SPEND NOW, EARN LATER: A BANDIT FRAMEWORK FOR DELAYED RESOURCE RENEWAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Bandits with Knapsacks (BwK) captures sequential decision-making under hard resource constraints, but standard models either treat resources as non-replenishable or assume that replenishment is immediately usable. We study *Delayed Replenishable BwK* (BwDRK), a stochastic multi-resource bandit model in which each productive pull consumes resources immediately while the induced replenishment is credited only after a tagged bounded delay. Delayed renewal creates a physical funds-in-flight effect: resources can be sufficient in aggregate but temporarily unavailable for the next decision. We propose a queue-aware primal-dual algorithm with tagged queues, a bounded-delay primal wrapper, and arrival-order dual updates. The analysis first gives a pathwise liquidity identity that converts non-active waiting/terminal rounds into a maximum active-prefix deficit, and then decomposes expected LP regret into liquidity, primal-learning, and dual-control terms. Under $E_t(a) \le \bar{d}$, the statistical learning terms scale as $\widetilde{\mathcal{O}}(\sqrt{K(1+\bar{d})T})$ up to resource-price constants. We also give a reserve-paced implementation based on the safe baseline, which enforces active-round slack and obtains complete sublinear expected regret against the corresponding reserve-aware comparator. The result separates the optimistic instantaneous-renewal LP from the operational price of maintaining lead-time liquidity.

1 INTRODUCTION

Resource constraints are central in online decision systems. Advertising platforms pace bids under campaign budgets, pricing and procurement systems spend inventory or cash, and compute platforms allocate scarce capacity. Bandits with Knapsacks (BwK) formalizes this tension by combining exploration with hard budget constraints (Badanidiyuru et al., 2018). The classical model is monotone: actions consume resources, budgets decrease, and the process stops when a resource is exhausted. Many systems are not monotone. Actions may also generate future resource renewal, as in conversion revenue returning to an advertising budget, replenishment shipments restoring inventory, or harvested energy restoring a storage device.

Replenishable BwK models account for this non-monotonicity by allowing actions to replenish resources (Kumar & Kleinberg, 2022; Bernasconi et al., 2023). A key idealization in these models is that generated replenishment becomes usable immediately. In operational systems, replenishment is often separated from the decision that creates it by a lead time. The cost of an impression is charged before conversion revenue is credited; inventory is spent before the replacement order arrives; energy is consumed before harvesting or charging completes. This temporal mismatch creates a *funds-in-flight* effect: a policy can be feasible in long-run net consumption and still run out of usable budget in the short run.

We study this effect through *Delayed Replenishable BwK* (BwDRK). In each productive round, the learner receives immediate bandit feedback on reward and consumption, while the associated replenishment vector arrives later through a tagged stochastic delay. The tag identifies the originating decision, so the learner can update the corresponding primal and dual information when the replenishment becomes available. The model is stochastic and multi-resource, with a null arm that can be used for pacing. Delayed replenishment is different from ordinary delayed feedback: it changes not only what the learner knows, but also which decisions are physically feasible at the next round.

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.

This distinction shapes both the algorithm and the benchmark. The instantaneous-renewal LP remains a useful canonical upper benchmark because it is the natural LP relaxation for BwK/BwRK after ignoring lead times. In BwDRK, however, the same LP is optimistic: it credits resources before they are usable. We therefore analyze two quantities in parallel. Against the instantaneous LP, we prove a decomposition into statistical learning error, dual-control error, and an explicit liquidity term. For a complete sublinear regret guarantee, we compare to a reserve-aware comparator produced by a safe-mixed implementation. This comparator reflects the resource reserve needed to operate under lead times, rather than treating all generated replenishment as immediately spendable.

Our algorithm is a queue-aware primal-dual method. The state machine invokes the learners when the current budget is safely usable, waits when a shortage occurs while replenishment remains in flight, and stops productive operation when no pending replenishment remains. The primal learning layer is modular: in the main instantiation we use a bounded-delay wrapper around EXP3-IX instances, so each base learner receives non-delayed feedback on its own local timeline. The dual layer processes complete net-violation vectors in tagged arrival order. The BwDRK-specific part of the analysis is a pathwise liquidity theorem: the number of non-active rounds is controlled by the maximum active-prefix excess deficit plus the bounded in-flight mass. A reserve-pacing layer based on a known safe baseline then makes this deficit controllable and yields a closed sublinear guarantee against the reserve-aware benchmark.

Contributions. Our contributions are threefold.

- We introduce BwDRK, a stochastic multi-armed and multi-resource bandit model with immediate consumption, tagged bounded replenishment delays, hard sequential feasibility, and a null action.
- We design a queue-aware primal-dual framework and prove a pathwise liquidity theorem showing that non-active rounds are controlled by active-prefix deficit and in-flight replenishment mass. This isolates the operational difficulty that is absent from ordinary delayed-feedback bandits.
- We instantiate the learning layer with a bounded-delay EXP3-IX wrapper and arrival-order projected dual updates. The statistical terms scale as $\widetilde{\mathcal{O}}(\sqrt{K(1+\bar{d})T})$ up to resource-price constants, and a safe-baseline mixing variant yields complete sublinear regret against a reserve-aware comparator.

2 Model and Benchmark

We consider a stochastic delayed-renewal bandit with a finite action set $\mathcal{A}$ of cardinality K and d replenishable resources. The action set contains a distinguished null arm a_0; the remaining arms $\mathcal{A}_+ := \mathcal{A} \setminus \{a_0\}$ are productive. Resource $j \in [d]$ starts with budget $B_j = \rho_j T$, and we write $\rho_{\min} := \min_j \rho_j$. For each productive arm a and round t, the potential outcome is $(R_t(a), \mathbf{C}_t(a), \mathbf{G}_t(a), E_t(a)) \sim P_a$, drawn independently across rounds before the learner's fresh randomization. Here $R_t(a) \in [0,1]$ is reward, $\mathbf{C}_t(a), \mathbf{G}_t(a) \in [0,1]^d$ are the immediate consumption and delayed replenishment vectors, and $E_t(a)$ is a non-negative integer delay. Only the outcome of the chosen arm is revealed. The null arm has $R_t(a_0) = 0$, $\mathbf{C}_t(a_0) = \mathbf{0}$, $\mathbf{G}_t(a_0) = \mathbf{0}$, and $E_t(a_0) = 0$.

Assumption 2.1 (Tagged bounded stochastic delays)**.** For every productive arm $a \in \mathcal{A}_+$, $0 \le E_t(a) \le \bar{d}$ almost surely for a known constant $\bar{d}$. When replenishment generated at round s, the learner observes the tagged tuple $(s, a_s, \mathbf{G}_s(a_s))$ and can match the feedback to its originating decision.

Assumption 2.2 (Budget regime)**.** Each resource budget scales linearly with the horizon: $B_j = \rho_j T$ with $\rho_j \in (0,1]$ independent of T.

Since each consumption coordinate is at most one, a resource with $B_j > T$ can never be exhausted even without replenishment; such resources can be removed.

Budget dynamics. At the beginning of round t, all tagged arrivals scheduled before that decision epoch are processed. A delay $E_s = 0$ means that the replenishment generated at round s is available before the next decision epoch. Thus, for every resource j,

$$B_{j,t} = B_j - \sum_{s=1}^{t-1} C_{j,s}(a_s) + \sum_{s=1}^{t-1} G_{j,s}(a_s)\mathbb{I}\{s + E_s(a_s) < t\}.$$

The analysis uses a conservative uniform safety rule: because $C_{j,t}(a) \le 1$, a productive pull is certainly safe whenever $B_{j,t} \ge 1$ for all j. If sharper deterministic upper bounds $u_j(a)$ with $C_{j,t}(a) \le u_j(a)$ are known, the same implementation can use the arm-specific safe set $\{a : u_j(a) \le B_{j,t}\ \forall j\}$; this may reduce unnecessary waiting, but the main LP-regret theorem is stated for the uniform rule so that the benchmark and liquidity accounting are aligned.

Pending queue and phases. Let $\mathcal{Q}_t := \{s < t :\ a_s \in \mathcal{A}_+,\ s + E_s(a_s) \ge t\}$ be the set of productive pulls whose replenishments are still in flight at the beginning of round t. The queue-aware state machine partitions the horizon into active rounds $\mathcal{T}_{\mathrm{a}}$, forced waiting rounds $\mathcal{T}_{\mathrm{w}}$, and terminal rounds $\mathcal{T}_{\varnothing}$:

$$\mathcal{T}_{\mathrm{a}} := \{t : B_{j,t} \ge 1,\ \forall j\},$$
$$\mathcal{T}_{\mathrm{w}} := \{t : \min_j B_{j,t} < 1 \text{ and } \mathcal{Q}_t \ne \emptyset\},$$
$$\mathcal{T}_{\varnothing} := \{t : \min_j B_{j,t} < 1 \text{ and } \mathcal{Q}_t = \emptyset\}.$$

On $\mathcal{T}_{\mathrm{a}}$ the primal learner is invoked and may choose either a productive arm or a_0. Rounds in $\mathcal{T}_{\mathrm{w}}$ are forced null rounds caused by in-flight replenishment, while $\mathcal{T}_{\varnothing}$ begins once a shortage occurs and no pending replenishment remains. We denote $N_{\mathrm{off}} := |\mathcal{T}_{\mathrm{w}}| + |\mathcal{T}_{\varnothing}|$. The realized cumulative delay of productive pulls is

$$D_T := \sum_{t=1}^{T} E_t(a_t)\mathbb{I}\{a_t \in \mathcal{A}_+\},$$

Instantaneous and reserve-aware LP benchmarks. Let $\bar{r}(a) = \mathbb{E}[R_t(a)]$, $\bar{\mathbf{c}}(a) = \mathbb{E}[\mathbf{C}_t(a)]$, and $\bar{\mathbf{g}}(a) = \mathbb{E}[\mathbf{G}_t(a)]$, with all three equal to zero for a_0. Here $\Delta(\mathcal{A}) := \{x \in \mathbb{R}_+^{|\mathcal{A}|} : \sum_a x(a) = 1\}$ is the probability simplex. The instantaneous-renewal LP is

$$\begin{aligned} \mathrm{OPT}^{\mathrm{LP}} := \max_{x \in \Delta(\mathcal{A})} & \sum_{a \in \mathcal{A}} x(a)\bar{r}(a) \\ \text{s.t.} & \sum_{a \in \mathcal{A}} x(a)(\bar{c}_j(a) - \bar{g}_j(a)) \le \rho_j,\ j \in [d]. \end{aligned}$$

This benchmark credits generated replenishment immediately. It is an optimistic upper benchmark for the delayed physical process. For a reserve vector $\mathbf{\Gamma} = (\Gamma_1, \dots, \Gamma_d) \in \mathbb{R}_+^d$ with $\Gamma_j < T\rho_j$ for every resource j, the reserve-adjusted value is

$$\begin{aligned} \mathrm{OPT}^{\mathrm{LP}}(\mathbf{\Gamma}) := \max_{x \in \Delta(\mathcal{A})} & \sum_{a \in \mathcal{A}} x(a)\bar{r}(a) \\ \text{s.t.} & \sum_{a \in \mathcal{A}} x(a)(\bar{c}_j(a) - \bar{g}_j(a)) \\ & \le \rho_j - \Gamma_j/T, \qquad j \in [d]. \end{aligned}$$

This delay-aware comparator sets aside Γ_j units of resource j for lead-time liquidity. The main decomposition applies to any feasible comparator of either LP.

