Multi-Task Gaussian Processes for Tactical Execution Under Regime Shifts

Álvaro Cartea* Fayçal Drissi† Gianluca Palmari‡

Latest version.
This version: December 23, 2025.

Abstract

In electronic markets, broker execution algorithms comprise a strategic layer that sets a trading schedule and a tactical layer that optimise child-order execution. This paper introduces the Multi-Task Gaussian Process bandit with Likelihood Ratio test (MTGP-LR) for the tactical layer. MTGP-LR maps market features to trading actions, adapts to regime shifts through change point detection, and employs a kernel that transfers learning across related actions. In stationary markets, MTGP-LR achieves sublinear regret. In nonstationary markets, we show how to control inference error and provide guarantees for detection delay. Our main result derives high-probability regret bounds in regime-switching markets. We illustrate the framework in two applications: child market orders timing and child limit order placement. Without assumptions on dynamics or regime transitions, MTGP-LR is accurate, computationally efficient, and outperforms control-based models when brokers misspecify model parameters.

Keywords: Gaussian processes, nonstationary bandits, transfer learning, change point detection, algorithmic trading

*Mathematical Institute and Oxford-Man Institute of Quantitative Finance, University of Oxford
†Oxford-Man Institute of Quantitative Finance, University of Oxford
‡Scuola Normale Superiore, Pisa, Italy

We thank Álvaro Arroyo, Sergio Calvo Ordoñez, Patrick Chang, Sebastian Jaimungal, and Anthony Ledford for helpful feedback and comments. We are also grateful to seminar participants at Oxford-Man Institute, and the Oxford Victoria Seminar.

An earlier version of this work was previously circulated under the title "Bandits for algorithmic trading with signals"; see Cartea et al. (2023b).

1 Introduction

In modern electronic financial markets, most execution algorithms used by brokers are organised in two layers: a strategic layer and a tactical layer (see Lehalle (2015), Tapia (2015), Chapters 3 and 7 of Guéant (2016), and Lehalle and Laruelle (2018)). The strategic layer determines the optimal trading schedule for a large *parent order*. The schedule accounts for the broker's objective, risk tolerance, and market impact. The tactical layer breaks the schedule into shorter time slices to execute *child orders* and optimises practical aspects such as placement, timing, and routing. A large literature develops optimal scheduling models including implementation shortfall (Almgren and Chriss (2000)), VWAP (Frei and Westray (2015)), and POV orders (Cartea et al. (2015)).[1] Other studies focus on aspects of the tactical layer (Wald and Horrigan (2005), Kovaleva and Iori (2012), Yam and Zhou (2017)). The performance and competitive edge of execution algorithms depend on the tactical layer's ability to obtain favorable prices, which requires incorporating market features and trading decisions into execution algorithms. However, classical approaches based on dynamic programming quickly become intractable as the dimensionality of the problem increases.

In this paper, we propose a data-driven approach for the tactical layer of execution algorithms. Given the strategic layer's schedule, we obtain the sequence of actions chosen by the tactical layer by solving a secondary optimisation problem using a new nonstationary contextual bandit, which we refer to as MTGP-LR. For each time slice of the strategic layer, the tactical layer's trading decisions correspond to the bandit's actions (arms), and market features are contextual information. Our framework is general; market features include future price movements, the fill probability of limit orders (LOs), or liquidity across venues. Trading decisions include the timing, placement, or routing of child orders.

MTGP-LR learns a reward function that maps market features to the financial performance of trading actions in noisy and non-stationary markets. The reward function is modeled as a sample from a multi-task Gaussian process (MTGP) that employs a parametric kernel for the contextual features and a non-parametric covariance matrix for the performance of different trading actions. Thus, MTGP-LR (i) learns the individual functions that map features to the rewards of each trading action, and (ii) exploits the correlation structure between these functions for transfer learning across trading actions that share common causal mechanisms. Importantly, performance is not degraded when the outcomes of different trading actions are uncorrelated. Our first contribution is an online change point detection mechanism based on a new likelihood ratio (LR) test for MTGPs.

[1] VWAP denotes volume-weighted average price orders, and POV denotes percentage of volume orders.

In stationary markets, we prove that MTGP-LR achieves sublinear regret with high probability. In non-stationary markets, MTGP-LR's regret is driven by three components: regret incurred within stationary segments, regret arising from detection delays following change points, and regret due to incorrect detection. Our main theoretical contribution is to provide regret guarantees for in regime-switching markets. We proceed in three steps. First, we derive dynamic LR thresholds that ensure high-probability control of both type-I and type-II errors. Second, we establish exponential tail bounds on the detection delay of change points. Third, we combine these results to derive high-probability regret bounds for MTGP-LR in regime-switching markets. The resulting bounds depend on the mutual information gain, the number of regime switches, and the LR thresholds.

