Stochastic Multi-Armed Bandits with Non-Stationary Rewards
Generated by a Linear Dynamical System

Jonathan Gornet, Mehdi Hosseinzadeh, Member, IEEE, and Bruno Sinopoli, Fellow, IEEE

Abstract—The stochastic multi-armed bandit has provided a framework for studying decision-making in unknown environments. We propose a variant of the stochastic multi-armed bandit where the rewards are sampled from a stochastic linear dynamical system. The proposed strategy for this variant is to learn a model of the dynamical system while choosing the optimal action based on the learned model. Motivated by mathematical finance areas such as Intertemporal capital asset pricing model proposed by Merton and Stochastic portfolio theory proposed by Fernholz that both model asset returns with stochastic differential equations, this strategy is applied to quantitative finance as a high-frequency trading strategy, where the goal is to maximize returns within a time period.

I. INTRODUCTION

The stochastic Multi-Armed Bandit (MAB) problem has provided a powerful modeling framework to investigate large class of decision making problems. In MAB, a learner interacts with the environment where in each interaction, a called a round, the learner chooses an action and receives a reward. The performance of policies is usually evaluated as the expectation of the cumulative difference between optimal and chosen actions, which is defined as the regret. In its basic formulation where rewards are sampled from a stationary distribution, a very popular algorithm, called the Upper Confidence Bound (UCB) algorithm [1], guarantees logarithmic growth of regret [2].

MAB has seen different applications in several areas such as machine learning, dynamic pricing, and portfolio management. In machine learning, the MAB formulation has been used to find a set of hyperparameters to increase the performance of the learning process [3]. This has been extended to algorithm selection problem, where a learner searches for a high-performing algorithm to use for training [4]. In dynamic pricing problems, MAB has been used to set prices when the model for demand is based on a low-dimensional demand model [5]. In portfolio selection, MAB formulation is natural in the case where the manager creates a portfolio with multiple assets [6], [7].

In the examples mentioned above, the environments that the learner interacts with has time correlations and correlations between the rewards for each action. Hyperparameter optimization can be viewed as optimizing a cost function with correlated decision variables [8] which implies that the training process is dynamic. In dynamic pricing problem, the demand changes over time [9]. As for portfolio management, Intertemporal capital asset pricing model [10] and Stochastic portfolio theory [11] model asset returns using stochastic differential equations. Therefore, there is a need to investigate non-stationary stochastic MAB where the rewards are sampled from a stochastic dynamical system. In this case, the reward can be expressed as the inner product between the action vector and a dynamic unknown parameter vector.

Previous work in non-stationary stochastic MAB studies the case where the change in magnitude of the unknown parameter vector is bounded [12], [13] or when such vector is sampled from a predefined set [14], [15]. In both cases, the unknown parameter vector changes the distribution of the reward, making the original stochastic MAB formulation non-stationary. To the best of our knowledge, most of the previous studies on non-stationary stochastic MAB focus on the piece-wise stationary case, where the distributions are stationary within set intervals. The studies on the piece-wise stationary case focus on remembering-vs-forgetting trade-off using either discounting or sliding windows [16] to forget early recorded rewards or detecting the change in the distribution to decide when to restart the learning process [17], [18], [19], [20]. A set of previous publications bound the cumulative change of the reward mean in the form of variational budget [21], [22]. In [23], the agent selects a controller each round to control a time-varying linear dynamical system where the reward is a cost function of the system’s state and inputted control signal. This system is driven by a bounded perturbation and the state is observed.

This paper tackles the stochastic MAB problem when the rewards are sampled from an unknown stochastic linear dynamical system driven by Gaussian noises. Since the rewards are now dynamic, taking inspiration from [24], we introduce a methodology that focuses on finding the optimal decision while learning the model of the system generating the rewards. The learned model is then used to design a decision-making algorithm. We verify the proposed algorithm on a simple high-frequency trading example.

This paper is organized as follows. Section II states the problem. In Section III, a methodology to model and predict rewards is presented. Section IV uses the model to develop a strategy to maximize cumulative reward over a horizon. Section V performs regret analysis and provides a theoretical upper bound for the regret of the proposed algorithm. Section VI shows an application of the proposed algorithm to a simple high-frequency trading example. Finally, Section VII
concludes the paper and provides future directions.