Proposition 2.3 (LP upper bound)**.** *For any algorithm operating in BwDRK, the following holds:* $\mathbb{E}[\mathrm{REW}(T)] \le T\mathrm{OPT}^{\mathrm{LP}}$.

Definition 2.4 (LP regret)**.** The total reward is $\mathrm{REW}(T) := \sum_{t=1}^{T} R_t(a_t)$. The instantaneous LP regret is $\mathrm{Reg}_{\mathrm{LP}}(T) := T\mathrm{OPT}^{\mathrm{LP}} - \mathrm{REW}(T)$, and the reserve-adjusted regret is $\mathrm{Reg}_{\mathrm{LP},\mathbf{\Gamma}}(T) := T\mathrm{OPT}^{\mathrm{LP}}(\mathbf{\Gamma}) - \mathrm{REW}(T)$.

3 Queue-Aware Primal-Dual Algorithm

The algorithm has three components: a pool of primal bandit learners, a full-information dual learner, and a tagged replenishment queue. The primal pool implements a bounded-delay wrapper: since at most $\bar{d}$ productive feedback items can be pending at any decision epoch, $M := 1 + \bar{d}$ independent non-delayed bandit learners ensure that an active decision always uses an instance that is not waiting for its own previous feedback. This gives a transparent primal-learning instantiation; the liquidity analysis below is independent of the particular no-regret learner used inside each free instance.

3.1 LAGRANGIAN LOSSES

The dual domain is $\Lambda := \{\boldsymbol{\lambda} \in \mathbb{R}^d_+ : \|\boldsymbol{\lambda}\|_1 \le 1/\rho_{\min}\}$. For $\boldsymbol{\lambda}_t \in \Lambda$, define the primal Lagrangian utility

$$U_t(a; \boldsymbol{\lambda}_t) := R_t(a) + \boldsymbol{\lambda}_t^\top (\boldsymbol{\rho} + \mathbf{G}_t(a) - \mathbf{C}_t(a)).$$

For a_0, $U_t(a_0; \boldsymbol{\lambda}_t) = \boldsymbol{\lambda}_t^\top \boldsymbol{\rho}$. Since $R_t \in [0,1]$ and $\mathbf{C}_t, \mathbf{G}_t \in [0,1]^d$,

$$-U_{\mathrm{P}} \le U_t(a; \boldsymbol{\lambda}_t) \le U_{\mathrm{P}}, \qquad U_{\mathrm{P}} := 1 + 2/\rho_{\min}.$$

The bandit learners run in loss form. Once the complete feedback of round t is available, define

$$\ell_t(a; \boldsymbol{\lambda}_t) := \frac{U_{\mathrm{P}} - U_t(a; \boldsymbol{\lambda}_t)}{2U_{\mathrm{P}}} \in [0,1] \tag{1}$$

and, for the primal instance that chose $a_t \sim \pi_t$, feed the implicit-exploration estimator

$$\widehat{\ell}_t(a) := \frac{\ell_t(a_t; \boldsymbol{\lambda}_t)\mathbb{I}\{a_t = a\}}{\pi_t(a_t) + \xi}, \qquad a \in \mathcal{A}. \tag{2}$$

The denominator is applied after the affine transformation to a non-negative loss. The dual learner receives the net-violation vector

$$\mathbf{v}_t := \mathbf{C}_t(a_t) - \mathbf{G}_t(a_t) - \boldsymbol{\rho}.$$

Choice of primal learner. We utilize the EXP3-IX (Neu, 2015) inside the wrapper because the wrapper converts delayed tagged feedback into ordinary non-delayed bandit subsequences: each base instance is used again only after its previous feedback has arrived. Thus the base learner only needs a standard adversarial bandit guarantee for adaptive non-anticipating losses in $[0,1]$.

3.2 PROTOCOL

Algorithm 1 gives the implementation. A productive pull makes its primal instance busy until its tagged replenishment arrives. An active null pull has zero delay and is updated immediately. Forced waiting and terminal rounds do not update the learners, since they are imposed by physical liquidity rather than by a learner choice.

Reserve-paced sampling. The complete sublinear theorem is stated for the following reserve-paced variant. If a safe baseline mixture $x^{\mathrm{s}} \in \Delta(\mathcal{A})$ is available, the implementation may replace the primal sampling distribution π_t by

$$\widetilde{\pi}_t = (1-\alpha)\pi_t + \alpha x^{\mathrm{s}}, \qquad \alpha \in [0,1).$$

The action is sampled from $\widetilde{\pi}_t$, while the primal weights are updated through the same shifted loss estimator with denominator $\widetilde{\pi}_t(a_t) + \xi$. Setting $\alpha = 0$ recovers Algorithm 1.

At the beginning of any round, a productive pull remains pending for at most $\bar{d}$ subsequent decision epochs. Since at most one productive pull is generated per epoch, at most $\bar{d}$ primal instances can be busy. With $M = 1 + \bar{d}$, at least one instance is free whenever the state machine enters an active round.

4 ANALYSIS

The analysis is organized around three questions. First, how many rounds are lost because delayed replenishment is physically unavailable? Second, how much reward is lost by learning the Lagrangian arm values under bandit feedback and delayed replenishment? Third, under what implementable reserve condition is the liquidity term itself controlled? The answers are pathwise liquidity accounting, wrapped primal/arrival-order dual regret bounds, and a safe-baseline mixing theorem.

Algorithm 1 Queue-aware primal-dual algorithm for BwDRK

Input: $M = 1 + \bar{d}$ primal learners $\{\mathcal{A}^{\mathrm{P}}_m\}_{m=1}^{M}$, dual learner $\mathcal{A}^{\mathrm{D}}$, regularization ξ.
Initialize: budget $\mathbf{B} \leftarrow (B_1, \ldots, B_d)$; queue $\mathcal{Q} \leftarrow \emptyset$; all primal instances are free; $\boldsymbol{\lambda}_1 = \mathbf{0}$.
for $t = 1, 2, \ldots, T$ **do**
  Receive tagged arrivals $\mathsf{Arr}_t \subseteq \{(s, a_s, \mathbf{G}_s(a_s)) : s \in \mathrm{idx}(\mathcal{Q})\}$.
  for each $(s, a_s, \mathbf{G}_s(a_s)) \in \mathsf{Arr}_t$ in a fixed tie-breaking order **do**
    Recover $(s, m_s, a_s, \pi_s, \boldsymbol{\lambda}_s, R_s(a_s), \mathbf{C}_s(a_s))$ from $\mathcal{Q}$.
    Update budget $\mathbf{B} \leftarrow \mathbf{B} + \mathbf{G}_s(a_s)$.
    Update dual learner with $\mathbf{C}_s(a_s) - \mathbf{G}_s(a_s) - \boldsymbol{\rho}$.
    Form $\widehat{\ell}_s$ from (1)–(2); update $\mathcal{A}^{\mathrm{P}}_{m_s}$ with $\widehat{\ell}_s$.
    Mark instance m_s free and remove index s from $\mathcal{Q}$.
  end for
  if $\min_j B_j < 1$ **then**
    if $\mathcal{Q} = \emptyset$ **then**
      Play a_0 for the remaining rounds and stop learner updates.
      break
    else
      Play forced null arm a_0 and continue.
    end if
  else
    Choose any free primal instance m_t; set $\boldsymbol{\lambda}_t \leftarrow \mathcal{A}^{\mathrm{D}}.\mathrm{next}()$ and $\pi_t \leftarrow \mathcal{A}^{\mathrm{P}}_{m_t}.\mathrm{next}()$.
    Sample $a_t \sim \pi_t$; observe $R_t(a_t)$ and $\mathbf{C}_t(a_t)$.
    Update budget $\mathbf{B} \leftarrow \mathbf{B} - \mathbf{C}_t(a_t)$.
    if $a_t = a_0$ **then**
      Form the loss estimator for a_0 using $\mathbf{G}_t(a_0) = \mathbf{0}$ and update $\mathcal{A}^{\mathrm{P}}_{m_t}$ immediately.
      Update dual learner immediately with $-\boldsymbol{\rho}$; keep instance m_t free.
    else
      Store $(t, m_t, a_t, \pi_t, \boldsymbol{\lambda}_t, R_t(a_t), \mathbf{C}_t(a_t))$ in $\mathcal{Q}$ and mark m_t busy.
    end if
  end if
end for

4.1 Liquidity accounting

For the realized action sequence, define the instantaneous-renewal shadow budget

$$\widetilde{B}_{j,1} = B_j, \qquad \widetilde{B}_{j,t+1} = \widetilde{B}_{j,t} - C_{j,t}(a_t) + G_{j,t}(a_t). \tag{3}$$

The shadow system uses the same actions and outcomes as the real system, but it credits replenishment immediately. The gap between $\widetilde{B}$ and the real budget is precisely the funds in flight. Define the maximum active-prefix excess deficit by

$$\Gamma_T := \max_{j\in[d]} \max_{\tau\in[T]} \left[\sum_{t\in\mathcal{T}_{\mathrm{a}}:\, t<\tau} \big(C_{j,t}(a_t) - G_{j,t}(a_t) - \rho_j\big) \right]_+ . \tag{4}$$

Thus Γ_T measures the largest prefix-level net over-spending generated by active decisions after subtracting the per-round budget rate ρ_j.

Proposition 4.1 (Pathwise liquidity bound)**.** *Under Assumption 2.1, every trajectory of Algorithm 1 satisfies*

$$0 \le \widetilde{B}_{j,t} - B_{j,t} \le \bar{d}, \qquad \forall j, t, \tag{5}$$

$$N_{\mathrm{off}} \le \frac{1 + \bar{d} + \Gamma_T}{\rho_{\min}}. \tag{6}$$

This result is the main structural step. What matters for non-active rounds is the combination of in-flight mass, bounded by $\bar{d}$, and active-prefix deficit Γ_T. The subsequent reserve-pacing theorem is designed precisely to keep this prefix deficit small.