As an application of our approach, we illustrate MTGP-LR as the tactical layer to implement a time weighted average price (TWAP) parent trade in two experimental setups. In the first example, the tactical layer uses price signals as contextual features to determine the timing of child market orders (MOs). The predictive power of the signals evolves according to a continuous-time Markov chain. We use stochastic control tools to derive the optimal tactical layer in closed form, which serves serves as an oracle when the broker correctly specifies all model parameters. We show that when the broker misspecifies model parameters, which frequently occurs in practice due to estimation errors and incorrect priors, MTGP-LR outperforms control-based methods. In the second example, the broker uses regime-switching signals for both price drift and volume imbalance in a limit order book (LOB), and she exploits this information for the optimal placement of child limit orders. This problem is intractable with stochastic control tools: it admits no closed-form solution, and a numerical approach requires solving an eleven-dimensional partial differential equation. We show that MTGP-LR performs well while remaining computationally efficient. Most importantly, in both examples, MTGP-LR does not require assumptions on the joint dynamics of features, on the number of market regimes, and on the transition probabilities between regimes.

Literature review Multiple works use Gaussian Processes (GPs) for financial decision-making. Ludkovski and Zail (2022) apply GPs to insurance loss modeling, Lyu et al. (2021), Lyu and Ludkovski (2022) explore their use in option pricing, and Ludkovski and Saporito (2021) propose GP-based methods for delta hedging in scenarios where models are computationally expensive. Closer to this work, Balata et al. (2021) propose a framework for approximate dynamic programming and Chen et al. (2024) for optimal execution. Other authors also study the use GPs in finance; Gonzalvez et al. (2019) employ uses Bayesian optimisation to study the term structure of interest rates, and Jafar (2022) uses GP regressions to study different option

management problems.

Our work contributes to the recent literature that proposes model-free approaches to scale to more complex financial environments; Guéant and Manziuk (2019) use deep reinforcement learning (RL) for optimal market making, Cartea et al. (2023c) use double-deep Q network learning for statistical arbitrage, Ning et al. (2021) use deep Q-learning for optimal execution, Arroyo et al. (2024) use a convolutional transformer to characterise fill probabilities, and Cartea et al. (2022) study RL algorithms for liquidity providers; see also Waldon et al. (2024). Our approach is related to the optimal placement and routing of child orders; see e.g., Wald and Horrigan (2005), Yingsaeree (2012), Kovaleva and Iori (2012) who study the choice between market and limit orders, to Guéant (2012) and Cartea and Jaimungal (2016) who study the optimal tracking of pre-computed schedules, and to Moallemi and Wang (2022) who explore the optimal timing of child orders based on unsupervised learning techniques. Finally, our application incorporates signals into optimal trading models, which has been extensively studied in the literature with the tools of dynamic programming; see Bechler and Ludkovski (2015), Belak et al. (2018), Lehalle and Neuman (2019), Bellani et al. (2021), Drissi (2022), Donnelly (2022), Cartea et al. (2018, 2024, 2025, 2023a).[2] In contrast to this literature, our approach does not require assumptions on the joint dynamics of the prices and the signals, on the number of market regimes, and on the transition probabilities between regimes.

In the RL literature, contextual bandits are applied in different settings such as healthcare (Durand et al., 2018), recommender systems (Zhou et al., 2017), dialogue systems (Liu et al., 2018), anomaly detection (Ding et al., 2019), and economics (Cohen and Treetanthiploet, 2021). Early bandit algorithms include Exp3, see Seldin et al. (2013), and LinUCB, see Chu et al. (2011), and those based on GPs are in Srinivas et al. (2009), Krause and Ong (2011) and the extensions in Bogunovic et al. (2016), Bogunovic and Krause (2021), Duran-Martin et al. (2022, 2024), and Waldon et al. (2024). Our work contributes to this literature by proposing a new algorithm that uses a sample from an MTGP to jointly model the reward functions of the bandit's arms.

Nonstationary bandits are difficult to study due to poor mathematical tractability. They are solved with either passive policies (e.g., the sliding window UCB in Garivier and Moulines (2008) and Rexp3 in Besbes et al. (2014)), or with active policies (e.g., Srivastava et al. (2014) and Cao et al. (2019)). The former decreases the weights of past rewards and the latter tracks reward distributions to reset the bandit when a change is detected. Recent work considers active policies that use LR tests to detect changes in the distribu-

[2] See also Bergault et al. (2022) for multi-asset extensions, and Arroyo et al. (2022), Scalzo et al. (2021) for applications to portfolio construction.

tion of the rewards, see Dette and Gösmann (2020). In particular, Caldarelli et al. (2022) consider a LR test for GP bandits with different hypotheses. In this paper, we contribute to the literature on active policies by proposing a new LR test for online change point detection for MTGPs and we derive regret guarantees for nonstationary bandits.

The remainder of this paper proceeds as follows. Section 2 outlines the class of problems addressed in this paper and presents some examples. Section 3 presents the MTGP bandit algorithm and derives regret bounds in stationary environments. Section 4 presents MTGP-LR, shows how to control inference error, and provides regret bounds in piecewise stationary markets. Finally, Section 5 presents extensive numerical experiments.

2 Strategic and tactical layers for execution

Here, Section 2.1 outlines the specific class of problems addressed in this paper and formalises the problem as a contextual bandit, Section 2.2 illustrates the use of the tactical layer to incorporate short term signals, and Section 2.3 illustrates the use of the tactical layer to incorporate order type selection.