II. PROBLEM STATEMENT

Suppose that for \( k \) given actions \( c_a \in \mathcal{A} \subset \mathbb{R}^d \), the reward \( X_t \) is sampled from the following stochastic linear dynamical system

\[
\begin{align*}
    z_{t+1} & = \Gamma z_t + \xi_t, \quad z_0 \sim \mathcal{N}(\mu_0, \Sigma_0) \\
    \theta_t & = C_a z_t + \phi_t \\
    X_t & = \langle c_a, z_t \rangle + \eta_t,
\end{align*}
\]  

(1)

where \( z_t \in \mathbb{R}^d \) is the state of the system. For each round \( t = 1, 2, \ldots, n \) and \( n > d \), the learner observes the reward \( X_t \in \mathbb{R} \) based on the chosen action and the context \( \theta_t \in \mathbb{R}^m \). The context \( \theta_t \) is a value that the learner always observes and its observation matrix \( C_\theta \) is constant. The noises \( \xi_t \in \mathbb{R}^d \), \( \phi_t \in \mathbb{R}^m \), and \( \eta_t \in \mathbb{R} \) are i.i.d. normally distributed, i.e., \( \xi_t \sim \mathcal{N}(0, \Sigma_\xi) \) with \( \Sigma_\xi > 0 \), \( \phi_t \sim \mathcal{N}(0, \Sigma_\phi) \), and \( \eta_t \sim \mathcal{N}(0, \Sigma_\eta) \). We pose the following assumptions on system (1).

**Assumption 1:** The matrices \( \Gamma, C_\theta, Q, R_\theta, \Sigma_\xi, \Sigma_\phi, \Sigma_\eta \), vectors \( \mu_\xi, \mu_\phi, \mu_\eta \), and scalar \( \sigma_\eta \) are unknown. The dimension \( d \) is unknown, but dimension \( m \) is known as it is the dimension of the context. Also, the vector \( c_a \) is unknown. Note that for notation, given that there are \( k \) vectors \( c_a \in \mathcal{A} \), we denote \( a \in \{1, 2, \ldots, k\} \) to be which vector \( c_a \) is chosen.

**Assumption 2:** The matrix \( \Gamma \) is Schur. Also, the pair \((\Gamma, C_\theta)\) is observable.

The goal of the learner is to maximize cumulative reward over a finite time horizon \( n \). To analyze the performance of the learner’s strategy, regret analysis is used. The regret is defined as the cumulative, over all rounds, expected difference between the highest reward (denoted as \( X^* \)) and the reward for the chosen action at time \( t \), i.e.,

\[
R_n = \sum_{t=1}^n \mathbb{E}[X^*_t - X_t].
\]  

(2)

III. LEARNING THE SYSTEM

If the learner knew (1), then the Kalman filter could be used to predict the state \( z_t \) and consequently the reward \( X_t \) for each action \( a \in \{1, 2, \ldots, k\} \).

\[
\begin{align*}
    z_{t+1|t} & = \Gamma z_{t|t} + \mu_\xi \\
    P_{t+1|t} & = \Gamma P_{t|t} \Gamma^\top + Q \\
    K_t & = P_{t|t-1} C_\theta^\top (C_\theta P_{t|t-1} C_\theta + R_\theta)^{-1} \\
    \hat{z}_{t|t} & = \hat{z}_{t|t-1} + K_t (\hat{\theta}_t - C_\theta \hat{z}_{t|t-1}) \\
    P_{t|t} & = P_{t|t-1} - K_t C_\theta P_{t|t-1} \\
    \hat{X}_{t|t-1} & = \langle c_a, \hat{z}_{t|t-1} \rangle
\end{align*}
\]

(3)

where \( \hat{z}_{t|t} \equiv \mathbb{E}[z_t|\mathcal{F}_t] \) and \( \mathcal{F}_t \) is the sigma algebra generated by previous contexts \( \theta_0, \ldots, \theta_t \). Since the Kalman gain converges thanks to assumption 2, it is reasonable [24] to use the steady-state Kalman filter:

\[
\begin{align*}
    \hat{z}_{t+1} & = \Gamma \hat{z}_t + \mu_\xi + \Gamma K (\hat{\theta}_t - C_\theta \hat{z}_t) \\
    \hat{X}_t & = \langle c_a, \hat{z}_t \rangle
\end{align*}
\]

(4)

where

\[
\begin{align*}
    K & \equiv PC_\theta^\top (C_\theta PC_\theta^\top + R_\theta)^{-1} \\
    P & \equiv \Gamma P \Gamma^\top + Q - \Gamma P C_\theta^\top (C_\theta PC_\theta^\top + R_\theta)^{-1} C_\theta P \Gamma^\top, \\
    \hat{z}_t & \equiv \hat{z}_{t|t-1}, \quad \hat{X}_t \equiv \hat{X}_{t|t-1}.
\end{align*}
\]