4.2 COMPARATOR REGRET AND RESERVE-AWARE BENCHMARKS

For a fixed comparator mixture $x \in \Delta(\mathcal{A})$, we denote that

$$\bar{r}(x) := \sum_a x(a)\bar{r}(a), \qquad U_t(x;\boldsymbol{\lambda}_t) := \sum_a x(a)U_t(a;\boldsymbol{\lambda}_t).$$

Concretely, x is LP-feasible if it satisfies the constraints in the instantaneous-renewal LP in Section 2. Define

$$R_P(x;T) := \mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{a}}} U_t(x;\boldsymbol{\lambda}_t) - U_t(a_t;\boldsymbol{\lambda}_t)\right], \tag{7}$$

$$R_D(T) := \mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{a}}} \boldsymbol{\lambda}_t^\top(\boldsymbol{\rho} + \mathbf{G}_t(a_t) - \mathbf{C}_t(a_t))\right]. \tag{8}$$

Here R_P is the primal learning loss on active rounds, while R_D is the residual dual-pressure term that appears when converting Lagrangian utility back to reward.

Lemma 4.2 (Feasible comparator utility)**.** *If x is LP-feasible, then for every round t,*

$$\mathbb{E}[\mathbb{I}\{t \in \mathcal{T}_{\mathrm{a}}\}U_t(x;\boldsymbol{\lambda}_t) \mid \mathcal{F}_{t-1}] \geq \bar{r}(x)\,\mathbb{I}\{t \in \mathcal{T}_{\mathrm{a}}\}.$$

Consequently,

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_{\mathrm{a}}} U_t(x;\boldsymbol{\lambda}_t)\right] \geq \bar{r}(x)\,\mathbb{E}[|\mathcal{T}_{\mathrm{a}}|].$$

Theorem 4.3 (Expected regret decomposition for any feasible comparator)**.** *For any LP-feasible comparator $x \in \Delta(\mathcal{A})$, Algorithm 1 satisfies*

$$\mathbb{E}[T\bar{r}(x) - \mathrm{REW}(T)] \leq \bar{r}(x)\,\mathbb{E}[N_{\mathrm{off}}] + R_P(x;T) + R_D(T).$$

In particular, if $x^\star$ solves the instantaneous-renewal LP, then

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{LP}}(T)] \leq \mathrm{OPT}^{\mathrm{LP}}\,\mathbb{E}[N_{\mathrm{off}}] + R_P(x^\star;T) + R_D(T).$$

Corollary 4.4 (Reserve-adjusted comparator)**.** *Let $x_{\boldsymbol{\Gamma}}$ be an optimizer of the reserve-adjusted LP defined in Section 2. Then*

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{LP},\boldsymbol{\Gamma}}(T)] \leq \mathrm{OPT}^{\mathrm{LP}}(\boldsymbol{\Gamma})\,\mathbb{E}[N_{\mathrm{off}}] + R_P(x_{\boldsymbol{\Gamma}};T) + R_D(T).$$

The corollary records that the same decomposition applies after replacing the optimistic LP by a reserve-adjusted comparator. The remaining liquidity term is then measured against the reward value of that reserve-aware target.

4.3 INSTANTIATING THE LEARNING TERMS

Each primal instance is EXP3-IX on $\mathcal{A}$ with

$$\eta^{\mathrm{P}} := \min\left\{\frac{1}{2}, \sqrt{\frac{(1+\bar{d})\ln K}{KT}}\right\}, \qquad \xi := \eta^{\mathrm{P}}/2.$$

The factor $1+\bar{d}$ is the number of non-delayed base instances in the wrapper; no distribution-ratio argument between decision and arrival times is used. The dual learner is projected gradient ascent on Λ, initialized at $\boldsymbol{\lambda}_1 = \mathbf{0}$, with

$$\eta^{\mathrm{D}} := \frac{1}{2\rho_{\min}\sqrt{d(1+\bar{d})T}}.$$

Proposition 4.5 (Primal wrapped bandit term)**.** *For every fixed comparator $x \in \Delta(\mathcal{A})$, there is a universal constant C_P such that*

$$R_P(x;T) \leq C_P U_{\mathrm{P}}\sqrt{K(1+\bar{d})T\ln K}.$$

Proposition 4.6 (Dual arrival-order term)**.** *There is a universal constant C_D such that*

$$R_D(T) \le C_D \frac{\sqrt{d(1+\bar{d})T}}{\rho_{\min}}.$$

Combining the previous results gives the unconditional decomposition

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{LP}}(T)] \le \mathrm{OPT}^{\mathrm{LP}}\, \mathbb{E}[N_{\mathrm{off}}] + C_P U_{\mathrm{P}} \sqrt{K(1+\bar{d})T \ln K} + C_D \frac{\sqrt{d(1+\bar{d})T}}{\rho_{\min}}.$$

The statistical part is therefore $\widetilde{\mathcal{O}}(\sqrt{K(1+\bar{d})T})$, up to resource-price constants. The term $\mathbb{E}[N_{\mathrm{off}}]$ is the operational price of physical liquidity shortage.

4.4 Algorithmically controlling liquidity by safe-baseline mixing

To obtain a complete sublinear theorem, the algorithm must keep Γ_T, and hence N_{off}, small. We do this through reserve-paced sampling with a known safe baseline mixture. The baseline can be the null arm itself, a conservative replenishing policy, or a domain-provided safety-stock policy.

Assumption 4.7 (Safe baseline)**.** There exists a known distribution $x^{\mathrm{s}} \in \Delta(\mathcal{A})$ and a margin $\gamma_0 > 0$ such that

$$\sum_a x^{\mathrm{s}}(a)\big(\bar{c}_j(a) - \bar{g}_j(a)\big) \le \rho_j - \gamma_0, \qquad j \in [d].$$

The null arm alone satisfies this assumption with any $\gamma_0 < \rho_{\min}$; stronger domain baselines improve the attainable comparator but are not needed for feasibility.

Given $\alpha \in [0,1)$, define

$$M_\alpha(z) := (1-\alpha)z + \alpha x^{\mathrm{s}}.$$

The reserve-paced implementation samples from $M_\alpha(\pi_t)$ in active rounds and uses importance weights relative to the actual sampling distribution. Its benchmark is

$$\mathrm{OPT}^{\mathrm{mix}}_\alpha := \max_{z \in \Delta(\mathcal{A})} \bar{r}(M_\alpha(z)).$$

Proposition 4.8 (Safe mixing enforces active-policy slack)**.** *Assume Assumption 4.7. Fix a target margin $\gamma \in (0, \gamma_0)$ and choose*

$$\alpha \ge \max_{j \in [d]} \frac{1-\rho_j+\gamma}{1-\rho_j+\gamma_0}.$$

Since $\gamma < \gamma_0$, the right-hand side of (4.8) *is strictly smaller than one, so such an $\alpha \in [0,1)$ exists. Then, for every history and every base distribution π_t, the safe-mixed active sampling distribution satisfies*

$$\mathbb{E}[C_{j,t}(a_t) - G_{j,t}(a_t) - \rho_j \mid \mathcal{F}_{t-1},\ t \in \mathcal{T}_{\mathrm{a}}] \le -\gamma, \qquad j \in [d].$$

Assumption 4.9 (Active-policy net slack)**.** There exists $\gamma \in (0,1]$ such that, on every active round and for every resource j,

$$\mathbb{E}[C_{j,t}(a_t) - G_{j,t}(a_t) - \rho_j \mid \mathcal{F}_{t-1},\ t \in \mathcal{T}_{\mathrm{a}}] \le -\gamma.$$

This condition is stated separately because it is the exact drift property needed for the liquidity proof. Proposition 4.8 gives an implementable sufficient condition.

Proposition 4.10 (Non-active rounds under active-policy slack)**.** *Under Assumption 4.9, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$N_{\mathrm{off}} \le \frac{1+\bar{d}+\frac{4}{\gamma}\log\Big(\frac{d}{\delta(1-e^{-\gamma^2/8})}\Big)}{\rho_{\min}}.$$

Consequently, setting $\delta = T^{-2}$ gives

$$\mathbb{E}[N_{\mathrm{off}}] \le \frac{1+\bar{d}+\frac{4}{\gamma}\log\Big(\frac{dT^2}{1-e^{-\gamma^2/8}}\Big)}{\rho_{\min}} + T^{-1}.$$

Theorem 4.11 (Complete reserve-aware guarantee)**.** *Under Assumptions 2.1 and 4.7, choose α satisfying* (4.8)*. Then the reserve-paced BwDRK implementation satisfies*

$$\mathbb{E}\big[T\mathrm{OPT}_{\alpha}^{\mathrm{mix}} - \mathrm{REW}(T)\big] \le \frac{\mathrm{OPT}_{\alpha}^{\mathrm{mix}}}{\rho_{\min}}\left(1 + \bar{d} + \frac{4}{\gamma}\log\left(\frac{dT^2}{1-e^{-\gamma^2/8}}\right)\right) + \frac{\mathrm{OPT}_{\alpha}^{\mathrm{mix}}}{T} + C'_P U_{\mathrm{P}}\sqrt{(1-\alpha)K(1+\bar{d})T\ln K} + C_D\frac{\sqrt{d(1+\bar{d})T}}{\rho_{\min}}. \quad (9)$$

Consequently, for fixed $K, d, \rho_{\min}^{-1}, U_{\mathrm{P}}, \gamma^{-1}$ and $\bar{d}$, the expected regret against $\mathrm{OPT}_{\alpha}^{\mathrm{mix}}$ is $\widetilde{O}(\sqrt{T})$.

The benchmark is reserve-aware: by Proposition 4.8, every $M_\alpha(z)$ has net consumption at most $\rho_j - \gamma$, and hence is feasible for $\mathrm{OPT}^{\mathrm{LP}}(\gamma T\mathbf{1})$. Thus

$$\mathrm{OPT}_{\alpha}^{\mathrm{mix}} \le \mathrm{OPT}^{\mathrm{LP}}(\gamma T\mathbf{1}) \le \mathrm{OPT}^{\mathrm{LP}}.$$

Theorem 4.11 controls the learning and operational loss relative to the reserve-aware target.

4.5 Dual pressure and active-prefix deficits

Dual prices guide the primal learner away from resource-expensive arms inside active rounds. The reserve-pacing layer provides the boundary safety used in the complete theorem, while the dual learner controls the Lagrangian trade-off within the active rounds. The exact algebraic bridge between dual regret and active-prefix deficit is the following.