2.1 Two-layered execution

In line with industry practice (see Lehalle (2015), Tapia (2015), Guéant (2016), Lehalle and Laruelle (2018)), we consider an execution algorithm consisting of a strategic and a tactical layer. The strategic layer is an execution strategy that encodes the urgency, the risk aversion, and the market impact of the trader, and we denote by $(\rho_t)_{t\in\mathbb{R}_+}$ the (predictable) speed of trading of this layer. The objective of the strategic layer is to liquidate $\mathcal{Q}$ shares throughout a trading window $[0,T]$, where $T > 0$. In practice, brokers discretise the optimal schedule ρ of the strategic layer and split it into child orders to be executed within time slices $[t_i, t_{i+1}]$, where $\mathcal{T} = \{t_1, \ldots, t_n\}$ is a set of discrete observation times and $t_n = T$. In contrast, the tactical layer uses $M \in \mathbb{N}$ contextual market features $\mathcal{I} = \{I^1, \ldots, I^M\}$ and a space of $N \in \mathbb{N}$ actions $\mathcal{A} = \{a^1, \ldots, a^N\}$ to target the execution strategy ρ and to improve the execution prices of child orders. We assume that each market feature takes values in a compact subset of $\mathbb{R}$.

To illustrate the use of the tactical layer, consider the following naive algorithm. Let $(q_t)_{t\in\mathcal{T}}$ be the discrete-time inventory process of the investor. At every observation time t_j, the investor compares the target *optimal* inventory $\sum_{i=1}^{j}(t_i - t_{i-1})\,\rho_i$ with her current inventory q_j and uses a market order to execute

a child order of size

$$\delta_j = \sum_{i=0}^{j} (t_i - t_{i-1})\, \rho_i - q_j \,;$$

when $\delta_j \leq 0$ the investor must sell the asset, and when $\delta_j \geq 0$ she buys the asset.

In this work, we model the tactical layer as an bandit algorithm that learns a *reward function* that maps market features (context) to the performance of a set of execution decisions for child orders (actions).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space supporting all the random variables defined below, and let $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ denote the natural filtration representing the information available to the agent up to time t. The reward function at each time $t \in \mathcal{T}$ is denoted by $f_t : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$. It is a measurable function defined on the joint context–action space, representing the expected reward associated with each pair $(a, \mathbf{x})$ at time t.[3] We define the bandit as the quadruple $\mathcal{B} = \left\{\mathcal{A}, \mathcal{X}, \mathcal{T}, (f_t)_{t\in\mathcal{T}}\right\}$.

At each time step $t \in \mathcal{T}$, the investor observes the realization of an $\mathcal{F}_{t-1}$-measurable context $\mathbf{x}_t : \Omega \to \mathcal{X}$, selects an $\mathcal{F}_{t-1}$-measurable action $a_t \in \mathcal{A}$, and observes an $\mathcal{F}_t$-measurable reward $y_t(a_t, \mathbf{x}_t) = f_t(a_t, \mathbf{x}_t) + \varepsilon_t$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. Gaussian noise. The objective is to minimise the expected cumulative regret

$$R(T) = \sum_{t=1}^{T} r_t = \sum_{t=1}^{T} \left(f_t(a_t^*, \mathbf{x}_t) - f_t(a_t, \mathbf{x}_t)\right), \tag{1}$$

where a_t^* is the optimal action at time t, i.e., $a_t^* = \arg\max_{a\in\mathcal{A}} f_t(a, \mathbf{x}_t)$. In our model, action a_t is selected using the upper confidence bound (UCB)

$$a_t = \arg\max_{a\in\mathcal{A}} \{\vartheta_{t-1}(a, \mathbf{x}_t) + \beta_t^{1/2}\, \varphi_{t-1}(a, \mathbf{x}_t)\}, \tag{2}$$

where $\vartheta_{t-1}(a, \mathbf{x}_t)$ and $\varphi_{t-1}(a, \mathbf{x}_t)$ are the posterior mean and posterior standard deviation for action a at the test point $\mathbf{x}_t$ conditioned on data observations up to time $t-1$, respectively, and $\{\beta_t\}_{t\in\mathcal{T}}$ is a deterministic, non-decreasing sequence controlling the exploration-exploitation trade-off. The UCB objective (2) balances exploration by selecting actions with large uncertainty, i.e., large value of the posterior variance $\varphi_{t-1}(a, \mathbf{x}_t)$ in (2), with exploitation by choosing actions with large potential reward, i.e., large value of the posterior mean $\vartheta_{t-1}(a, \mathbf{x}_t)$ in (2). At time t, the parameter β_t modulates the relative importance of one component relative to the other.

[3] We index the function by time because we consider nonstationary environments in the following sections.

The posterior mean and posterior standard deviation depend on the specification of the reward function. In this work, we model the reward function as a sample from an MTGP. In the next two sections, we present two practical settings to illustrate the use of the tactical layer. Later, we use these settings in our numerical experiments.

2.2 Example one: market order timing

Here, the tactical layer uses short-term signals to time child market orders. The contextual market features are a set of N short-term predictive signals for the future price, and the actions are the timing of market orders in a LOB.

