feedback.

Each historical visit records available activities, eligible intervals, and completion predictions at candidate offers, conditional on task acceptance. An activity's offer–probability pairs within a visit form one response curve, and repeated visits provide multiple curves. We retain offers up to 10

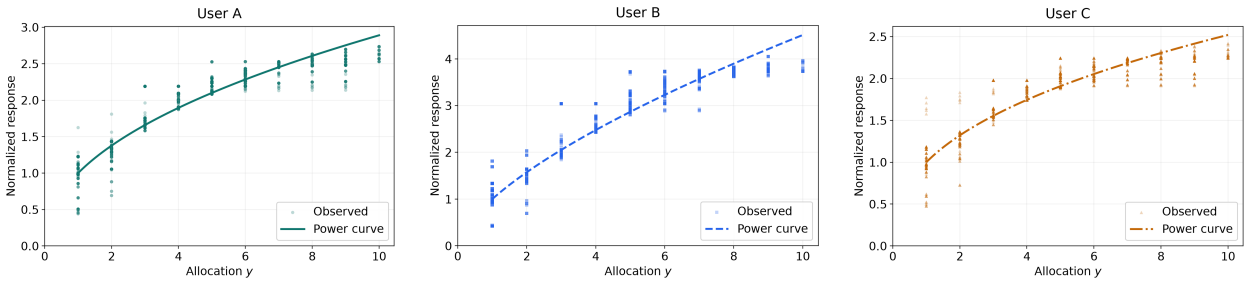


Figure 1 Normalized completion predictions and the fitted power response for each user.

incentive units, cap upper bounds at 10, and intersect eligible intervals. Their largest lower and smallest upper bounds define ℓ_a, n_a, keeping simulated positive offers eligible for every retained curve. We fit $P_{c,a}(y) = k_{c,a} y^q$ with one exponent per user and curve-specific coefficients. Log-linear regression with curve-specific intercepts estimates the common exponent.

We select cases jointly by repeated visits, activity coverage, and fitted offer sensitivity. Of 245,805 users, 722 have at least 50 historical visits. The same filtering and fitting rules yield a 5th–95th percentile range of 0.361–0.674 for q. Users A, B, and C have 86, 68, and 51 visits and nine retained activities each, permitting repeated fitting and varied activation and scaling combinations. Their exponents 0.461, 0.654, and 0.401 lie near the 44th, 92nd, and 18th percentiles, representing moderate, higher, and lower sensitivity of conditional completion predictions to offers in this cohort. Figure 1 uses linear axes: offer y horizontally and recorded prediction divided by $k_{c,a}$ vertically. Observed thus denotes normalized responses and can exceed one even when the recorded probability is valid. To obtain valid simulation probabilities from fits that can exceed one near capacity, we apply each activity's largest uniform factor at most one that keeps every curve at most one throughout $[\ell_a, n_a]$. This preserves the power shape and relative coefficients. Their equally weighted adjusted average $\bar{k}_a$ gives completion probability $\bar{k}_a y^q$.

We simulate completion conditional on acceptance. At active offer y, take a fresh draw $C \sim \text{Bernoulli}(\bar{k}_a y^q)$ and return $X = C/(s y^q)$, where $s = \max_a \ell_a^{-q}$ and $\mu_a = \bar{k}_a / s$. Its conditional mean is μ_a and $0 \le X \le 1/(s\ell_a^q) \le 1$, satisfying the bounded-feedback and fixed-mean assumptions even as offers change its law. Only active activities reveal feedback. Valid completion probability at capacity gives public bound $M_a = 1/(s n_a^q)$. Appendix H.1 gives details.

6.2. Baselines

The five baselines compare ways of learning and allocating. Two ranking rules restrict selection or filling and form Hoeffding bounds over $[0, 1]$ from their own observations. Three others use PriceResolve with point estimates, initial exploration followed by commitment, or continuing random exploration.

- **UCB Rank-and-Activate.** Rank by confidence bound times ℓ_a^{q-1}, the optimistic reward per unit of minimum commitment, and reserve each minimum that fits. Maximize reward within the selected set using those bounds and the budget and activity constraints. This generalizes Proposition 3's rule to heterogeneous minima and power rewards.
- **UCB Rank-and-Fill.** Rank confidence bounds in descending order. Assign the smaller of capacity and remaining budget when the minimum fits, skipping other activities. Leave the final remainder unallocated.
- **Empirical Mean + PriceResolve.** Solve the full allocation problem each period using raw empirical means, with zero for unseen activities.
- **Explore-Then-Commit (ETC).** For m periods, use balanced exploration $x_a = \zeta \ell_a$, where $\zeta = \min\{1, B/\sum_a \ell_a\}$, giving every activity the same activation probability at its minimum. Then call PriceResolve once with exploration-period means clipped at M_a, using zero for unseen activities, and hold the expected allocation fixed.
- **ε-Greedy.** Each period, choose that balanced exploration with fixed probability ε and otherwise use PriceResolve. Use $M_a/2$ before observation and clip empirical means at M_a. Both branches update estimates. The initialization differs from Empirical Mean + PriceResolve.