Proposition 4.12 (Dual-deficit bridge)**.** *For any active prefix $I \subseteq \mathcal{T}_{\mathrm{a}}$ and any resource j,*

$$\sum_{t\in I}(C_{j,t}(a_t) - G_{j,t}(a_t) - \rho_j) \le \rho_{\min}R_I^D(e_j/\rho_{\min}) + \rho_{\min}\sum_{t\in I}\boldsymbol{\lambda}_t^\top(\mathbf{C}_t(a_t) - \mathbf{G}_t(a_t) - \boldsymbol{\rho}).$$

Thus a large active-prefix deficit can occur only if the dual regret on that prefix is large or if the realized dual pressure term is large. This is the sense in which dual prices contribute to resource control inside the active layer; reserve pacing controls boundary safety.

5 Experiments

We evaluate BwDRK on a multi-dimensional resource environment ($d = 2$) designed to isolate the funds-in-flight effect and test our algorithm's ability to achieve dynamic load balancing.

Setup and baselines. We set the horizon $T = 1000$ and initialize two independent resources at $B_{1,0} = B_{2,0} = 42$ to deliberately make the system vulnerable to liquidity crises. The action set contains $K = 4$ arms, with feedback drawn independently from continuous uniform distributions:

- Arm 0 (Null Arm): Zero reward, consumption, and replenishment.
- Arm 1 (R_1-Aggressive): High reward $\mathcal{U}(0.7, 0.9)$, heavy R_1 usage (consumption/replenishment $\sim \mathcal{U}(0.8, 1.0)$), and light R_2 usage ($\sim \mathcal{U}(0.1, 0.3)$ and $\sim \mathcal{U}(0.2, 0.4)$, respectively).
- Arm 2 (R_2-Aggressive): High reward $\mathcal{U}(0.7, 0.9)$, light R_1 usage ($\sim \mathcal{U}(0.1, 0.2)$), and heavy R_2 usage ($\sim \mathcal{U}(0.8, 1.0)$).
- Arm 3 (Conservative): Moderate reward $\mathcal{U}(0.4, 0.6)$, balanced usage for both resources (all metrics $\sim \mathcal{U}(0.3, 0.5)$).

By default, the multi-dimensional replenishment vector arrives after a stochastic delay $e_t \sim \mathcal{U}(50, 150)$. We compare BwDRK against: (1) **Standard PD**, which ignores delay and replenishment; (2) **Naive PD**, which allows replenishment but is queue-unaware; (3) **PD-BwK**, standard multi-resource primal-dual BwK that ignores in-flight queues; (4) **Greedy**, which maximizes immediate return under current budgets; and (5) **Oracle**, the no-delay LP upper bound.

Long-term reward and dynamic load balancing. As shown in Figure 1, the 1st panel demonstrates that while most baselines experience a brief initial gain, they eventually flatline and permanently

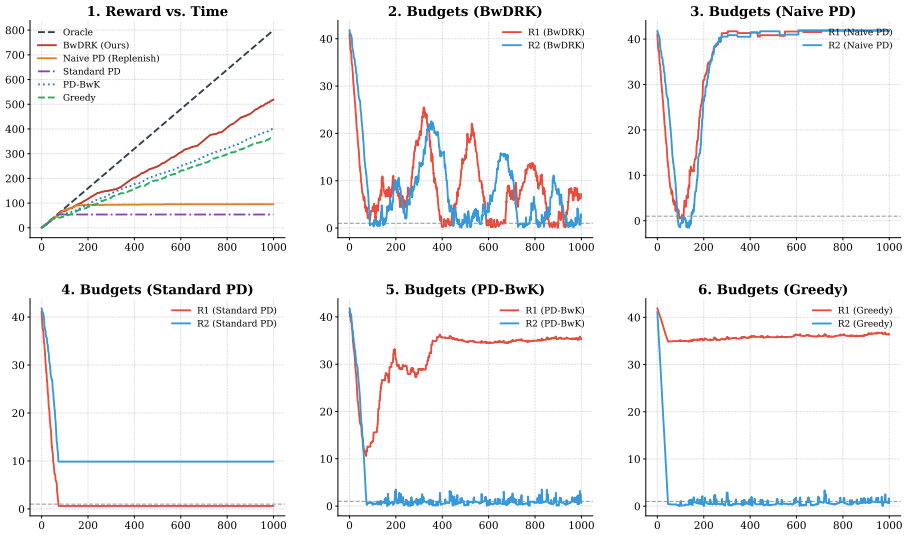


Figure 1: Cumulative reward and the budget trajectories under delay $e_t \sim \mathcal{U}(50, 150)$, comparing BwDRK to Naive PD with replenishment, Standard PD, PD-BwK, Greedy, and a no-delay LP Oracle.

stall. The underlying reason is revealed in the multidimensional budget trajectories. For delay-blind baselines like PD-BwK and Greedy, the early aggressive pursuit of high rewards depletes R_2 to zero.

In contrast, BwDRK successfully maintains a proper spend-wait-replenish cycle. Its budget trajectory exhibits periodic oscillations, safely keeping both resource balances. When R_1 approaches depletion but R_2 is sufficient, the queue-aware mechanism sharply increases the shadow price λ_1. This forces the algorithm to proactively avoid the R_1-Aggressive arm and pivot to the R_2-Aggressive or Conservative arm, allowing BwDRK to closely track the Oracle's growth.

Robustness to extreme delays and horizon scaling. As shown in Figure 2, the delay-sweep panel shows that as the delay interval expands, the margin for error sharply decreases. While all baselines suffer from severe degradation, BwDRK demonstrates graceful degradation, accumulating more than twice the final reward of the best baseline in the extreme regime. The N_{off} panel quantifies this resilience. Finally, the horizon-scaling plot extends to $T = 20,000$, confirming that BwDRK's average regret strictly decreases and converges toward zero, validating our theoretical sublinear regret bounds in operational environments.

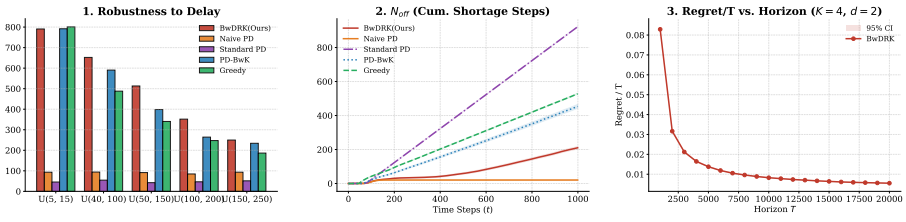


Figure 2: Final reward under increasing delay magnitudes, cumulative non-active rounds N_{off}, and normalized regret as the horizon grows.

6 Conclusion

In this work, we introduced BwDRK to address the practical temporal mismatch between resource consumption and renewal. Rather than assuming replenishment is immediately spendable, our model treats delayed renewal as a liquidity constraint driven by funds-in-flight. To navigate this, we developed a queue-aware primal-dual framework that separates optimistic-LP decomposition, pathwise liquidity accounting, and reserve-paced learning, which disentangles statistical learning error from the operational price of waiting, ultimately yielding robust sublinear regret.

AI USE STATEMENT

In this work, we used generative AI tools for cleaning and reformatting dataset, assisting with translation, cleaning and reformatting dataset. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, supporting qualitative and thematic data analysis, generating synthetic data sets, interpreting results and formulating mathematical claims. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the 15th ACM Conference on Economics and Computation (EC)*, 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 *Proceedings of the 29th Conference on Learning Theory (COLT)*, pp. 4–18, 2016.

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

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. Bandits with replenishable knapsacks: the best of both worlds. *arXiv preprint arXiv:2306.08470*, 2023.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1097–1105, 2014.

Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk Van Der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: when do intermediate observations help? *arXiv preprint arXiv:2305.19036*, 2023.

Giannis Fikioris, Robert Kleinberg, Yoav Kolumbus, Raunak Kumar, Yishay Mansour, and Éva Tardos. Learning in budgeted auctions with spacing objectives. *arXiv preprint arXiv:2411.04843*, 2024.

Arthur Flajolet and Patrick Jaillet. Logarithmic regret bounds for bandits with knapsacks. *arXiv preprint arXiv:1510.01800*, 2015.

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

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International conference on machine learning*, pp. 1453–1461. PMLR, 2013.

Thomas Kesselheim and Sahil Singla. Online learning with vector costs and bandits with knapsacks. In *Conference on Learning Theory*, pp. 2286–2305. PMLR, 2020.

Raunak Kumar and Robert Kleinberg. Non-monotonic resource utilization in the bandits with knapsacks problem. *Advances in Neural Information Processing Systems*, 35:19248–19259, 2022.

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pp. 5969–5978. PMLR, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Bingcong Li, Tianyi Chen, and Georgios B Giannakis. Bandit online learning with unknown delays. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 993–1002. PMLR, 2019.

Xiaocheng Li, Chunlin Sun, and Yinyu Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *International Conference on Machine Learning*, pp. 6483–6492. PMLR, 2021.

Kweiguu Liu, Setareh Maghsudi, and Makoto Yokoo. Budgeted recommendation with delayed feedback. In *World Conference on Information Systems and Technologies*, pp. 202–213. Springer, 2024.

Anne-Gaël Manegueu, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pp. 3348–3356. PMLR, 2020.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. *Advances in Neural Information Processing Systems*, 35:11752–11762, 2022.

Michael J Neely. *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool Publishers, 2010.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pp. 4105–4113. PMLR, 2018.

Hao Qiu, Emmanuel Esposito, and Mengxiao Zhang. Exploiting curvature in online convex optimization with delayed feedback. *arXiv preprint arXiv:2506.07595*, 2025.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Bandits with knapsacks beyond the worst case. *Advances in Neural Information Processing Systems*, 34:23191–23204, 2021.

Abdalaziz Sawwan and Jie Wu. Budget-constrained and deadline-driven multi-armed bandits with delays. In *2024 21st Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 1–9. IEEE, 2024.

Ofir Schlisselberg, Tal Lancewicki, Peter Auer, and Yishay Mansour. Improved best-of-both-worlds regret for bandits with delayed feedback. *arXiv preprint arXiv:2505.24193*, 2025.

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

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Dirk Van Der Hoeven and Nicolo Cesa-Bianchi. Nonstochastic bandits and experts with arm-dependent delays. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*, 2017.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pp. 9712–9721. PMLR, 2020.

Siwei Wang, Haoyun Wang, and Longbo Huang. Adaptive algorithms for multi-armed bandit with composite and anonymous feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10210–10217, 2021.

Jianyi Yang, Pengfei Li, Mohammad Jaminur Islam, and Shaolei Ren. Online allocation with replenishable budgets: Worst case and beyond. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 8(1):1–34, 2024.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3285–3294. PMLR, 2020.

Paul H Zipkin. *Foundations of Inventory Management*. McGraw-Hill, 2000.