At every observation time $t_j \in \mathcal{T}$, action $a^\iota \in \mathcal{A}$, for $\iota \in \{1, \ldots, N\}$, uses a signal I^ι to decide whether to send the market order at the beginning or at the end of each time slice $[t_j, t_{j+1}]$. More precisely, assume the investor wishes to sell a quantity $-\delta_j$ over $[t_j, t_{j+1}]$ to track the inventory of the strategic layer. If the signal with the best UCB score predicts a price increase over the slice $[t_j, t_{j+1}]$, then it is more profitable to sell at time t_{j+1}. Similarly, if the signal predicts a decrease in the price over $[t_j, t_{j+1}]$, it is more profitable to sell at time t_j.[4]

The N signals, together with additional relevant market features (such as volatility, bid-ask spread, or time of day), constitute the *contextual information* $\mathcal{X}$ used by the bandit algorithm. After executing a market order at time t_j based on a signal $I^\star$, the investor observes a reward

$$y_{j+1} = \left(\tilde{S}_{j+1} - S^\star_{j+1}\right) \times \text{sign}\left(\delta_j\right) \tag{3}$$

at time t_{j+1}, where $\tilde{S}_{j+1}$ is the execution price of a benchmark tracking strategy (e.g., the TWAP strategy over the interval $[t_j, t_{j+1}]$), and $S^\star_{j+1}$ is the execution price received by the investor. As describe below, the reward (3) is used to update the hyper-parameters of the algorithm.

2.3 Example two: limit order placement

To reduce trading costs, investors use LOs to avoid crossing the spread. However, they face execution risk. Here, we showcase the use of the tactical layer to incorporate order type selection between passive LOs and

[4]Here, we assume that the predictive horizon of the signals coincides with the timescale of the time slice used to execute the child order.

aggressive MOs. The contextual market features are a set of short-term predictive signals for liquidity, and the actions are the choice between LOs of multiple levels of depth or an MO.

Assume the investor wishes to buy a quantity δ_j over the slice $[t_j, t_{j+1}]$ and let S_j denote the midprice at time t_j. A buy LO of size δ_j posted at $S_j - \mu$, where $\mu > 0$ is the LO's depth, improves the execution price if filled compared to that of an MO. Larger values of the depth μ decrease the *fill probability*, i.e., the probability of being executed over $[t_j, t_{j+1}]$. The fill probability of LOs depends on the level of liquidity and other market features; see Arroyo et al. (2024). Assuming a flow of MOs with sizes exponentially distributed with mean $\overline{v}$, and a block-shaped LOB with volume A at each price level of the LOB, the fill probability decays exponentially with μ, i.e.,

$$\mathbb{P}[\text{execution of LO of depth } \mu] = \exp\left\{-\frac{A}{\overline{v}}\,\mu\right\}.$$

Let $\{\mu^1, \ldots, \mu^N\}$ denote a set of levels for the depth of an LO, i.e., number of price ticks from the best bid or best ask. In the tactical layer, action a^ι for $\iota \in \{1, \ldots, N\}$, denotes posting an LO at depth μ^ι, while actions a_0^{MO} and a_1^{MO} denote respectively sending an MO at the beginning and at the end of the trading slice. LOs are not guaranteed to be filled within the slice $[t_j, t_{j+1}]$, so action a^ι includes completing any unfilled quantity with a *fallback MO* at time t_{j+1}.

Let $S^\star_{j+1}$ denote the execution price received by the investor by choosing the optimal action $a^\star_j$ at time t_j.[5] The reward observed at time t_{j+1} is

$$y_{j+1} = \left(\tilde{S}_{j+1} - S^\star_{j+1}\right) \times \text{sign}(\delta_j)\,,$$

where $\tilde{S}_{j+1}$ denotes the benchmark price (e.g., TWAP over $[t_j, t_{j+1}]$).

3 MTGP bandit

Here, we study the tactical layer when the reward function of the tactical layer is sampled from an MTGP defined over a joint context-action space. Below, Section 3.1 introduces GPs, and Section 3.2 formalises the MTGP contextual bandit algorithm and proves regret guarantees in stationary environments.

[5] For an LO, let $S^{\star,\text{LO}}_{j+1}$ and $S^{\star,\text{MO}}_{j+1}$ denote the volume-weighted prices of the filled LO and fallback MO, respectively. The execution price is $S^\star_{j+1} = \alpha^\star_{j+1} S^{\star,\text{LO}}_{j+1} - (1 - \alpha^\star_{j+1}) S^{\star,\text{MO}}_{j+1}$, where $\alpha^\star_{j+1} \in [0, 1]$ is the fraction of the order filled by the LO.

3.1 Gaussian processes

In this paper, we use lowercase bold symbols to denote vectors, bold uppercase symbols to denote matrices, the subscript $\star$ on random vector variables to denote inference or posterior prediction, and the subscript $\star,\star$ on matrices to denote inference.