  All oracle and ranking ties follow the fixed activity order.

6.3. Experimental Design

Budgets $B = 6.25, 12.5, 18.75$ are $0.25, 0.5, 0.75$ times $\sum_a \ell_a = 25$, varying scarcity below total minimum commitment for each fixed calibration. Each period implements one allocation through (12), potentially activating several activities. The budget controls conditional expected total offers, while payment depends on completion. Policies know intervals, exponent, and public mean bounds and learn reward means through activation.

All seven policies run for $T = 100{,}000$ periods in each of nine settings. VA-UCB and NCMD use the generalized specifications in Section 5.1. Per setting, eight prespecified candidates are evaluated on the same 20 tuning seeds for each of VA-UCB's b, NCMD's c, ETC's exploration length m, and ε-Greedy's probability ε. Minimum terminal mean regret selects parameters, breaking exact ties toward smaller values. We freeze these choices before evaluating all policies on the same 80 formal seeds, disjoint from tuning. Regret counts expected completed tasks, equals s times normalized regret with $s = 1$ here, and includes exploration. Each recorded loss subtracts expected reward under the chosen allocation from the known-mean optimum. For ε-Greedy, this conditions on its current exploration or optimization branch. Averaging over that branch yields the period's conditional expected regret in (3), after reward conversion. Figure 2 averages cumulative recorded losses over all 80 seeds into 63