A Omitted Proofs

This appendix gives the details omitted from the main text. We first prove the LP upper bound, then the liquidity identities, then the regret decomposition, and finally the learning and slack bounds. Throughout the proof, $\mathcal{F}_t$ denotes the sigma-field generated by all decisions, realized immediate feedback, tagged arrivals, and learner randomization up to the beginning of round $t+1$. Potential outcomes for round t are drawn after the beginning-of-round state is fixed and before the learner observes the chosen outcome. Hence, conditional on the history immediately before the fresh arm draw, the counterfactual loss vector used in the bandit analysis is fixed and non-anticipating.

A.1 LP upper bound

Proof of Proposition 2.3. Let $N_a(T) := \sum_{t=1}^{T} \mathbb{I}\{a_t = a\}$ be the number of times arm a is selected, including possible null pulls. Define the random empirical mixture $\widehat{x}(a) := N_a(T)/T$ and its expectation $x(a) := \mathbb{E}[\widehat{x}(a)]$. Then $x \in \Delta(\mathcal{A})$.

For each resource j, hard feasibility implies that total consumption cannot exceed the initial budget plus all replenishment generated by selected arms. Since ignoring whether some generated replenishment arrives after the horizon can only increase the right-hand side,

$$\begin{aligned} 0 &\le B_j + \mathbb{E}\left[\sum_{t=1}^{T} G_{j,t}(a_t) - C_{j,t}(a_t)\right] \\ &= \rho_j T + \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \mathbb{E}[\mathbb{I}\{a_t = a\}(\bar{g}_j(a) - \bar{c}_j(a))]\,. \end{aligned}$$

The equality uses that a_t is measurable before the round-t outcome is drawn. Dividing by T gives

$$\sum_{a \in \mathcal{A}} x(a)(\bar{c}_j(a) - \bar{g}_j(a)) \le \rho_j, \qquad j \in [d],$$

so x is feasible for the instantaneous-renewal LP. Similarly,

$$\mathbb{E}[\mathrm{REW}(T)] = \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \mathbb{E}[\mathbb{I}\{a_t = a\}\bar{r}(a)] = T \sum_{a} x(a)\bar{r}(a) \le T\mathrm{OPT}^{\mathrm{LP}}.$$

□

A.2 Liquidity accounting

For a resource j, define the in-flight replenishment mass at the beginning of round t by

$$F_{j,t} := \sum_{s<t:\, a_s \in \mathcal{A}_+,\, s+E_s(a_s) \ge t} G_{j,s}(a_s).$$

Lemma A.1 (Shadow-real gap)**.** *For every j, t, $\widetilde{B}_{j,t} - B_{j,t} = F_{j,t}$ and $0 \le F_{j,t} \le \bar{d}$.*

Proof. Subtracting the real budget formula (2) from the shadow recursion (3) gives

$$\begin{aligned}\widetilde{B}_{j,t} - B_{j,t} &= \sum_{s=1}^{t-1} G_{j,s}(a_s) - \sum_{s=1}^{t-1} G_{j,s}(a_s)\mathbb{I}\{s + E_s(a_s) < t\} \\ &= \sum_{s<t:\, s+E_s(a_s)\geq t} G_{j,s}(a_s) = F_{j,t}.\end{aligned}$$

The sum is non-negative. Since each productive pull contributes at most one unit to coordinate j, and only pulls from the last $\bar{d}$ decision epochs can still be in flight at the beginning of round t, $F_{j,t} \leq \bar{d}$. □

Lemma A.2 (Non-active rounds from prefix deficit)**.** *Every trajectory satisfies*

$$N_{\text{off}} \leq \frac{1 + \bar{d} + \Gamma_T}{\rho_{\min}}.$$

Proof. If $N_{\text{off}} = 0$, the claim is immediate. Otherwise let $t_\star$ be the first non-active round. Then there exists a resource $j_\star$ with $B_{j_\star,t_\star} < 1$. Let

$$n_\star := |\{t \in \mathcal{T}_{\text{a}} : t < t_\star\}|$$

be the number of active rounds before $t_\star$. Using (2), Lemma A.1, and $B_{j_\star} = \rho_{j_\star} T$,

$$\begin{aligned}1 &> B_{j_\star,t_\star} \\ &= \rho_{j_\star} T - \sum_{t\in\mathcal{T}_{\text{a}}:\, t<t_\star} C_{j_\star,t}(a_t) + \sum_{t\in\mathcal{T}_{\text{a}}:\, t<t_\star} G_{j_\star,t}(a_t) - F_{j_\star,t_\star} \\ &= \rho_{j_\star} T - \sum_{t\in\mathcal{T}_{\text{a}}:\, t<t_\star} \big(C_{j_\star,t}(a_t) - G_{j_\star,t}(a_t)\big) - F_{j_\star,t_\star} \\ &= \rho_{j_\star}(T - n_\star) - \sum_{t\in\mathcal{T}_{\text{a}}:\, t<t_\star} \big(C_{j_\star,t}(a_t) - G_{j_\star,t}(a_t) - \rho_{j_\star}\big) - F_{j_\star,t_\star} \\ &\geq \rho_{j_\star}(T - n_\star) - \Gamma_T - \bar{d}.\end{aligned}$$

Thus

$$T - n_\star < \frac{1 + \bar{d} + \Gamma_T}{\rho_{j_\star}} \leq \frac{1 + \bar{d} + \Gamma_T}{\rho_{\min}}.$$

Let $N_a := |\mathcal{T}_{\text{a}}|$ be the total number of active rounds. Since $N_a \geq n_\star$,

$$N_{\text{off}} = T - N_a \leq T - n_\star \leq \frac{1 + \bar{d} + \Gamma_T}{\rho_{\min}}.$$

□

Proof of Proposition 4.1. Equation (5) is Lemma A.1; equation (6) is Lemma A.2. □

A.3 Comparator utility and regret decomposition

Proof of Lemma 4.2. Condition on $\mathcal{F}_{t-1}$. On the event $\{t \in \mathcal{T}_{\text{a}}\}$, the multiplier $\boldsymbol{\lambda}_t$ is fixed before the round-t outcome is drawn. For any LP-feasible x,

$$\begin{aligned}\mathbb{E}[U_t(x;\boldsymbol{\lambda}_t) \mid \mathcal{F}_{t-1}] &= \sum_a x(a)\bar{r}(a) + \boldsymbol{\lambda}_t^\top \left(\boldsymbol{\rho} + \sum_a x(a)(\bar{\mathbf{g}}(a) - \bar{\mathbf{c}}(a))\right) \\ &= \bar{r}(x) + \boldsymbol{\lambda}_t^\top \left(\boldsymbol{\rho} - \sum_a x(a)(\bar{\mathbf{c}}(a) - \bar{\mathbf{g}}(a))\right) \\ &\geq \bar{r}(x),\end{aligned}$$

where the last inequality uses feasibility of x and $\boldsymbol{\lambda}_t \geq 0$. Multiplying by the $\mathcal{F}_{t-1}$-measurable indicator $\mathbb{I}\{t \in \mathcal{T}_{\text{a}}\}$ gives the conditional statement. Summing over t and taking expectations gives (4.2). □

Proof of Theorem 4.3. On active rounds,

$$R_t(a_t) = U_t(a_t; \boldsymbol{\lambda}_t) - \boldsymbol{\lambda}_t^\top (\boldsymbol{\rho} + \mathbf{G}_t(a_t) - \mathbf{C}_t(a_t)).$$

Forced waiting and terminal rounds earn zero reward. Hence

$$\begin{aligned}\mathbb{E}[\mathrm{REW}(T)] &= \mathbb{E}\left[\sum_{t\in\mathcal{T}_\mathrm{a}} U_t(a_t;\boldsymbol{\lambda}_t)\right] - \mathbb{E}\left[\sum_{t\in\mathcal{T}_\mathrm{a}} \boldsymbol{\lambda}_t^\top(\boldsymbol{\rho} + \mathbf{G}_t(a_t) - \mathbf{C}_t(a_t))\right] \\ &\geq \mathbb{E}\left[\sum_{t\in\mathcal{T}_\mathrm{a}} U_t(x;\boldsymbol{\lambda}_t)\right] - R_P(x;T) - R_D(T) \\ &\geq \bar{r}(x)\,\mathbb{E}[|\mathcal{T}_\mathrm{a}|] - R_P(x;T) - R_D(T).\end{aligned}$$

Since $|\mathcal{T}_\mathrm{a}| = T - N_\mathrm{off}$,

$$\mathbb{E}[\mathrm{REW}(T)] \geq T\bar{r}(x) - \bar{r}(x)\,\mathbb{E}[N_\mathrm{off}] - R_P(x;T) - R_D(T).$$

Rearranging proves (4.3). Choosing an optimal LP solution $x^\star$ gives (4.3). □

A.4 Primal learning bound via a delay wrapper

We first prove a non-delayed EXP3-IX bound for one primal instance, then sum it over the $M = 1+\bar{d}$ instances. The proof is written for an adaptive loss sequence: at the moment an instance chooses p_s, the loss vector $\ell_s \in [0,1]^K$ may depend on the past but not on the fresh random arm draw of that instance. This is the only property needed by the exponential-weights proof.

Lemma A.3 (Delay-wrapper invariant)**.** *With $M = 1 + \bar{d}$ primal instances, Algorithm 1 always has a free primal instance at the beginning of every active decision. Moreover, for each fixed instance, its local feedback sequence is non-delayed: before the instance is used for its next decision, the complete loss vector estimator from its previous decision has already been processed.*

Proof. Consider the beginning of a real round t after all arrivals scheduled for that epoch have been processed. A productive decision at round $s < t$ remains pending exactly when $s + E_s(a_s) \geq t$. Since $E_s(a_s) \leq \bar{d}$, only productive decisions from rounds $t-\bar{d}, t-\bar{d}+1, \ldots, t-1$ can still be pending. At most one productive decision is made per real round, hence at most $\bar{d}$ productive decisions are pending. Active null pulls have delay zero and are updated immediately, so they never make an instance busy. Therefore at most $\bar{d}$ of the $1+\bar{d}$ instances are busy, and at least one instance is free.