GPs are a flexible and a non-parametric class of models for nonlinear random functions. They are notably used for Bayesian inference and Bayesian optimisation in machine learning; see Williams and Rasmussen (2006). Formally, a GP is a random function $f : \mathcal{X} \to \mathbb{R}$, such that, for any finite set of points $\mathbf{X}_\star \subseteq \mathcal{X}$, the random vector $\boldsymbol{f}_\star = \{f(\mathbf{x})\}_{\mathbf{x}\in\mathbf{X}_\star}$ follows a multivariate Gaussian distribution.[6] The shape of the function f is determined by a finite set of (training) observations $\boldsymbol{y} = \{y_i\}_{i\in\{1,\dots,n\}}$ collected at the (training) observation points $\mathbf{X} = \{\mathbf{x}_i\}_{i\in\{1,\dots,n\}}$, where $y_i = f(\mathbf{x}_i) + \varepsilon_i$ is subject to i.i.d. Gaussian measurement noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $\sigma > 0$. GPs are fully specified by a mean function $\mu : \mathcal{X} \to \mathbb{R}$ and a covariance (kernel) function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e.,

$$\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \quad \text{and} \quad k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))] \,.$$

In particular, if $f \sim \mathcal{GP}(\mu, k)$ and $\mathbf{X}_\star$ is a set of test points in the domain $\mathcal{X}$ of the GP, then the set of random variables $\boldsymbol{f}_\star$ is Gaussian with parameters $\mathcal{N}(\boldsymbol{\mu}_\star, \boldsymbol{K}_{\star,\star})$, where

$$\boldsymbol{\mu}_\star = \{\mu(\mathbf{x})\}_{\mathbf{x}\in\mathbf{X}_\star} \quad \text{and} \quad \boldsymbol{K}_{\star,\star} = \{k(\mathbf{x}, \mathbf{x}')\}_{(\mathbf{x},\mathbf{x}')\in\mathbf{X}_\star} \,.$$

Without loss of generality, we assume throughout this section that $\mu = 0$. The kernel function k is defined with a vector of hyper-parameters $\boldsymbol{\theta}$ and specifies the correlation behaviour of every finite subset of points. Thus, the kernel function encodes the smoothness and regularity properties of sample functions drawn from the GP. In this work, we motivate our model with applications to finance where a sample from a GP is the reward function that maps market features to trading performance; see Hollifield et al. (2004) and Cartea et al. (2018) for examples. We restrict the choice of kernels to stationary differentiable kernels with bounded variance.[7]

Assumption 1. *The kernel satisfies* $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}') \leq \kappa \leq 1$, $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

A convenient property of GPs is that one computes the posterior distribution with analytic formulae.

[6] The domain $\mathcal{X}$ of the function f is an arbitrary subset (possibly discrete) of $\mathbb{R}^m$, where $m > 0$.

[7] This condition is not restrictive and is a standard assumption in the literature; see Rasmussen and Ghahramani (2000).

Suppose we collect n noisy observations $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ at the domain points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $y_i = f(\mathbf{x}_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Then, the posterior distribution over f given the previous (training) observations $\mathbf{X}$ and $\boldsymbol{y}$, is also a GP with mean function μ_{post} and covariance function k_{post} given by

$$\begin{cases} \mu_{\text{post}}(\mathbf{x}_\star) & = \boldsymbol{k}(\mathbf{x}_\star, \mathbf{X}) \left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}\,, \\ k_{\text{post}}(\mathbf{x}_\star, \mathbf{x}'_\star) & = k(\mathbf{x}_\star, \mathbf{x}'_\star) - \boldsymbol{k}(\mathbf{x}_\star, \mathbf{X}) \left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}(\mathbf{X}, \mathbf{x}'_\star)\,, \end{cases} \tag{4}$$

where

$$\boldsymbol{k}(\mathbf{x}_\star, \mathbf{X}) = \boldsymbol{k}(\mathbf{X}, \mathbf{x}_\star)^\intercal = \left(k(\mathbf{x}_\star, \mathbf{x}_1), \ldots, k(\mathbf{x}_\star, \mathbf{x}_n)\right)$$

is the n-dimensional covariance vector of the test point $\mathbf{x}_\star$ with training points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\boldsymbol{K} = \left(k(\mathbf{x}_i, \mathbf{x}_j)\right)_{i,j \in \{1,\ldots,n\}}$ is the positive semi-definite kernel matrix from training data, and $\boldsymbol{I}$ is the n-dimensional identity matrix.

The elements of the vector $\boldsymbol{\theta} \in \Theta$ are the hyper-parameters of the prior's kernel function and σ^2 is the variance of the i.i.d. Gaussian noise that corrupts reward observations. Both $\boldsymbol{\theta}$ and σ^2 are inferred with the log marginal likelihood of the data given by

$$\begin{aligned} L(\boldsymbol{\theta}, \sigma) &= \log p(\, \boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma) \\ &= -\frac{1}{2} \log\Big(\det\left(\boldsymbol{K}_{\boldsymbol{\theta}} + \sigma^2 \boldsymbol{I}\right)\Big) - \frac{1}{2}\, \boldsymbol{y}^\intercal \left(\boldsymbol{K}_{\boldsymbol{\theta}} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y} - \frac{n}{2} \log(2\,\pi)\,, \end{aligned} \tag{5}$$

for a zero-mean GP, where $\mathbf{X}$ and $\boldsymbol{y}$ are the n training samples and $\boldsymbol{K}_{\boldsymbol{\theta}}$ is the prior's positive covariance matrix with kernel $k_{\boldsymbol{\theta}}$. The vector of hyper-parameters $\boldsymbol{\theta}$ and the variance σ^2 maximise the quantity in (5), i.e., $(\boldsymbol{\theta}^\star, \sigma^\star) \in \underset{\boldsymbol{\theta} \in \Theta, \sigma \in \mathbb{R}^+}{\arg\max}\, L(\boldsymbol{\theta}, \sigma)$, which one solves with classical gradient descent optimisation.