Since using the steady-state Kalman filter prediction \( \hat{X}_t \) provides a good prediction of the reward \( X_t \), then learning the steady-state Kalman filter will intuitively provide a good prediction of the reward \( X_t \) for each action \( a \in \{1, 2, \ldots, k\} \). Therefore, to learn the steady-state Kalman filter, a variation of the method presented in [24] is used. Let \( s > 0 \) be the horizon length of how far we look into the past. We define a matrix \( G_a \) for each \( a \in \{1, 2, \ldots, k\} \) and a vector \( \Theta_a \), as follows:

\[
G_a \equiv \left[ \langle c_a, (\Gamma - \Gamma K C_\theta)^s \hat{z}_{t-s} \rangle + \varepsilon_{ax} \right],
\]

(5)

where \( \varepsilon_{ax} \equiv X_t - \hat{X}_t \) is defined above, it can be shown that the reward \( X_t \) can be expressed as

\[
X_t = G_a \Theta_t + \langle c_a, (\Gamma - \Gamma K C_\theta)^s \hat{z}_{t-s} \rangle + \varepsilon_{ax},
\]

(6)

The equation (6) is set such that if a set of contexts \( \theta_{t-s}, \ldots, \theta_{t-1} \) are provided, their linear combination using the constants in the vector \( G_a \) provide a prediction for \( X_t \) where the accuracy is impacted by the terms \( \langle c_a, (\Gamma - \Gamma K C_\theta)^s \hat{z}_{t-s} \rangle \) and \( \varepsilon_{ax} \). Note that since \( \Gamma - \Gamma K C_\theta \) is Schur by construction, then the magnitude of the term \( \langle c_a, (\Gamma - \Gamma K C_\theta)^s \hat{z}_{t-s} \rangle \) decreases as \( s \) increases. Given a set of time instants \( \mathcal{F}_a = \{t_1, t_{N_a}\} \), the reward \( X_t \) can be rewritten as

\[
\begin{bmatrix}
    X^\top_{t_1} \\
    \vdots \\
    X^\top_{t_{N_a}}
\end{bmatrix} = G_a \begin{bmatrix}
    \Theta^\top_{t_1} \\
    \vdots \\
    \Theta^\top_{t_{N_a}}
\end{bmatrix} + \begin{bmatrix}
    \langle c_a, (\Gamma - \Gamma K C_\theta)^s \hat{z}_{t_1-s} \rangle + \varepsilon_{ax_{t_1}} \\
    \vdots \\
    \langle c_a, (\Gamma - \Gamma K C_\theta)^s \hat{z}_{t_{N_a-s}} \rangle + \varepsilon_{ax_{t_{N_a}}}
\end{bmatrix},
\]

(8)

A regularized least squares estimate for (8) is

\[
\hat{G}_a = \sum_{t \in \mathcal{F}_a} X_t \Theta_t \Theta_t^\top V_a^{-1},
\]

(9)

where \( V_a > 0 \) is given by

\[
V_a = \lambda I + \sum_{t \in \mathcal{F}_a} \Theta_t \Theta_t^\top,
\]

(10)

with \( \lambda > 0 \) as the regularization parameter and \( I \) as the identity matrix with appropriate dimension.
Algorithm 1 Phased Initial Exploration of System (PIES)

\[ t \leftarrow 1 \]
\[ \text{for } a \in \{1, 2, \ldots, k\} \text{ do} \]
\[ \hat{G}_a \leftarrow 0_{1 \times m s + 1} \]
\[ \mathcal{F}_a \leftarrow \{} \]
\[ \text{end for} \]
\[ \text{for } t = 1, 2, \ldots, n \text{ do} \]
\[ \text{if } t > ks \text{ then} \]
\[ a \leftarrow \arg \max_{a \in \{1, 2, \ldots, k\}} \hat{G}_a \Theta_t \]
\[ \text{else} \]
\[ a \leftarrow (t \mod k) + 1 \]
\[ \text{end if} \]
\[ \text{if } t > s \text{ then} \]
\[ \text{for } a \in \{1, 2, \ldots, k\} \text{ do} \]
\[ \text{Update } \mathcal{F}_a \text{ and learn } \hat{G}_a \text{ based on (9) and (10)} \]
\[ \text{end for} \]
\[ \text{end if} \]
\[ \text{Sample } (\theta_t, X_t) \text{ based on (1)} \]
\[ \text{end for} \]

IV. Bandit Strategy

The action the learner ought to choose is the action that the learner predicts will output the highest reward. This prediction is based on the matrix \( \hat{G}_a \) in (9), which is an estimate of the matrix \( G_a \). Therefore, the learner should focus on learning \( \hat{G}_a \) for each action \( a \in \{1, 2, \ldots, k\} \) and choosing an action based on the learned \( \hat{G}_a \).