When an instance is used for a productive pull, Algorithm 1 marks it busy and does not use it again until the tagged replenishment of that pull arrives. At that arrival time the algorithm reconstructs the shifted loss estimator, updates that instance, and only then marks it free. Therefore, in the local timeline of any fixed instance, each new action is selected after the feedback from the previous local action has been processed. Multiple arrivals in the same real round are handled by the fixed tie-breaking order in the algorithm, which simply induces a deterministic order of local updates and does not create delayed feedback within any instance. □

Lemma A.4 (EXP3-IX on one non-delayed subsequence)**.** *Consider any non-anticipating sequence of losses $\ell_1, \ldots, \ell_n \in [0,1]^K$. EXP3-IX with learning rate $\eta \leq 1/2$, estimator $\widehat{\ell}_s(a) = \ell_s(a_s)\mathbb{I}\{a_s = a\}/(p_s(a_s) + \xi)$, and $\xi = \eta/2$ satisfies, for every $x \in \Delta(\mathcal{A})$,*

$$\mathbb{E}\left[\sum_{s=1}^{n} \langle p_s - x, \ell_s\rangle\right] \leq \frac{\ln K}{\eta} + 2\eta K n. \tag{10}$$

Proof. Let $w_1(a) = 1$ and $p_s(a) = w_s(a)/W_s$, where $W_s = \sum_a w_s(a)$. The update is $w_{s+1}(a) = w_s(a)\exp(-\eta\widehat{\ell}_s(a))$. Since $\widehat{\ell}_s(a) \geq 0$ and $\eta \leq 1/2$, $e^{-z} \leq 1 - z + z^2$ gives

$$\begin{aligned}\ln\frac{W_{s+1}}{W_s} &= \ln\sum_a p_s(a) e^{-\eta\widehat{\ell}_s(a)} \\ &\leq \ln\left(1 - \eta\langle p_s, \widehat{\ell}_s\rangle + \eta^2\sum_a p_s(a)\widehat{\ell}_s(a)^2\right) \\ &\leq -\eta\langle p_s, \widehat{\ell}_s\rangle + \eta^2\sum_a p_s(a)\widehat{\ell}_s(a)^2.\end{aligned} \tag{11}$$

For any comparator distribution x, the log-sum lower bound gives

$$\ln \frac{W_{n+1}}{W_1} \geq \sum_a x(a) \ln \frac{w_{n+1}(a)}{w_1(a)} - \ln K = -\eta \sum_{s=1}^{n} \langle x, \widehat{\ell}_s \rangle - \ln K. \tag{12}$$

Combining (11) and (12),

$$\sum_{s=1}^{n} \langle p_s, \widehat{\ell}_s \rangle - \sum_{s=1}^{n} \langle x, \widehat{\ell}_s \rangle \leq \frac{\ln K}{\eta} + \eta \sum_{s=1}^{n} \sum_a p_s(a) \widehat{\ell}_s(a)^2. \tag{13}$$

Condition on the history before the fresh draw $a_s \sim p_s$. Then

$$\mathbb{E}[\langle p_s, \widehat{\ell}_s \rangle \mid \mathcal{H}_{s-1}] = \sum_a \frac{p_s(a)^2 \ell_s(a)}{p_s(a) + \xi} \geq \langle p_s, \ell_s \rangle - K\xi, \tag{14}$$

$$\mathbb{E}[\langle x, \widehat{\ell}_s \rangle \mid \mathcal{H}_{s-1}] = \sum_a \frac{x(a) p_s(a) \ell_s(a)}{p_s(a) + \xi} \leq \langle x, \ell_s \rangle, \tag{15}$$

$$\mathbb{E}\left[\sum_a p_s(a) \widehat{\ell}_s(a)^2 \mid \mathcal{H}_{s-1} \right] = \sum_a \frac{p_s(a)^2 \ell_s(a)^2}{(p_s(a) + \xi)^2} \leq K. \tag{16}$$

The first inequality uses $p^2/(p+\xi) = p - p\xi/(p+\xi) \geq p - \xi$ and $\ell_s(a) \leq 1$. Taking expectations in (13) and using (14)–(16),

$$\mathbb{E}\left[\sum_{s=1}^{n} \langle p_s - x, \ell_s \rangle \right] \leq \frac{\ln K}{\eta} + \eta K n + K\xi n.$$

With $\xi = \eta/2$, the right-hand side is at most $\ln K/\eta + 2\eta K n$. □

Lemma A.5 (Primal wrapper regret)**.** *Let $M = 1 + \bar{d}$ primal instances be used as in Algorithm 1, each with learning rate* (4.3)*. Then for every $x \in \Delta(\mathcal{A})$,*

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_{\mathrm{a}}} \langle \pi_t - x, \ell_t(\cdot; \boldsymbol{\lambda}_t) \rangle \right] \leq C \sqrt{K(1+\bar{d}) T \ln K} \tag{17}$$

for a universal constant C.

Proof. By Lemma A.3, the local sequence seen by each fixed instance is non-delayed. Let $\mathcal{H}_{s-1}^{(m)}$ be the sigma-field generated by all global randomness and all local feedback processed by instance m before its s-th local decision. The dual vector, the safe-mixing distribution if used, and the counterfactual potential outcomes for that real round are fixed before the fresh local arm draw. Hence the local shifted loss vector is $\mathcal{H}_{s-1}^{(m)}$-measurable, and the non-anticipation condition of Lemma A.4 holds.

Let n_m be the number of active decisions assigned to instance m. Applying Lemma A.4 conditionally to each instance and summing over m,

$$\begin{aligned} \mathbb{E}\left[\sum_{t \in \mathcal{T}_{\mathrm{a}}} \langle \pi_t - x, \ell_t \rangle \right] &\leq \sum_{m=1}^{M} \left(\frac{\ln K}{\eta^{\mathrm{P}}} + 2\eta^{\mathrm{P}} K n_m \right) \\ &= \frac{M \ln K}{\eta^{\mathrm{P}}} + 2\eta^{\mathrm{P}} K |\mathcal{T}_{\mathrm{a}}| \\ &\leq \frac{(1+\bar{d}) \ln K}{\eta^{\mathrm{P}}} + 2\eta^{\mathrm{P}} K T. \end{aligned}$$

Substituting the chosen value of η^P gives (17) after increasing constants. If the minimum in (4.3) selects $1/2$, the same bound follows because this can only happen for small $KT/((1+\bar{d})\ln K)$, which is absorbed by the universal constant. □

Proof of Proposition 4.5. The affine definition of ℓ_t in Section 3.1 implies, conditional on $\mathcal{F}_{t-1}$ and $a_t \sim \pi_t$,

$$\mathbb{E}[U_t(x; \boldsymbol{\lambda}_t) - U_t(a_t; \boldsymbol{\lambda}_t) \mid \mathcal{F}_{t-1}] = 2U_{\mathrm{P}} \langle \pi_t - x, \ell_t(\cdot; \boldsymbol{\lambda}_t) \rangle.$$

Taking expectations and applying Lemma A.5 proves (4.5). □

A.5 DUAL ARRIVAL-ORDER PROJECTED-GRADIENT BOUND

List active decisions in chronological order by n and write $t(n)$ for their real round. Let

$$\mathbf{v}_n := \mathbf{C}_{t(n)}(a_{t(n)}) - \mathbf{G}_{t(n)}(a_{t(n)}) - \boldsymbol{\rho}.$$

Then $\|\mathbf{v}_n\|_2 \le 2\sqrt{d}$. The dual learner processes these vectors in tagged arrival order; active null pulls have zero delay.

Lemma A.6 (Intervening-update count)**.** *For each active item n, let $\mathcal{J}_n$ be the set of processed dual updates that occur after the decision at $t(n)$ and before item n is processed. Then $|\mathcal{J}_n| \le 2(1+\bar{d})$, and hence $\sum_n |\mathcal{J}_n| \le 2(1+\bar{d})T$.*

Proof. Let $r(n)$ be the real round in which item n is processed. Since $r(n) - t(n) \le 1 + \bar{d}$ under the arrival-first convention, every update in $\mathcal{J}_n$ is processed in the real-time interval $[t(n), r(n)]$. An item processed in this interval must have been generated no earlier than $t(n) - \bar{d}$ and no later than $r(n)$. The number of real rounds in $[t(n) - \bar{d}, r(n)]$ is at most $2\bar{d} + 2$, and at most one active item is generated per real round. Therefore $|\mathcal{J}_n| \le 2(1+\bar{d})$. Summing over $n \le T$ proves the second claim. □

Lemma A.7 (Arrival-order telescope)**.** *Projected gradient ascent on Λ with step size η^{D} satisfies, for any $\lambda \in \Lambda$,*

$$\sum_n \langle \lambda - \lambda_n, \mathbf{v}_n \rangle \le \frac{\|\lambda\|_2^2}{2\eta^{\mathrm{D}}} + C\eta^{\mathrm{D}} d(1+\bar{d})T. \tag{18}$$

Proof. Let $\sigma(m)$ be the original active index of the m-th processed dual feedback item, and let μ_m be the dual vector immediately before this processed update. The actual update is

$$\mu_{m+1} = \Pi_\Lambda(\mu_m + \eta^{\mathrm{D}} \mathbf{v}_{\sigma(m)}).$$

By non-expansiveness of projection, for any fixed $\lambda \in \Lambda$,

$$\begin{aligned} \|\mu_{m+1} - \lambda\|_2^2 &\le \|\mu_m + \eta^{\mathrm{D}} \mathbf{v}_{\sigma(m)} - \lambda\|_2^2 \\ &= \|\mu_m - \lambda\|_2^2 + 2\eta^{\mathrm{D}} \langle \mathbf{v}_{\sigma(m)}, \mu_m - \lambda \rangle + (\eta^{\mathrm{D}})^2 \|\mathbf{v}_{\sigma(m)}\|_2^2. \end{aligned}$$

Rearranging and summing over processed updates gives

$$\sum_m \langle \lambda - \mu_m, \mathbf{v}_{\sigma(m)} \rangle \le \frac{\|\lambda\|_2^2}{2\eta^{\mathrm{D}}} + 2\eta^{\mathrm{D}} dT, \tag{19}$$

where $\mu_1 = 0$ and $\|\mathbf{v}_n\|_2^2 \le 4d$.