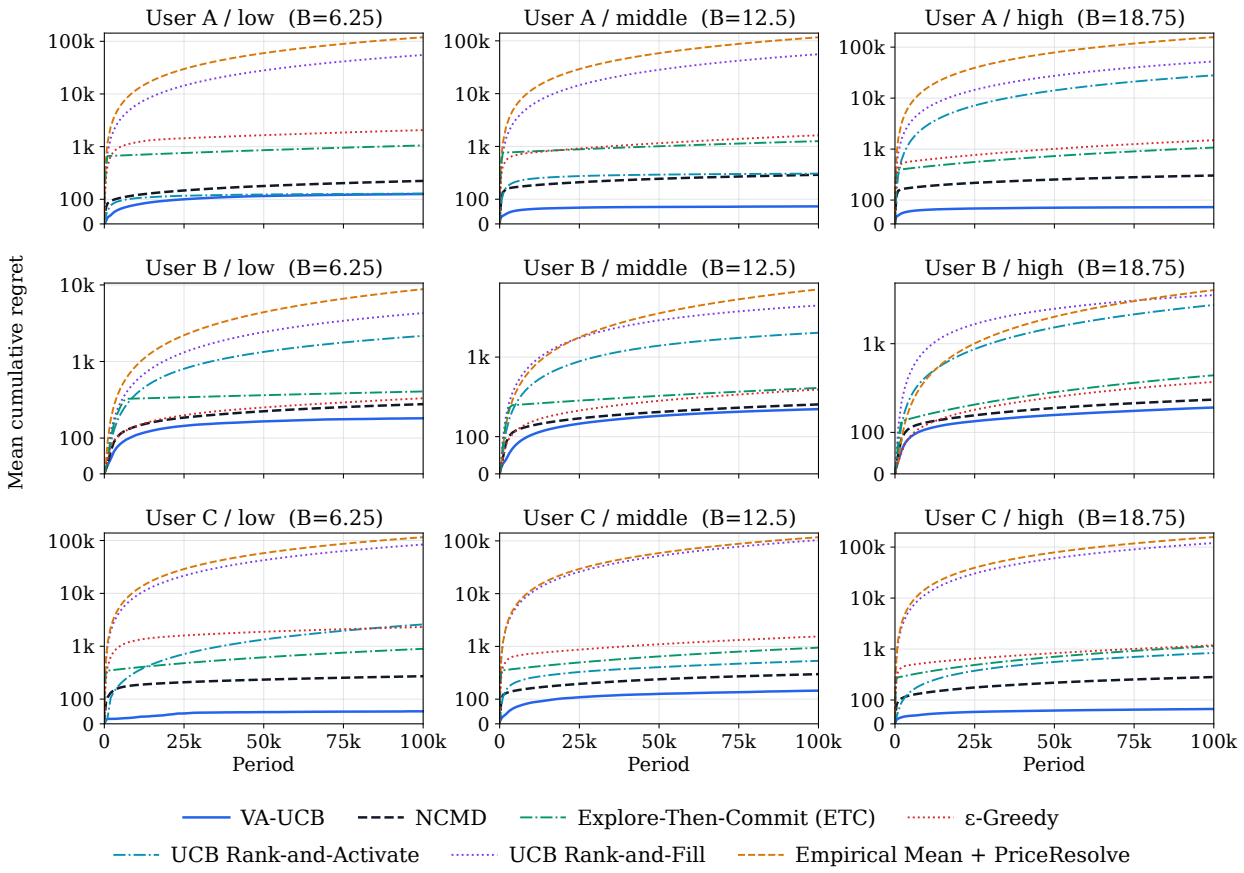


Figure 2 Mean cumulative regret of seven policies for three users and three budgets, in expected completed tasks.

mean curves as point estimates. Its vertical axis preserves the initial zero in a linear region and uses a logarithmic region for larger values.

Separate serial runs time allocation selection and state updates for the two proposed policies. Each calibrated setting has five repetitions, all covering 100,000 periods from the first decision and update. We average full-run action-plus-update totals and divide by the horizon for per-period times. Appendix H.3 gives the protocol.

The synthetic study uses $K \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$, $\ell_a = 1$, $q = 1/2$, and $B = K/2$, with capacities and means drawn to ensure valid completion probabilities (Appendix H.3). One instance per size remains fixed across ten full-horizon timing repetitions. We transfer b, c selected for User A at the middle budget to every size, recomputing VA-UCB's confidence factor for activity count and NCMD's schedule for the budget, minima, capacities, feedback bounds, and exponent. Appendix H gives implementations, grids, seeds, aggregation, and complete terminal means.

6.4. Learning and Computational Performance

Known-mean allocations combine inactivity, minima, and further investment in proportions varying by user and budget. For A at the low and middle budgets, Rank-and-Activate selects the optimal active set and attains the full optimum by optimizing within it. These settings compare learning without a known-mean active-set loss for this rule. Learning performance still depends on each policy's observations and estimates.

Figure 2 shows mean regret trajectories. Appendix H's terminal means place VA-UCB first in all nine settings and NCMD second in eight. At A's low budget, Rank-and-Activate is close to VA-UCB and below NCMD. At A's middle budget, NCMD has a small terminal advantage over Rank-and-Activate. The horizon also matters: NCMD is below VA-UCB at the first four recorded positive checkpoints there, before their order reverses. The strongest other baseline varies, including ε-Greedy for B and ETC for C at the low budget. Under its stated initialization and tie rule, Empirical Mean + PriceResolve has higher terminal mean regret than VA-UCB in all nine settings.

NCMD has lower mean per-period computation time in all nine calibrated settings: 0.610–0.741 milliseconds, compared with 0.956–1.269 for VA-UCB. The nine VA-UCB-to-NCMD time ratios average 1.744. Table 4 in the electronic companion gives full-run totals.

Table 1 shows NCMD faster at every tested synthetic size, with varying speedups generally higher at larger K.

Table 1 Mean full-run computation time in seconds on synthetic instances of increasing size.

K	VA-UCB	NCMD	Speedup factor
8	97.467	55.541	1.75×
16	101.710	66.021	1.54×
32	107.084	69.546	1.54×
64	133.824	88.425	1.51×
128	200.011	114.109	1.75×
256	300.167	158.156	1.90×
512	595.596	295.020	2.02×
1024	1343.993	664.526	2.02×

Each run measures all 100,000 periods, beginning with the first decision and update. Times are arithmetic means of ten full-run action-plus-update totals. Speedup is VA-UCB's mean divided by NCMD's mean, using unrounded values.

These finite measurements accord with its lower operation count. Across calibrated settings, VA-UCB achieves the lowest terminal mean regret, while NCMD uses less computation per decision and update. Its computation advantage also persists over the tested activity counts.

7. Conclusions and Future Work

Minimum commitments make activation a reward and information decision. The exact concave benchmark separates activation from further investment through a common optimal price, explaining why an activity can remain at its minimum. VA-UCB uses this structure for optimistic reoptimization, while NCMD reduces work through incremental updates and rescaling. Under unit minima and square-root rewards with fixed tuning, both approach the known-mean benchmark in expected average reward and are near-minimax over the stated scarce integer-budget range. The local pair identifies additional learning costs near activation thresholds. At a positive price within fixed strict regimes, small estimation errors can change interior allocations while preserving zero, minimum, and capacity allocations. The calibrated comparison offers a practical choice: VA-UCB has the lowest terminal mean regret among seven policies in all nine settings, while NCMD uses less per-period computation there and at every tested synthetic size.