3.2 MTGP in stationary markets

In contextual bandits, it is essential to model regularity of the reward function f_t jointly over contexts and actions. Throughout this section, let the context space be $\mathcal{X} \subset \mathbb{R}^m$ with $m \in \mathbb{N}$, and assume that $\mathcal{X}$ is compact and convex; moreover, assume $\mathcal{X} \subset [-s, s]^m$ for some $s > 0$. Let the action set $\mathcal{A}$ be finite with $|\mathcal{A}| = N$, and define $\Omega = \mathcal{A} \times \mathcal{X}$. A stationary regime is a time interval on which the latent reward function is time-invariant, i.e., there exists a function $f : \Omega \to \mathbb{R}$ such that $f_t \equiv f$ for all t in the regime. Within a stationary regime, we write f instead of f_t and place a multi-task Gaussian process prior on f, denoted

$f \sim \mathcal{GP}(0, k)$, with separable kernel

$$k\big((a, \mathbf{x}), (a', \mathbf{x}')\big) \;=\; k^{\mathcal{X}}(\mathbf{x}, \mathbf{x}')\, \mathbf{K}^{\mathcal{A}}_{a,a'}\,, \qquad \forall (a, x), (a', x') \in \Omega, \tag{6}$$

where $k^{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semi-definite kernel on $\mathcal{X}$ and $\mathbf{K}^{\mathcal{A}} = (\mathbf{K}^{\mathcal{A}}_{a,a'})_{(a,a')\in\mathcal{A}\times\mathcal{A}} \in \mathbb{R}^{N\times N}$ is a learned positive semi-definite action-similarity matrix. Let $\mathcal{D}_t = \{(\mathbf{w}_i, y_i)\}_{i=1}^{n}$ be the data available at time t within the current stationary regime, consisting of the $n = n(t)$ observation pairs collected since the most recent change point; here each $\mathbf{w}_i = (a_i, \mathbf{x}_i) \in \Omega$ and $y_i \in \mathbb{R}$.[8] Assume the observation model $y_i = f(\mathbf{w}_i) + \varepsilon_i$ with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, and define $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and $\boldsymbol{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$. For any $\mathbf{w}_\star, \mathbf{w}'_\star \in \Omega$, the MTGP posterior at time t (see (4)) is given by

$$\begin{aligned}
\mu_{\text{post}}(\mathbf{w}_\star) &= \boldsymbol{k}(\mathbf{w}_\star, \mathbf{W})\, (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}, \\
k_{\text{post}}(\mathbf{w}_\star, \mathbf{w}'_\star) &= k(\mathbf{w}_\star, \mathbf{w}'_\star) - \boldsymbol{k}(\mathbf{w}_\star, \mathbf{W})\, (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}(\mathbf{W}, \mathbf{w}'_\star),
\end{aligned}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix, $\boldsymbol{k}(\mathbf{w}_\star, \mathbf{W}) = \big(k(\mathbf{w}_\star, \mathbf{w}_1), \ldots, k(\mathbf{w}_\star, \mathbf{w}_n)\big) \in \mathbb{R}^{1\times n}$, $\boldsymbol{k}(\mathbf{W}, \mathbf{w}_\star) = \boldsymbol{k}(\mathbf{w}_\star, \mathbf{W})^\top \in \mathbb{R}^{n\times 1}$, and $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ is the Gram matrix with entries $\boldsymbol{K}_{ij} = k(\mathbf{w}_i, \mathbf{w}_j)$. With the separable kernel (6), define the matrices

$$\boldsymbol{K}^{\mathcal{X}}_{\boldsymbol{\theta}} = \big(k^{\mathcal{X}}_{\boldsymbol{\theta}}(x_i, x_j)\big)_{i,j=1}^{n} \qquad \text{and} \qquad \boldsymbol{K}^{\mathcal{A}}_{\boldsymbol{a}} = \big(K^{\mathcal{A}}_{a_i,a_j}\big)_{i,j=1}^{n},$$

so that $\boldsymbol{K} = \boldsymbol{K}^{\mathcal{X}}_{\boldsymbol{\theta}} \odot \boldsymbol{K}^{\mathcal{A}}_{\boldsymbol{a}}$, where $\odot$ denotes the Hadamard (element-wise) product.[9] The action-similarity matrix $K^{\mathcal{A}}$ is learned jointly with the hyper-parameters $\boldsymbol{\theta} \in \Theta$ of $k^{\mathcal{X}}_{\boldsymbol{\theta}}$ and the noise level $\sigma \in \mathbb{R}_+$ by maximizing the log marginal likelihood

$$(\boldsymbol{\theta}^\star, \sigma^\star, K^{\mathcal{A},\star}) \in \underset{\boldsymbol{\theta}\in\Theta,\ \sigma\in\mathbb{R}_+,\ K^{\mathcal{A}}\in\mathcal{S}^N_+(\mathbb{R})}{\arg\max} L(\boldsymbol{\theta}, \sigma, K^{\mathcal{A}}),$$

[8] The number n of observations available at time t need not equal t due to possible change points; see Subsection 4.