Let $m(n)$ be the processing index of item n. Decompose

$$\sum_n \langle \lambda - \lambda_n, \mathbf{v}_n \rangle = \sum_n \langle \lambda - \mu_{m(n)}, \mathbf{v}_n \rangle + \sum_n \langle \mu_{m(n)} - \lambda_n, \mathbf{v}_n \rangle. \tag{20}$$

The first term equals the left side of (19). For the second term, repeated projection non-expansiveness along the actual arrival-order path gives

$$\|\mu_{m(n)} - \lambda_n\|_2 \le \eta^{\mathrm{D}} \sum_{u \in \mathcal{J}_n} \|\mathbf{v}_u\|_2.$$

Thus, using $\|\mathbf{v}_u\|_2 \le 2\sqrt{d}$ and Lemma A.6,

$$\begin{aligned} \sum_n \langle \mu_{m(n)} - \lambda_n, \mathbf{v}_n \rangle &\le \eta^{\mathrm{D}} \sum_n \sum_{u \in \mathcal{J}_n} \|\mathbf{v}_u\|_2 \|\mathbf{v}_n\|_2 \\ &\le 4\eta^{\mathrm{D}} d \sum_n |\mathcal{J}_n| \\ &\le 8\eta^{\mathrm{D}} d(1+\bar{d})T. \end{aligned}$$

Combining this with (19) proves (18). □

Proof of Proposition 4.6. Apply Lemma A.7 with comparator $\lambda = 0$. The left side becomes

$$\sum_n \langle 0 - \lambda_n, \mathbf{v}_n \rangle = \sum_{t \in \mathcal{T}_\mathrm{a}} \boldsymbol{\lambda}_t^\top (\boldsymbol{\rho} + \mathbf{G}_t(a_t) - \mathbf{C}_t(a_t)).$$

The comparator term vanishes. Substituting the chosen value of η^D yields the claimed bound, after increasing the universal constant. Taking expectations preserves the inequality. □

A.6 Controlling non-active rounds under active-policy slack

Lemma A.8 (Exponential maximal inequality for a slack prefix)**.** *Let $(Y_n)_{n\geq 1}$ be adapted to a filtration $(\mathcal{H}_n)$, satisfy $Y_n \in [-2, 2]$, and obey $\mathbb{E}[Y_n \mid \mathcal{H}_{n-1}] \leq -\gamma$ for some $\gamma \in (0, 1]$. Then, for every $u > 0$ and every horizon T,*

$$\Pr\left(\max_{m \leq T} \sum_{n=1}^{m} Y_n \geq u\right) \leq \frac{\exp(-\gamma u/4)}{1 - e^{-\gamma^2/8}}.$$

Proof. Set $\theta = \gamma/4$. By Hoeffding's lemma for variables supported in an interval of length at most 4,

$$\begin{aligned}\mathbb{E}[e^{\theta Y_n} \mid \mathcal{H}_{n-1}] &= e^{\theta \mathbb{E}[Y_n \mid \mathcal{H}_{n-1}]} \mathbb{E}\big[e^{\theta(Y_n - \mathbb{E}[Y_n \mid \mathcal{H}_{n-1}])} \mid \mathcal{H}_{n-1}\big] \\ &\leq e^{-\theta\gamma + 2\theta^2} = e^{-\gamma^2/8} =: q < 1.\end{aligned}$$

Consequently, for every fixed m, iterating conditional expectations gives

$$\mathbb{E} \exp\left(\theta \sum_{n=1}^{m} Y_n\right) \leq q^m.$$

By Markov's inequality and a union bound over $m \leq T$,

$$\begin{aligned}\Pr\left(\max_{m \leq T} \sum_{n=1}^{m} Y_n \geq u\right) &\leq \sum_{m=1}^{T} \Pr\left(\sum_{n=1}^{m} Y_n \geq u\right) \\ &\leq e^{-\theta u} \sum_{m=1}^{T} q^m \\ &\leq \frac{e^{-\gamma u/4}}{1 - e^{-\gamma^2/8}}.\end{aligned}$$

□

Proof of Proposition 4.10. List active rounds in chronological order as $t(1), t(2), \ldots, t(N)$. For each resource j, define

$$Y_{n,j} := C_{j,t(n)}(a_{t(n)}) - G_{j,t(n)}(a_{t(n)}) - \rho_j.$$

By Assumption 4.9, $\mathbb{E}[Y_{n,j} \mid \mathcal{F}_{t(n)-1}] \leq -\gamma$, and $Y_{n,j} \in [-2, 2]$. Applying Lemma A.8 to each resource and taking a union bound over $j \in [d]$ gives, with probability at least $1 - \delta$,

$$\Gamma_T = \max_{j \in [d]} \max_{m \leq T} \left[\sum_{n=1}^{m} Y_{n,j}\right]_+ \leq \frac{4}{\gamma} \log\left(\frac{d}{\delta(1 - e^{-\gamma^2/8})}\right).$$

Substituting this event into Proposition 4.1 proves (4.10). For the expectation bound, set $\delta = T^{-2}$ in (4.10). On the failure event use the deterministic bound $N_{\mathrm{off}} \leq T$. This yields (4.10). □

A.7 Safe-baseline mixing

Proof of Proposition 4.8. Fix a history, a base distribution π_t, and a resource j. Since $C_j - G_j \leq 1$ coordinate-wise for every arm,

$$\begin{aligned}\mathbb{E}[C_{j,t}(a_t) - G_{j,t}(a_t) \mid \mathcal{F}_{t-1}, t \in \mathcal{T}_\mathrm{a}] &\leq (1 - \alpha) \cdot 1 + \alpha \sum_a x^\mathrm{s}(a)(\bar{c}_j(a) - \bar{g}_j(a)) \\ &\leq 1 - \alpha + \alpha(\rho_j - \gamma_0).\end{aligned}$$

Condition (4.8) is exactly

$$1 - \alpha + \alpha(\rho_j - \gamma_0) \le \rho_j - \gamma,$$

for every j. Therefore the safe-mixed sampling distribution satisfies Assumption 4.9 with margin γ. □

Lemma A.9 (EXP3-IX with external safe mixing)**.** *Let a primal instance maintain weights p_s but sample from $\widetilde{p}_s = (1-\alpha)p_s + \alpha x^{\mathrm{s}}$, with $\alpha \in [0,1)$. The estimator is $\widehat{\ell}_s(a) = \ell_s(a_s)\mathbb{I}\{a_s = a\}/(\widetilde{p}_s(a_s) + \xi)$, and the weights are updated using $\widehat{\ell}_s$. For any non-anticipating loss sequence $\ell_s \in [0,1]^K$ and any $z \in \Delta(\mathcal{A})$,*

$$\mathbb{E}\left[\sum_{s=1}^{n} \langle p_s - z, \ell_s \rangle\right] \le \frac{\ln K}{\eta} + \frac{2\eta K n}{1-\alpha},$$

when $\eta \le 1/2$ and $\xi = \eta/2$.

Proof. The potential argument is identical to Lemma A.4 up to the moment where expectations of the estimator are evaluated. Conditional on the past before the fresh draw $a_s \sim \widetilde{p}_s$,

$$\begin{aligned}
\mathbb{E}[\langle p_s, \widehat{\ell}_s \rangle] &= \sum_a \frac{p_s(a)\widetilde{p}_s(a)\ell_s(a)}{\widetilde{p}_s(a) + \xi} \ge \langle p_s, \ell_s \rangle - \sum_a \frac{p_s(a)\xi}{\widetilde{p}_s(a) + \xi} \\
&\ge \langle p_s, \ell_s \rangle - \frac{K\xi}{1-\alpha}, \\
\mathbb{E}[\langle z, \widehat{\ell}_s \rangle] &= \sum_a \frac{z(a)\widetilde{p}_s(a)\ell_s(a)}{\widetilde{p}_s(a) + \xi} \le \langle z, \ell_s \rangle, \\
\mathbb{E}\left[\sum_a p_s(a)\widehat{\ell}_s(a)^2\right] &= \sum_a \frac{p_s(a)\widetilde{p}_s(a)\ell_s(a)^2}{(\widetilde{p}_s(a) + \xi)^2} \le \sum_a \frac{p_s(a)}{\widetilde{p}_s(a)} \le \frac{K}{1-\alpha}.
\end{aligned}$$

The last inequalities use $\widetilde{p}_s(a) \ge (1-\alpha)p_s(a)$. Substituting these three estimates into the same potential inequality as in (13) gives

$$\mathbb{E}\left[\sum_{s=1}^{n} \langle p_s - z, \ell_s \rangle\right] \le \frac{\ln K}{\eta} + \frac{\eta K n}{1-\alpha} + \frac{K\xi n}{1-\alpha}.$$

With $\xi = \eta/2$ and a harmless constant enlargement, the displayed bound follows. □

Proof of Theorem 4.11. Let $z^\star \in \arg\max_z \bar{r}(M_\alpha(z))$ and set $x^\star_\alpha = M_\alpha(z^\star)$. By Proposition 4.8, the safe-mixed algorithm satisfies Assumption 4.9 with margin γ, so Proposition 4.10 controls $\mathbb{E}[N_{\mathrm{off}}]$.

It remains to state the primal term. On each active round, the actual sampling distribution and the safe-mixed comparator differ by

$$M_\alpha(\pi_t) - M_\alpha(z^\star) = (1-\alpha)(\pi_t - z^\star).$$

Applying Lemma A.9 on each of the $M = 1 + \bar{d}$ wrapper instances and summing as in Lemma A.5, with

$$\eta = \min\left\{\frac{1}{2}, \sqrt{\frac{(1-\alpha)(1+\bar{d})\ln K}{KT}}\right\},$$

gives the utility regret bound

$$R^\alpha_P(z^\star; T) \le C'_P U_{\mathrm{P}} \sqrt{(1-\alpha)K(1+\bar{d})T \ln K}.$$

The regret decomposition Theorem 4.3 applies to the feasible comparator $x^\star_\alpha$; Proposition 4.6 controls the dual term. Combining these three bounds yields (9). □

Lemma A.10 (Dual-deficit bridge)**.** *For any active prefix $I \subseteq \mathcal{T}_{\mathrm{a}}$ and any resource j,*

$$\sum_{t \in I}(C_{j,t}(a_t) - G_{j,t}(a_t) - \rho_j) \le \rho_{\min} R^D_I(e_j/\rho_{\min}) + \rho_{\min} \sum_{t \in I} \boldsymbol{\lambda}_t^\top (\mathbf{C}_t(a_t) - \mathbf{G}_t(a_t) - \boldsymbol{\rho}).$$

Proof. The dual regret on I against $e_j/\rho_{\min}$ is

$$R_I^D(e_j/\rho_{\min}) \geq \sum_{t\in I}\left\langle \frac{e_j}{\rho_{\min}} - \boldsymbol{\lambda}_t, \mathbf{C}_t(a_t) - \mathbf{G}_t(a_t) - \boldsymbol{\rho}\right\rangle$$
$$= \frac{1}{\rho_{\min}}\sum_{t\in I}(C_{j,t}(a_t) - G_{j,t}(a_t) - \rho_j) - \sum_{t\in I}\boldsymbol{\lambda}_t^\top(\mathbf{C}_t(a_t) - \mathbf{G}_t(a_t) - \boldsymbol{\rho}).$$

Rearranging gives the claim. □

A.8 A simple sufficient condition: uniform net slack

Assumption 4.9 is policy-level. A stronger but easy-to-check sufficient condition is that every arm is slack: there exists $\gamma > 0$ such that $\bar{c}_j(a) - \bar{g}_j(a) \leq \rho_j - \gamma$ for every a and j. Then any active randomized policy automatically satisfies Assumption 4.9. This stronger condition is not needed for the main statement and is mainly useful as a transparent sanity check.