Inspired by the Explore-then-Commit (ETC) method discussed in [25], we propose the Phased Initial Exploration of System (PIES) which is shown in Algorithm 1. The parameter to set is \( s \). At the start of the algorithm, the learner will cycle through each action \( a \in \{1, 2, \ldots, k\} \) from round \( t = 1 \) to round \( t = ks \). The learner will start computing \( \hat{G}_a \) once \( t > s \). After round \( t = ks \), the learner will choose the action \( a \in \{1, 2, \ldots, k\} \) that has the largest \( \hat{G}_a \Theta_t \) value while keep updating \( \hat{G}_a \). The matrix \( \hat{G}_a \) is constantly updated after \( t = ks \) to minimize identification error \( \|G_a - \hat{G}_a\|_2 \).

V. Regret Analysis of the Proposed Algorithm

The following theorem provides an upper bound for regret defined in (2).

**Theorem 1:** Given a failure rate of \( \delta \in (0, 1) \), regret as in (2) is bounded as follows with a probability of at least \( 1 - \delta \):

\[
R_n \leq \sum_{t=1}^{n} \sum_{a \neq a^*} \mathbb{E}[(\Delta c_a, z_t)] + \left( \sum_{t=ks+1}^{n} \mathbb{E}[(\Delta c_a, z_t)] a \right) \min \left\{ B_a \mathbb{E}[\Theta_{t}^\top \frac{\|\Theta_{t}\|_2}{\|\Delta G_a \Theta_{t}\|_2}], 1 \right\}. \tag{11}
\]

where \( \Delta c_a \) and \( \Delta G_a \) are defined as

\[
\Delta c_a \triangleq c_a - c_{a^*}, \tag{12}
\]
\[
\Delta G_a \triangleq G_a^* - G_a, \tag{13}
\]

and \( B_a \) is such that the following inequality holds with a probability of at least \( 1 - \delta \):

\[
\|\hat{G}_a - G_a\|_2 + \|G_a^* - \hat{G}_a^*\|_2 \leq B_a. \tag{14}
\]

**Proof:** Using the law of iterated expectations [26], the instantaneous regret for one round \( t \in \{ks+1, \ldots, n\} \) is

\[
E[X_t^s - X_t] = E[(\Delta c_a, z_t)] = E_a[E_c[(\Delta c_a, z_t)|a]] = \sum_{a=1}^{k} E_c[(\Delta c_a, z_t)|a] P[a], \tag{15}
\]

where \( \Delta c_a \) is defined in (12). In the following, we will provide an upper bound for \( P[a] \). Consider the event \( \mathcal{E}_a \) defined as

\[
\mathcal{E}_a \triangleq \{ \hat{G}_a \Theta_t \geq G_a \Theta_t \}, \tag{16}
\]

which implies that modeling error leads to selecting an suboptimal action. Note that \( P[a] \) is

\[
P[a] = P\left[ \bigcap_{i \neq a} \{ \hat{G}_a \Theta_t \geq G_a \Theta_t \} \right] = P\left[ \bigcap_{i \neq a, a^*} \{ \hat{G}_a \Theta_t \geq G_a \Theta_t \} \cap \{ G_a \Theta_t \geq G_a^* \Theta_t \} \right], \tag{17}
\]

which implies that that \( P[a] \leq P[\mathcal{E}_a] \). Adjusting the inequality in \( \mathcal{E}_a \) yields

\[
G_a \Theta_t + (\hat{G}_a - G_a) \Theta_t \geq G_a \Theta_t - (G_a - G_a^*) \Theta_t, \tag{18}
\]

Let \( \Delta G_a \) be as in (13). Then (18) can be rewritten as follows:

\[
\Delta G_a \Theta_t \leq (\hat{G}_a - G_a) \Theta_t + (G_a^* - \hat{G}_a^*) \Theta_t, \tag{19}
\]

At this stage, we use (9) and (10) to express \( \hat{G}_a \) for any action as follows:

\[
\hat{G}_a = \sum_{t \in \mathcal{F}_a} G_a \Theta_t \Theta_t^\top V_a^{-1} + \sum_{t \in \mathcal{F}_a} \varepsilon_a \Theta_t \Theta_t^\top V_a^{-1} + \sum_{t \in \mathcal{F}_a} (c_{a^*}(\Gamma - \Gamma K C_{\theta})^2 \xi_{t-s}) \Theta_t \Theta_t^\top V_a^{-1}.
\]

Thus, \( \|\hat{G}_a - G_a\|_2 \) can be upper-bounded as

\[
\|\hat{G}_a - G_a\|_2 \leq \|\Lambda G_a V_a^{-1}\|_2 + \sum_{t \in \mathcal{F}_a} \varepsilon_a |\Theta_t \Theta_t^\top V_a^{-1}|_2 \leq \sum_{t \in \mathcal{F}_a} (c_{a^*}(\Gamma - \Gamma K C_{\theta})^2 \xi_{t-s}) |\Theta_t \Theta_t^\top V_a^{-1}|_2. \tag{21}
\]

