some stuff

Arshia and Mirzadi

July 24, 2025

Abstract

Your abstract.

1 Introduction

we build a bandit-style portfolio manager that only learns when we actually trade a stock: each stock gets a risk-adjusted score q_i (regularized squared sharpe), and its ucb exploration bonus shrinks purely with its own trade count. an "anytime" average risk budget controls how often we trade—if a trade would push risk too high, we skip. this ties exploration directly to trading activity while keeping portfolio-level risk in check, and can extend to trading small bundles of stocks using the same scoring and budget rule.

2 Market, Actions, Observability

Notation.

- $K \in N$: number of stocks (arms).
- Horizon $T \in N$; rounds $t = 1, \ldots, T$.
- Decision $A_t \in \{1, \ldots, K\}$ or *skip* (null action).
- Realized return when trading i at t: $R_{i,t} \in [\ell, u]$, i.i.d. across t with unknown mean μ_i and variance σ_i^2.

Observability / Exploration. Although full price paths may be observable, the *learning sample size* updates only when a trade is executed:

$$N_i(t) = \sum_{s \le t} \mathbf{1}\{A_s = i\}.$$

All estimators and exploration bonuses depend solely on $N_i(t)$ (no exploration on passive observability).

3 Risk / Budget Structure (Anytime Cost Constraint)

Each trade of stock i incurs deterministic cost $c_i \ge 0$. Define cumulative and average costs:

$$C(t) = \sum_{s \le t} c_{A_s}, \qquad \bar{C}(t) = \frac{C(t)}{t}.$$

Fix budget level $c > 0$. The anytime constraint requires

$$\bar{C}(t) \le c \quad \forall t.$$

Under-utilization Target. Let $\omega(t)$ denote a lower confidence bound on the minimum cost gap (problem-dependent). Define a shrinking target average cost

$$c_t = c - \frac{\log t}{\omega^2(t)\, t}.$$

Skipping (null action) is permitted to enforce $\bar{C}(t) \le c_t$.

Multi-resource Extension. For m resource types, let $\mathbf{c}_i \in R^m$ and maintain $\mathbf{C}(t) = \sum_{s \le t} \mathbf{c}_{A_s}$; impose componentwise anytime constraints. Negative components allow replenishment.

4 Risk-Adjusted Value

For stock i, define the *regularized squared Sharpe ratio* (RSSR)

$$q_i = \frac{\mu_i^2}{L + \sigma_i^2}, \qquad L > 0.$$

Empirical statistics over $N_i(t)$ trades:

$$\bar{X}_i(t) = \frac{1}{N_i(t)} \sum_{s:A_s=i} R_{i,s}, \quad \bar{V}_i(t) = \frac{1}{N_i(t)} \sum_{s:A_s=i} \big(R_{i,s} - \bar{X}_i(t)\big)^2, \quad \tilde{X}_i(t) = \frac{1}{N_i(t)} \sum_{s:A_s=i} R_{i,s}^2.$$

Confidence radius:

$$\epsilon_i(t) = \sqrt{\frac{2\log t}{N_i(t)}}.$$

Empirical RSSR:

$$\widehat{q}_i(t) = \frac{\bar{X}_i^2(t)}{L + \bar{V}_i(t)}.$$

Define the UCB-RSSR index

$$B_i(t) = \frac{\bar{X}_i^2(t)}{L + \bar{V}_i(t)} + \frac{\big(\bar{V}_i(t) + \tilde{X}_i(t) + 2\epsilon_i(t) + L\big)\epsilon_i(t)}{\big(L + \bar{V}_i(t)\big)\big(L + \bar{V}_i(t) - 3\epsilon_i(t)\big)}. \tag{1}$$

A deviation bound (problem constants suppressed) has the form

$$\Pr(|\widehat{q}_i(t) - q_i| > \hat{\epsilon}_i(t)) \le 2\exp\left(-\frac{2N_i(t)\epsilon_i^2(t)}{u^2}\right),$$

with

$$\hat{\epsilon}_i(t) = \frac{\big(\bar{V}_i(t) + \tilde{X}_i(t) + 2\epsilon_i(t) + L\big)\epsilon_i(t)}{\big(L + \bar{V}_i(t)\big)\big(L + \bar{V}_i(t) - 3\epsilon_i(t)\big)}.$$

5 Single-Stock Trading Algorithm (Anytime-Cost UCB-RSSR)

Initialization. Trade each stock δ times (round-robin), updating statistics.

[h] Anytime-Cost UCB-RSSR [1] **Input:** Budget c, regularization L, initialization length δ, horizon T. **Init:** For $t = 1$ to δK, set $A_t = ((t-1) \bmod K) + 1$; update N_i, $\bar{X}_i$, $\bar{V}_i$, $\tilde{X}_i$, B_i. $t = \delta K + 1$ to T Compute $c_t = c - \frac{\log t}{\omega^2(t)t}$. $C(t-1) + \min_i c_i > c_t\, t$ $A_t \leftarrow$ skip (null action); continue. Select $i_t = \arg\max_i B_i(t-1)$. Execute trade i_t: observe $R_{i_t,t}$. Update N_{i_t}, $\bar{X}_{i_t}$, $\bar{V}_{i_t}$, $\tilde{X}_{i_t}$. Recompute $\epsilon_{i_t}(t)$ and $B_{i_t}(t)$ via (1). Update cost $C(t) = C(t-1) + c_{i_t}$.

6 Multi-Stock (Base) Trades

Let $\mathcal{B}$ denote a collection of bases $b \subseteq \{1, \dots, K\}$ with weight vectors $w_i \geq 0$ ($i \in b$). Portfolio empirical RSSR:

$$\widehat{q}_b(t) = \frac{\Big(\sum_{i \in b} w_i \bar{X}_i(t)\Big)^2}{L + \sum_{i \in b} w_i^2 \bar{V}_i(t)}.$$

Conservative bonus:

$$\mathrm{CB}_b(t) = \sum_{i \in b} w_i \big(B_i(t) - \widehat{q}_i(t)\big),$$

Index:

$$B_b(t) = \widehat{q}_b(t) + \mathrm{CB}_b(t).$$

At round t choose

$$b_t = \arg\max_{b \in \mathcal{B}} B_b(t-1)$$

subject to cost/budget checks and minimum inclusion probability $\omega_{\min}$ (each arm in a base must appear with probability $\geq \omega_{\min}$). Statistics of arms in b_t are updated as in the single-stock case.

7 Regret / Guarantees

Let $i^\star = \arg\max_i q_i$ and $\Delta_i = q_{i^\star} - q_i$. Under standard regularity and gap assumptions:

- UCB-RSSR (single resource): $\sum_{t=1}^{T} (q_{i^\star} - q_{A_t}) = O\Big(\sum_{i:\Delta_i > 0} \frac{\log T}{\Delta_i}\Big)$.
- Anytime cost with under-utilization: additional $O(K \log T)$ term; number of skips sublinear.
- Multi-resource / replenishable (block scheduling or primal–dual): analogous logarithmic or $\tilde{O}(\sqrt{KT})$ bounds depending on sign structure of costs.

8 Implementation Parameters

- Regularization $L > 0$ controls stability of RSSR.
- Initialization length δ ensures sufficient samples to estimate variances.
- Under-utilization schedule $c_t = c - \frac{\log t}{\omega^2(t) t}$.

9 Exploration Principle

Exploration bonus depends only on trade counts:

$$\epsilon_i(t) = \sqrt{\frac{2 \log t}{N_i(t)}}.$$

Passive observability does not alter $N_i(t)$; thus no exploration pressure from merely observing prices.

References