[9] A Kronecker representation $K^{\mathcal{A}} \otimes K^{\mathcal{X}}$ corresponds to a separable covariance defined on a full action–context Cartesian grid after vectorizing the associated multi-output process; in contrast, here the covariance matrix is constructed directly over the observed pairs (a_i, x_i), yielding $\boldsymbol{K}_{ij} = K^{\mathcal{A}}_{a_i,a_j} k^{\mathcal{X}}(x_i, x_j)$ and thus the Hadamard form above.

where $\mathcal{S}_+^N(\mathbb{R})$ is the cone of real $N \times N$ positive semi-definite matrices and

$$L(\boldsymbol{\theta}, \sigma, K^{\mathcal{A}}) = -\frac{1}{2}\log\det\left(\boldsymbol{K}_{\boldsymbol{\theta}}^{\mathcal{X}} \odot \boldsymbol{K}_{\boldsymbol{a}}^{\mathcal{A}} + \sigma^2 \boldsymbol{I}\right) - \frac{1}{2}\boldsymbol{y}^\top \left(\boldsymbol{K}_{\boldsymbol{\theta}}^{\mathcal{X}} \odot \boldsymbol{K}_{\boldsymbol{a}}^{\mathcal{A}} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y} - \frac{n}{2}\log(2\pi).$$

Consequently, the MTGP prior is parametrised by the context kernel $k^{\mathcal{X}}$ (controlling regularity over $\mathcal{X}$) and the action similarity matrix $\mathbf{K}^{\mathcal{A}}$ (encoding cross-action dependence), and learning $\mathbf{K}^{\mathcal{A}}$ enables information sharing across actions through the induced correlations of their latent reward functions.

In contrast with our approach, the classical approach in the literature on contextual GP bandits (see eg. Krause and Ong (2011)) is to model f as a sample from a known GP and to decompose the GP's covariance function k into two parametric kernel functions, $k^{\mathcal{X}}$ over contexts $\mathcal{X}$ and $k^{\mathcal{A}}$ over actions $\mathcal{A}$, both of which encode the smoothness of the reward function f. The classical GP bandit with parametric kernels also accommodates transfer learning. However, in many practical problems and in the financial application that we study, it is undesirable to consider a unique parametric kernel function $k^{\mathcal{A}}$ that encodes the global similarity between actions. Often, the outcomes of a subset of actions share common causal effects while the outcomes of another subset of actions are unrelated. Thus, forcing the algorithm to learn similarity between actions with a single parametric kernel function will harm performance as a result of spurious transfer learning that is enforced between actions that are unrelated. In particular, the aim of our algorithm is to transfer learning only within clusters of similar actions. There are other techniques that solve similar issues; see Bonilla et al. (2007) and references therein.

Bandits are *no-regret* or achieve *sublinear* regret when $\lim_{T\to\infty} R(T)/T = 0$ (see (1)). Bounds on the regret impose bounds on the optimisation problem solved by the bandit. In stationary environments, Theorem 1 gives a sub-linear bound with probability $1-\delta$ for any value of $\delta \in (0,1)$.

Theorem 1. *Fix $\delta \in (0,1)$, and suppose $\mathcal{X} \subset [-s,s]^m$ is compact, where $m \in \mathbb{N}$ and $s > 0$. Assume that the reward function does not change (stationary regime) and write $f_t = f$ for all $t \in \mathcal{T}$. Let f be sampled from a known MTGP prior with known covariance function k of the form (6) and known noise variance σ^2. Then, for any set of observations of contexts $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, the contextual regret (see (1)) of the Contextual Gaussian Process Bandit with UCB is bounded with high probability and we write*

$$\mathbb{P}\left[R(T) \leq \sqrt{C_1\, T\, \beta_T\, \gamma_T} + 2, \quad \forall T \geq 1\right] \geq 1-\delta\,, \tag{7}$$

where

$$C_1 = \frac{8}{\log(1+\sigma^{-2})}, \quad \beta_T = 2\log(N\,T^2\,\pi^2/3\,\delta\,), \quad \gamma_{T,\mathbf{x}_{1:T}} = \max_{A\subset\mathcal{A}:\#A=T} I\left(\boldsymbol{y}_{A,\mathbf{x}_{1:T}};\mathbf{f}_{A,\mathbf{x}_{1:T}}\right), \tag{8}$$

and $\mathbf{f}_{A,\mathbf{x}_{1:T}} = [f\,(a_t,\mathbf{x}_t)]_{t\in[T]}\,, \quad \boldsymbol{y}_{A,\mathbf{x}_{1:T}} = \mathbf{f}_{A,\mathbf{x}_{1:T}} + \boldsymbol{\varepsilon}_A, \quad A = \{a_1,\cdots,a_T\}, \quad$ *and* $\boldsymbol{\varepsilon}_A \sim \mathcal{N}\left(\mathbf{0},\sigma^2\boldsymbol{I}\right)$. *Moreover, if the MTGP's kernel* $k^{\mathcal{X}}$ *is the Squared Exponential, Matérn, or Linear kernel, then, the regret bound in* (7) *is sub-linear.*

For a proof, see Appendix B.1. The quantity $\gamma_{T,\mathbf{x}_{1:T}}$ in (7) is called the maximal information gain due to sampling, conditioned on the observed context $\mathbf{x}_{1:T}$, which quantifies the informativeness of a set of sampling points about the reward function f_t, or equivalently the mutual information between the true function f and the noisy observations $y_t = f\,(a_t,\mathbf{x}_t) + \varepsilon_t$.