Since \( \Gamma \) is Schur, the estimate \( (c_{a^*}(\Gamma - \Gamma K C_{\theta})^2 \xi_{t-s}) \) is bounded. For the product \( \varepsilon_a \Theta_t \Theta_t^\top V_a^{-1} \), based on Theorem 1 in [27], since \( \varepsilon_a \) is conditionally \( c_{a^*} P c_a + c_{\theta} \)-sub-Gaussian and \( \Theta_t \Theta_t^\top \) is measurable, then given a failure rate \( \delta \in (0, 1) \), this term is bounded with a probability of at least \( 1 - \delta \). Therefore, (14) is satisfied with a probability of at least \( 1 - \delta \).
Now, assuming that $E_1$ is given, the inequality (19) can be rewritten as
\[ B_a \| \Theta_t \|_2 \geq \Delta G_a \Theta_t \Rightarrow B_a \| \Theta_t \|_2 \geq 1. \] (22)

Let (22) be denoted as $\mathcal{E}_2|\mathcal{E}_1$. Based on the Markov inequality [28], the following concentration bound is given
\[ \mathbb{P}[\mathcal{E}_2|\mathcal{E}_1] \leq \min \left\{ B_a \mathbb{E}_{\Theta} \left[ \frac{1}{\| \Theta_t \|_2} \right], 1 \right\}. \] (23)

Also, (16) and (22) can be rewritten, respectively, as
\[ \mathcal{E}_a = \{ \hat{G}_a - G_a \Theta_t + (G_a^* - \hat{G}_a^*) \Theta_t \geq 1 \}, \] (24)
and
\[ \mathcal{E}_2|\mathcal{E}_1 = \{ B_a \| \Theta_t \|_2 \geq 1 \}. \] (25)

Note that $\mathbb{P}[\mathcal{E}_a] \leq \mathbb{P}[\mathcal{E}_2|\mathcal{E}_1]$ is true if the following inequality holds:
\[ B_a \| \Theta_t \|_2 \geq \frac{(\hat{G}_a - G_a) \Theta_t + (G_a^* - \hat{G}_a^*) \Theta_t}{\Delta G_a \Theta_t}. \] (26)

Since (26) is satisfied with a probability of at least $1 - \delta$ according to (14), then $\mathbb{P}[\mathcal{E}_a] \leq \mathbb{P}[\mathcal{E}_2|\mathcal{E}_1]$ is satisfied with a probability of at least $1 - \delta$. Therefore, (15) can be upper-bounded as
\[
\mathbb{E}[X_t^* - X_t] = \sum_{a=1}^{k} \mathbb{E}_c[|\Delta a, \xi_t|] a \mathbb{P}[a] \leq \sum_{a\neq a^*} \frac{\mathbb{E}_c[|\Delta a, \xi_t|] a \mathbb{P}[\mathcal{E}_a]}{\mathbb{P}[\mathcal{E}_a]} \leq \sum_{a\neq a^*} \frac{\mathbb{E}_c[|\Delta a, \xi_t|] a \mathbb{P}[\mathcal{E}_2|\mathcal{E}_1]}{\mathbb{P}[\mathcal{E}_2|\mathcal{E}_1]}. \] (27)

Finally, using (23) in (27) and summing over $t$ rounds completes the proof.

Remark 1: As shown in Theorem 1, the regret performance is based on the bound for model error $B_a$. In particular, if the model is known (i.e., $B_a = 0$), the upper-bound will be zero after $t = ks$ rounds, which is reasonable.

Remark 2: The model error $B_a$ depends on the number of times action $a$ is chosen. For now, the exploration period is set to $ks$ so that the learner has $s$ samples for each action $a \in \{1, 2, \ldots, k\}$. Future work will focus on what is a more effective length for the exploration period.

VI. EVALUATION STUDY—HIGH-FREQUENCY TRADING

This section will exemplify the use of the proposed framework. Suppose there are two stocks a trader is interested in. The trader can either buy then sell either stock 1 or 2, or refrain from trading at each round.

The context $\Theta$ represents the price change for the stocks. The reward $X_t$ is the financial gain (loss) deriving from trading a stock. Both variables are sampled from the following stochastic linear dynamical system (see Appendix):
\[ z_{t+1} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9512 & 0 \\ 0 & 0 & 0 & 0.6065 \end{bmatrix} z_t + \xi_t, \] (28)
\[ \Theta_t = \begin{bmatrix} -1 & 0 & 0.0353 & 0 \\ 0 & -1 & 0 & 0.2987 \end{bmatrix} z_t, \]
\[ X_t = \{ c_a, \xi_t \}. \]

For the model, the number of previous contexts $\Theta_t$ that are used is $s = 10$ and the regularization parameter $\lambda = 10^{-1}$. Since $\Gamma$ is Schur, one could argue that the reward $X_t$ is just Gaussian distributed as $t \gg 0$. Therefore, UCB is a reasonable method to use as a comparison. The parameter used in UCB is $\delta = 0.1$. To provide an upper bound on the algorithm’s performance we use an oracle that leverages optimal predicted awards generated by the Kalman filter $\hat{X}_t$ as in (4) to choose the action $c_a \in \mathcal{A}$.