B Buffered Shadow-Feasible Case

This section records a stronger sufficient condition under which the explicit liquidity term in Section 4.3 disappears. The main text does not rely on this condition; it is included here to clarify the relationship between the general decomposition and the cleanest sublinear regime.

Corollary B.1 (Buffered shadow regime)**.** *If the realized shadow process generated by Algorithm 1 satisfies*

$$\min_{j\in[d],\, t\in[T]} \widetilde{B}_{j,t} \geq 1 + \bar{d},$$

then $\mathcal{T}_{\mathrm{w}} = \mathcal{T}_{\varnothing} = \emptyset$. Consequently,

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{LP}}(T)] \leq C_P U_{\mathrm{P}}\sqrt{K(1+\bar{d})T(1+\ln K)} + C_D\,\frac{\sqrt{d(1+\bar{d})T}}{\rho_{\min}}.$$

In particular, if $\bar{d} = \mathcal{O}(1)$, the statistical part is $\widetilde{\mathcal{O}}(\sqrt{T})$ up to problem-dependent constants.

Proof. By Lemma A.1, for every j, t,

$$B_{j,t} = \widetilde{B}_{j,t} - (\widetilde{B}_{j,t} - B_{j,t}) \geq (1+\bar{d}) - \bar{d} = 1.$$

Therefore the state machine never enters the waiting condition $\min_j B_{j,t} < 1$ with a nonempty queue, and it never enters the terminal condition $\min_j B_{j,t} < 1$ with an empty queue. Thus $N_{\mathrm{off}} = 0$. Substituting $\mathbb{E}[N_{\mathrm{off}}] = 0$ into Section 4.3 gives the displayed inequality. □

The preceding condition is pathwise and may be stronger than necessary. The next proposition gives one simple stochastic condition under which it holds with high probability for a fixed operating mixture.

Proposition B.2 (Strict-slack shadow operation)**.** *Let $x \in \Delta(\mathcal{A})$ be any fixed mixture such that for some $\gamma \in (0, \rho_{\min}]$,*

$$\sum_{a\in\mathcal{A}} x(a)\big(\bar{c}_j(a) - \bar{g}_j(a)\big) \leq \rho_j - \gamma, \qquad \forall j \in [d]. \tag{21}$$

If the shadow process repeatedly samples from x, then with probability at least $1-\delta$,

$$\min_{j\in[d],\, t\in[T]} \widetilde{B}_{j,t} \geq \gamma T - \sqrt{8T\log\left(\frac{2dT}{\delta}\right)}.$$

Hence the buffered condition in Corollary B.1 holds whenever

$$\gamma T \geq 1 + \bar{d} + \sqrt{8T\log\left(\frac{2dT}{\delta}\right)}.$$

Proof. Fix a resource $j \in [d]$ and define the net-consumption random variables

$$X_{s,j} := C_{j,s}(a_s) - G_{j,s}(a_s), \qquad \mu_j := \mathbb{E}[X_{s,j}].$$

Under the fixed mixture x, condition (21) gives $\mu_j \le \rho_j - \gamma$. For the shadow process,

$$\begin{aligned}
\widetilde{B}_{j,t} &= B_j - \sum_{s=1}^{t-1} X_{s,j} \\
&= \rho_j T - (t-1)\mu_j - \sum_{s=1}^{t-1}(X_{s,j} - \mu_j) \\
&\ge \rho_j T - (t-1)(\rho_j - \gamma) - \sum_{s=1}^{t-1}(X_{s,j} - \mu_j) \\
&= \rho_j (T - t + 1) + \gamma(t-1) - \sum_{s=1}^{t-1}(X_{s,j} - \mu_j) \\
&\ge \gamma T - \sum_{s=1}^{t-1}(X_{s,j} - \mu_j),
\end{aligned}$$

where the last inequality uses $\gamma \le \rho_j$. The centered increments lie in $[-2, 2]$. Hence, for any fixed j, t, Hoeffding's inequality yields

$$\Pr\left(\sum_{s=1}^{t-1}(X_{s,j} - \mu_j) > u\right) \le \exp\left(-\frac{u^2}{8T}\right).$$

Taking

$$u := \sqrt{8T \log\left(\frac{2dT}{\delta}\right)}$$

and applying a union bound over all dT pairs (j, t), we obtain with probability at least $1 - \delta$ that

$$\sum_{s=1}^{t-1}(X_{s,j} - \mu_j) \le u, \qquad \forall j, t.$$

On this event, $\widetilde{B}_{j,t} \ge \gamma T - u$ for every j, t. The stated sufficient condition then implies $\widetilde{B}_{j,t} \ge 1 + \bar{d}$ uniformly, completing the proof. □

C Related Works

Stochastic multi-armed bandits. Stochastic multi-armed bandits are a foundational model for sequential decision-making under uncertainty (Lattimore & Szepesvári, 2020; Slivkins, 2019). Classical algorithms such as UCB and Thompson sampling achieve logarithmic instance-dependent regret under standard gap conditions, while adversarial algorithms such as EXP3 provide worst-case guarantees when reward sequences are not stochastic (Bubeck & Cesa-Bianchi, 2012). BwDRK remains a bandit problem in the usual sense: in each productive round the learner selects one arm and observes only the outcome of that arm. The novelty is not the feedback model alone, but the coupling between bandit learning and delayed hard feasibility.

Bandits with knapsacks and replenishable resources. The BwK framework was introduced to model bandit learning with limited resources and LP-based benchmarks (Badanidiyuru et al., 2018). Subsequent work refined stochastic BwK through primal-dual methods, logarithmic instance-dependent analyses, and extensions to contextual or combinatorial settings (Li et al., 2021; Flajolet & Jaillet, 2015; Sankararaman & Slivkins, 2021; Agrawal & Devanur, 2014; Agrawal et al., 2016). Adversarial BwK requires different performance criteria because regret minimization against strong benchmarks is generally impossible (Immorlica et al., 2022; Kesselheim & Singla, 2020). Replenishable or non-monotone variants allow resources to increase as well as decrease (Kumar & Kleinberg,

2022; Bernasconi et al., 2023; Yang et al., 2024). These models capture renewal but typically credit replenishment immediately. BwDRK differs by inserting a stochastic lead time between the action that generates replenishment and the time at which that replenishment becomes available for future feasibility.

Online learning with delayed feedback. Delayed online learning studies how to learn when the feedback generated at a decision time is observed later, possibly out of order. Time-stamped delayed-feedback models show that multiple feedback items may arrive in the same round and that tags are needed to match each arrival to its originating decision (Joulani et al., 2013). Delayed bandit algorithms further show how virtual-time arguments and biased estimators control the regret impact of unknown delays (Li et al., 2019; Thune et al., 2019; Bistritz et al., 2019; Zimmert & Seldin, 2020). Other work considers stochastic arm-dependent delays, censored conversions, anonymous or composite feedback, and best-of-both-worlds delayed bandits (Manegueu et al., 2020; Pike-Burke et al., 2018; Wang et al., 2021; Lancewicki et al., 2021; Van Der Hoeven & Cesa-Bianchi, 2022; Esposito et al., 2023; Masoudian et al., 2022; Schlisselberg et al., 2025). These results mainly treat delay as an informational obstacle. In BwDRK, delayed replenishment also changes the feasible action set by locking resources in transit.

Stochastic networks, inventory lead times, and Lyapunov control. The funds-in-flight viewpoint is also related to stochastic network optimization and drift-plus-penalty methods, where queue backlogs are used as dynamic prices to trade reward against stability (Neely, 2010). Inventory theory with lead times studies how base-stock or safety-stock policies protect against delayed replenishment and demand uncertainty (Zipkin, 2000). BwDRK differs in both objective and information structure: the learner receives bandit feedback about arm outcomes, competes with LP-type regret benchmarks, and must attribute each delayed replenishment to its originating action. Thus, the queue is not only a controlled state variable but also a tagged feedback mechanism for learning.

Delayed budgets and constrained dynamics. Delayed conversions and delayed linear bandit feedback have been studied in models where delay primarily affects observation (Vernade et al., 2017; 2020). Other recent works combine delays with budgeted recommendation, deadline constraints, auctions, or online convex optimization (Sawwan & Wu, 2024; Liu et al., 2024; Fikioris et al., 2024; Qiu et al., 2025). These settings are close in motivation, but they do not isolate the specific BwK-type phenomenon studied here: an action consumes hard budget immediately, creates endogenous replenishment, and makes the replenishment available only after a stochastic delay. The queue and shadow-process analysis in BwDRK is designed precisely to quantify this funds-in-flight effect.

D Additional Application Scenarios

BwDRK should be read as a first-order abstraction of tagged delayed replenishment. The examples below illustrate where the abstraction is useful; they do not claim to cover anonymous attribution, censored feedback, heavy-tailed delays, or capacity clipping without additional modeling.

Inventory management with lead times. Consider a retailer managing inventories B_j for commodities $j \in [d]$. At round t, the retailer selects a sales action a_t, collects revenue $R_t(a_t)$, and incurs inventory consumption $C_{j,t}(a_t)$. Replenishment orders generate inventory inflow $G_{j,t}(a_t)$ but become available only after a stochastic delivery lag $E_t(a_t)$. When order identities and arrivals are observable, the tagged queue distinguishes frictional stockouts (inventory is in transit) from structural stockouts (no future replenishment remains). Systems with supplier congestion, order cancellation, or unknown heavy-tailed lead times require extensions beyond the present model.

Real-time bidding with delayed conversion revenue. In RTB, a demand-side platform spends budget immediately when it wins impressions, while conversion revenue or budget recycling may be realized later (Vernade et al., 2017; Chapelle, 2014). In the simplified tagged abstraction, advertiser budgets are resources, impression value is the immediate reward, bid spend is immediate consumption, and attributed conversion revenue is delayed replenishment. A delay-unaware bidding policy can satisfy an average return-on-spend relation and still stop bidding prematurely because too much revenue is still in flight. In practice, conversions may be anonymous, censored, or delayed together with reward; these complications are outside the current theory.

Renewable-energy dispatch with storage limits. In a virtual power plant, the controller chooses dispatch actions that consume stored energy now and replenish storage later through harvesting or market purchases. Delays arise because replenishment is credited after physical transfer or charging time. If storage has a hard capacity ceiling, delayed replenishment may also be clipped when it arrives to a saturated battery. With the same arrival-first convention as in the main model, this can be represented by

$$B_{j,t}^{+} = \min\left\{ B_{j,t} + \sum_{s:\, s+E_s(a_s)+1=t} G_{j,s}(a_s),\, B_{j,\max} \right\}, \qquad B_{j,t+1} = B_{j,t}^{+} - C_{j,t}(a_t).$$

Only replenishments that actually arrive at the beginning of round t are credited. The clipping term introduces an additional overflow trade-off not analyzed in the main theorem.