The analysis uses a conditional expected budget and one observation per activation, and the comparisons condition on fixed user calibrations. One direction is to develop learning policies for general hard per-period budgets beyond the special benchmarks in Appendix G. Such policies must coordinate realized activations to remain feasible. A second direction is to quantify how investment changes feedback precision. The bounded-feedback model permits allocation-dependent variance. Specifying this relation would let policies account for information gained from scaling.

References

Agrawal P, Tulabandhula T (2020) Learning by repetition: Stochastic multi-armed bandits under priming effect. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, 470–479 (PMLR).

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47(2–3):235–256.

Badanidiyuru A, Kleinberg R, Slivkins A (2018) Bandits with knapsacks. *Journal of the ACM* 65(3):13:1–13:55.

Balseiro SR, Feldman J, Mirrokni V, Muthukrishnan S (2014) Yield optimization of display advertising with ad exchange. *Management Science* 60(12):2886–2907.

Balseiro SR, Lu H, Mirrokni V (2025) Regularized online allocation problems: Fairness and beyond. *Manufacturing & Service Operations Management* 27(3):720–735.

Bhaskaran S, Erat S (2025) Optimal prototyping with noisy measurements. *Manufacturing & Service Operations Management* 27(1):322–338.

Chen H, Chen YJ, Park SH, Shin D (2025) Multichannel advertising: Budget allocation in the presence of spillover and carryover effects. *Manufacturing & Service Operations Management* 27(3):862–880.

Chen W, Wang Y, Yuan Y, Wang Q (2016) Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research* 17(50):1–33.

Combes R, Talebi MS, Proutiere A, Lelarge M (2015) Combinatorial bandits revisited. *Advances in Neural Information Processing Systems*, volume 28, 2116–2124.

Dagan Y, Crammer K (2018) A better resource allocation algorithm with semi-bandit feedback. *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, 268–320 (PMLR).

Dudík M, Langford J, Li L (2011) Doubly robust policy evaluation and learning. *Proceedings of the 28th International Conference on Machine Learning*, 1097–1104.

Einbu JM (1978) Optimal allocations of continuous resources to several activities with a concave return function—some theoretical results. *Mathematics of Operations Research* 3(1):82–88.

Fontaine X, Mannor S, Perchet V (2020) An adaptive stochastic optimization algorithm for resource allocation. *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, 319–363 (PMLR).

Galgana R, Golrezaei N (2025) Learning in repeated multiunit pay-as-bid auctions. *Manufacturing & Service Operations Management* 27(1):200–229.

Howard SR, Ramdas A, McAuliffe J, Sekhon J (2021) Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* 49(2):1055–1080.

Kveton B, Wen Z, Ashkan A, Szepesvári C (2015) Tight regret bounds for stochastic combinatorial semi-bandits. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, 535–543 (PMLR).

Lattimore T, Crammer K, Szepesvári C (2014) Optimal resource allocation with semi-bandit feedback. *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, 477–486.

Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web*, 661–670 (Association for Computing Machinery).

Li X, Ye Y (2022) Online linear programming: Dual convergence, new algorithms, and regret bounds. *Operations Research* 70(5):2948–2966.

Little JDC (1979) Feature article—aggregate advertising models: The state of the art. *Operations Research* 27(4):629–667.

Moreno-Bote R, Ramírez-Ruiz J, Drugowitsch J, Hayden BY (2020) Heuristics and optimal solutions to the breadth–depth dilemma. *Proceedings of the National Academy of Sciences* 117(33):19799–19808.

Mourelatos E, Giannakopoulos N, Tzagarakis M (2025) Productivity and task heterogeneity in online labor markets: A bonus payment experiment. *Bulletin of Economic Research* 77(2):198–218.

Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19(4):1574–1609.

Rajaram K (2001) Assortment planning in fashion retailing: Methodology, application and analysis. *European Journal of Operational Research* 129(1):186–208.

Sankararaman KA, Slivkins A (2018) Combinatorial semi-bandits with knapsacks. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, 1760–1770 (PMLR).

Slivkins A (2019) Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning* 12(1–2):1–286.

Sun X, Zheng X, Li D (2013) Recent advances in mathematical programming with semi-continuous variables and cardinality constraint. *Journal of the Operations Research Society of China* 1(1):55–77.

Tang J, Chen B, Shi C (2024) Online learning for dual-index policies in dual-sourcing systems. *Manufacturing & Service Operations Management* 26(2):758–774.

Warmuth MK, Kuzmin D (2008) Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research* 9:2287–2320.

Wen Y (2025) Adversarial combinatorial semi-bandits with graph feedback. *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 66419–66436 (PMLR).

Zuo J, Joe-Wong C (2021) Combinatorial multi-armed bandits for resource allocation. *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 1–4 (IEEE).