Figure 1 shows the instantaneous regret and the cumulative regret of the learner using UCB, Algorithm 1, and the oracle averaged across 1,000 different simulations. As seen in this figure, even though the system converges to the steady-state, UCB still performs sub-par compared to Algorithm 1.

Remark 3: There are other well-known quantitative finance methods such as universal portfolio selection [29] that could be compared with Algorithm 1. These comparisons will be considered in future work.

A. Sensitivity Analysis—Impact of Parameter $s$

In this subsection, we conduct sensitivity analysis of the performance of Algorithm 1 with respect to the parameter $s$. To do so, system (28) was sampled 1,000 different times. For each sample, the regret was computed for each $s \in \{1, 2, \ldots, 10\}$. Figure 2 shows how the parameter $s$ impacts the average regret. As seen in this figure, when $s = 3$, the regret on average is the lowest, implying that this is most likely the ideal $s$ to choose for this system. As $s$ increases, the regret increases steadily. In Figure 2, it appears that $s = 1$ and $s = 8$ have approximately the same large regret values. The intuition behind this observation is that not enough previous observations are used for providing an accurate prediction for $s = 1$, while there are a lot of unknown terms to estimate which degrades performance for $s = 8$.

VII. CONCLUSION

This paper introduced a new variant of the stochastic multi-armed bandit where the rewards are sampled from an unknown stochastic linear dynamical system. To approach this problem, the learner explores all the actions to learn the underlying dynamical model and predict the action with
the highest reward, simultaneously. An application to high-frequency trading is used to illustrate the method. Future work will provide a theoretical analysis for the impact of the parameter $s$ and will focus on creating an adaptive exploration period to avoid the dependency on the horizon length $n$.

**REFERENCES**

[1] R. Agrawal, “Sample mean based index policies by o (log n) regret for the multi-armed bandit problem,” *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[3] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6765–6816, 2017.

[4] M. Gagliolo and J. Schmidhuber, “Algorithm selection as a bandit problem with unbounded losses,” in *International conference on learning and intelligent optimization*. Springer, 2010, pp. 82–96.

[5] J. W. Mueller, V. Syrgkanis, and M. Taddy, “Low-rank bandit methods for high-dimensional dynamic pricing,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] W. Shen, J. Wang, Y.-G. Jiang, and H. Hua, “Portfolio choices with orthogonal bandit learning,” in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[7] X. Huo and F. Fu, “Risk-aware multi-armed bandit problem with application to portfolio selection,” *Royal Society open science*, vol. 4, no. 11, p. 171377, 2017.

[8] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, “Taking the human out of the loop: A review of bayesian optimization,” *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2015.

[9] S. Agrawal, S. Yin, and A. Zeevi, “Dynamic pricing and learning under the Bass model,” in *Proceedings of the 22nd ACM Conference on Economics and Computation*, 2021, pp. 2–3.

[10] R. C. Merton, “An intertemporal capital asset pricing model,” *Econometrica: Journal of the Econometric Society*, pp. 867–887, 1973.

[11] R. Fernholz and B. Shy, “Stochastic portfolio theory and stock market equilibrium,” *The Journal of Finance*, vol. 37, no. 2, pp. 615–624, 1982.

[12] W. C. Cheung, D. Simchi-Levi, and R. Zhu, “Learning to optimize under non-stationarity,” in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1079–1087.

[13] A. Javanmard, “Perishability of data: dynamic pricing under varying-coefficient models,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1714–1744, 2017.

[14] Y. Qin, T. Menara, S. Oymak, S. Ching, and F. Pasqualetti, “Non-stationary representation learning in sequential linear bandits,” *IEEE Open Journal of Control Systems*, vol. 1, pp. 41–56, 2022.

[15] Y. Qin, T. Menara, S. Oymak, and F. Pasqualetti, “Representation learning for stochastic sequential linear bandits,” *IEEE Conference on Decision and Control*, 2021.

[16] A. Garivier and E. Moulines, “On upper-confidence bound policies for non-stationary bandit problems,” *arXiv preprint arXiv:1005.3415*, 2008.

[17] C. Hartland, N. Baskiotis, S. Gelly, M. Sebag, and O. Teyyaud, “Change point detection and meta-bandits for online learning in dynamic environments,” in *CIAP 2007: 9th Conference francophone sur l’apprentissage automatique*, 2007, pp. 237–250.

[18] F. Liu, J. Lee, and N. Shroff, “A change-detection based framework for piecewise-stationary multi-armed bandit problem,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[19] Y. Cao, Z. Wen, B. Kveton, and Y. Xie, “Nearly optimal adaptive procedure with change detection for piecewise-stationary bandits,” in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 418–427.

[20] J. Mellor and J. Shapiro, “Thompson sampling in switching environments with bayesian online change detection,” in *Artificial intelligence and statistics*. PMLR, 2013, pp. 442–450.

[21] L. Wei and V. Srivastava, “Nonstationary stochastic multi-armed bandits: Ucb policies and minimax regret,” *arXiv preprint arXiv:2101.08980*, 2021.

[22] O. Bbesbes, Y. Gur, and A. Zeevi, “Stochastic multi-armed bandit problem with non-stationary rewards,” *Advances in neural information processing systems*, vol. 27, pp. 199–207, 2014.

[23] P. Gradu, E. Hazan, and E. Minasian, “Adaptive regret for control of time-varying dynamics.” 2020. [Online]. Available: https://arxiv.org/abs/2007.04393

[24] T. Csiszar and G. J. Pappas, “Nearby sample analysis of stochastic system identification,” in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3648–3654.

[25] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[26] J. M. Wooldridge, *Econometric analysis of cross section and panel data*. MIT press, 2010.

[27] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári, “Improved algorithms for linear stochastic bandits,” in *Advances in Neural Information Processing Systems*. J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011.

[28] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[29] A. Gelb et al., *Applied optimal estimation*. MIT press, 1974.

[30] C. Van Loan, “Computing integrals involving the matrix exponential,” *IEEE transactions on automatic control*, vol. 23, no. 3, pp. 395–404, 1978.

**APPENDIX**

The trader models the price evolution of stock $i$ (denoted as $S_i^t$) using the following stochastic differential equation
for $\tau \in [0,T]$ based on [10].

\[
\begin{align*}
\frac{dS^i_\tau}{S^i_\tau} &= M^i_\tau d\tau + dW^i_\tau \\
\frac{dM^i_\tau}{M^i_\tau} &= \kappa^i_\tau \left( \frac{1}{2} - M^i_\tau \right) d\tau + \sigma^i_\tau dV^i_\tau,
\end{align*}
\]

(29)

where $\kappa^i_\tau$ and $\sigma^i_\tau$, $i = 1, 2$, are defined to be

\[
\begin{bmatrix}
\kappa^1_\tau \\
\kappa^2_\tau
\end{bmatrix} = \begin{bmatrix}
10^{-1} \\
1
\end{bmatrix}, \quad
\begin{bmatrix}
\sigma^1_\tau \\
\sigma^2_\tau
\end{bmatrix} = \begin{bmatrix}
10 \\
1
\end{bmatrix}.
\]

(30)

The variable $M^i_\tau$ is the drift rate of stock $i$, $\kappa^i_\tau$ is the speed of reversion (the rate $M^i_\tau$ returns to its mean), and $\sigma^i_\tau$ sets the magnitude of $dV^i_\tau$. Both $dW^i_\tau$ and $dV^i_\tau$ are independent Gaussian distributed random variables with a variance of 1 with no time correlation, i.e. $dW^i_\tau \sim \mathcal{N}(0, \delta^i_\tau)$ and $dV^i_\tau \sim \mathcal{N}(0, \delta^i_\tau)$ where $\delta^i_\tau$ is the delta dirac function. Let $y^i_\tau = \log(S^i_\tau)$. Using Itô’s lemma, the stochastic differential equation for $\log(S^i_\tau)$ is

\[
\begin{align*}
\frac{d}{\tau} \log(S^i_\tau) &= \left\langle \frac{\partial \log(S^i_\tau)}{\partial S^i_\tau}, S^i_\tau M^i_\tau \right\rangle d\tau \\
&\quad + \frac{1}{2} \left\langle \frac{\partial^2 \log(S^i_\tau)}{\partial (S^i_\tau)^2}, (S^i_\tau)^2 \right\rangle d\tau \\
&\quad + \left\langle \frac{\partial \log(S^i_\tau)}{\partial S^i_\tau}, S^i_\tau \right\rangle dW^i_\tau,
\end{align*}
\]

(31)

\[
\begin{align*}
\frac{d}{\tau} \log(S^i_\tau) &= \left( M^i_\tau - \frac{1}{2} \right) d\tau + dW^i_\tau.
\end{align*}
\]

(32)

\[
\begin{align*}
\frac{d}{\tau} \log(S^i_\tau) &= \left( M^i_\tau - \frac{1}{2} \right) d\tau + dV^i_\tau.
\end{align*}
\]

(33)

This leads to the following stochastic differential equations:

\[
\begin{align*}
y^i_\tau &= \left( Y^i_\tau \right) {\left[ Y^1_\tau \ Y^2_\tau \right]}^\top, \\
\frac{d}{\tau} \log(S^i_\tau) &= \left( M^i_\tau - \frac{1}{2} \right) d\tau + dW^i_\tau.
\end{align*}
\]

(34)

Consider the following matrices and vectors:

\[
\begin{align*}
\gamma(\tau) &\triangleq \left[ Y^1_\tau \ Y^2_\tau \right]^\top, \\
\kappa &\triangleq \begin{bmatrix} \kappa^1_\tau & 0 \\ 0 & \kappa^2_\tau \end{bmatrix}, \\
\sigma &\triangleq \begin{bmatrix} 0 & I_2 \\ 0 & -\kappa \end{bmatrix}, \\
\Gamma &\triangleq \begin{bmatrix} F \ 0 \\ 0 \ I_2 \end{bmatrix}, \\
\mu^i_\tau &\triangleq \left[ \Delta B^i_\mu \ 0 \right]^\top, \\
\xi^i_\tau &\sim \mathcal{N}(\mu^i_\tau, Q^i), \\
Q^i &\triangleq \begin{bmatrix} \Phi_1 \ 0 \\ 0 \end{bmatrix}, \\
C_\theta &\triangleq \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}, \\
c^i_\alpha &\in \left\{ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}^\top \right\}. \\
\end{align*}
\]

(43)

Finally, the eigenvectors $[U \ U]'$ of $\Gamma$ are computed, where $(C_\theta U, U^{-1} \Gamma' U)$ is observable and $U^{-1} \Gamma' U$ is Schur. This transformation provides (28).

System (36) is discretized with step $\Delta T$, which gives the following discrete-time stochastic linear dynamical system:

\[
\begin{align*}
\mathbf{y}(t\Delta T + \Delta T) &= \exp(F \Delta T) \mathbf{y}(t\Delta T) \\
\mathbf{m}(t\Delta T + \Delta T) &= \mathbf{m}(t\Delta T) + \Delta B^i_\mu + \Delta w(t\Delta T),
\end{align*}
\]

(37)

where $\Delta B^i_\mu$ and $\Delta w(t\Delta T)$ are derived using reference [30]:

\[
\begin{align*}
\Delta B^i_\mu &\triangleq \sum_{i=0}^m F^i B^i_\mu \frac{\Delta T^{i+1}}{(i+1)!}, \\
\Delta w(t\Delta T) &\sim \mathcal{N}(0, \Sigma),
\end{align*}
\]

(39)

with

\[
\Sigma = \int_{t\Delta T}^{(t+1)\Delta T} e^{F((t+1)\Delta T-\tau)} \Sigma e^{F^\top((t+1)\Delta T-\tau)} d\tau.
\]

(40)

Evaluating (39) is analytically intractable; therefore, the method in [31] is used to approximate $\Sigma$, i.e.,

\[
\begin{align*}
\Phi_{1,1} &\triangleq \begin{bmatrix} \Phi_{1,2} \end{bmatrix}, \\
\Phi_{1,2} &\triangleq \begin{bmatrix} \Phi_{1,1}^\top \ 0 \\ 0 \end{bmatrix}, \\
\end{align*}
\]

(41)

\[
\begin{align*}
\mathbf{y}_\Delta &\triangleq \mathbf{y}(t\Delta T) - \mathbf{y}(t\Delta T - \Delta T) = \log \left( \frac{S^i_{\Delta T}}{S^i_{(t-1)\Delta T}} \right).
\end{align*}
\]

(42)

Therefore, the difference $\mathbf{y}_\Delta = Y^i_\tau - Y^i_{\tau - \Delta T}$ is the logarithm of the percentage increase/decrease of buying at $t\Delta T - \Delta T$ and then selling at $t\Delta T$. We extend (37) by using the following matrices and vectors:

\[
\begin{align*}
\zeta^i_\tau &\triangleq \begin{bmatrix} y(t\Delta T) \\
\mathbf{m}(t\Delta T) \\
\end{bmatrix}, \\
\Gamma &\triangleq \begin{bmatrix} F \ 0 \\ 0 \ I_2 \end{bmatrix}, \\
\mu^i_\tau &\triangleq \left[ \Delta B^i_\mu \ 0 \right]^\top, \\
\xi^i_\tau &\sim \mathcal{N}(\mu^i_\tau, Q^i), \\
Q^i &\triangleq \begin{bmatrix} \Phi_1 \ 0 \\ 0 \end{bmatrix}, \\
C_\theta &\triangleq \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}, \\
c^i_\alpha &\in \left\{ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}^\top \right\}.
\end{align*}
\]

(43)