Theorem 1 shows that the regret incurred in stationary environments depends on how efficiently one can explore the reward function in action and context space to gather information. The maximal information gain is bounded for a large class of kernels (see Srinivas et al. (2009)), in particular, they are bounded for stationary kernels with bounded variance, such as the finite-dimensional linear kernel and squared exponential kernel.[10] This result uses the sub-modularity of the information gain with respect to context, which allows one to bound the information gain with its discretised version. We use the discrete nature of our action space to extend the result in Srinivas et al. (2009) to the multi-task GP case.

4 MTGP-LR in nonstationary markets

In this section we introduce MTGP-LR for the tactical layer when the mapping from market features to trading decisions goes through regime changes. Section 4.1 describes the algorithm and introduces the change point detection LR test.

In nonstationary markets, regret is driven by three components: regret incurred within stationary segments (analysed above), regret arising from detection delays following change points, and regret due to incorrect detection. Sections 4.2, 4.3, and 4.4 study the latter two components. Specifically, Section 4.2 provides dynamic thresholds that bound type-I and type-II error probabilities. Section 4.3 establishes an exponential tail bound on the detection delay (type-II error) following true change points. Section 4.4 uses

[10] We use the squared exponential kernel in Subsection 5 as an application of our results to algorithmic trading.

the dynamic thresholds of Section 4.2 to derive bounds on the cumulative probability of incorrect detection (type-I error) over time.

Finally, Section 4.5 combines these results with the sublinear regret bounds for stationary segments to obtain high-probability regret bounds for MTGP-LR in regime-switching environments. These bounds are expressed in terms of mutual information gain, the number of regime switches, and the LR thresholds.

4.1 The MTGP-LR algorithm

MTGP-LR is designed for regime-switching reward functions $\{f_t\}_{t=1}^T$. Let ℓ be the number of stationary regimes, defined as

$$\ell \;=\; 1 + \sum_{t=2}^{T} \mathbb{1}_{\{f_t \neq f_{t-1}\}}\,.$$

Let $\{b_j\}_{j=0}^{\ell} \subset \mathcal{T}$ be the (true) change point sequence with $0 = b_0 < b_1 < \cdots < b_\ell = T$, so that f_t is stationary on each regime $t \in [b_{j-1}+1,\ b_j]$, with $j = 1, \ldots, \ell$. Thus, $\{b_j\}_{j=1}^{\ell-1}$ are the (true) change points, i.e., the observation times at which the reward function changes. A regime change modifies the mapping from market features to rewards of trading actions, and can therefore change the optimal action. After a regime change, only the most recent observations are informative for inferring the new reward function. Below, we introduce an online LR test for change point detection to make the tactical layer robust to regime changes: when a change is detected, MTGP-LR resets and discards pre-change reward observations.

Let $\tau_1 < \tau_2 < \cdots$ be the (random) trigger times of the sequential LR test, and define the associated (random) reset times $\{\hat{b}_k\}_{k\geq 0}$ by $\hat{b}_0 = 0$ and $\hat{b}_k = \tau_k$ whenever the test triggers at time τ_k. (If the test never triggers again, the sequence stops; in particular, we later set $\hat{b}_{\hat{\ell}} = T$ by convention.) For any $t \in \mathcal{T}$, let $\hat{b}(t) \;=\; \max\{\hat{b}_k :\ \hat{b}_k < t\}$ denote the most recent reset time strictly before t. We define the number of post-reset observations available at time t as $P_t \;=\; t - \hat{b}(t)$. Let $\mathcal{Y}_t$ be the time window containing the last P_t rewards since the most recent reset, $\mathcal{Y}_t \;=\; \{y_s\}_{s=\hat{b}(t)+1}^{t}$, and let $\mathcal{W}_t$ be the corresponding time window of action–context tuples,

$$\mathcal{W}_t \;=\; \{\mathbf{w}_s\}_{s=\hat{b}(t)+1}^{t}, \qquad \mathbf{w}_s = (a_s, \mathbf{x}_s).$$

We write $\mathbf{W}_t$ for the *set* of tuples in the window $\mathcal{W}_t$, i.e., $\mathbf{W}_t = \{\mathbf{w}_s :\ s = \hat{b}(t)+1, \ldots, t\}$. Fix a block size $p \leq P_t$. Let $\overline{\mathcal{Y}}_t \subset \mathcal{Y}_t$ be the sub-window of the most recent p rewards, and $\underline{\mathcal{Y}}_t = \mathcal{Y}_t \setminus \overline{\mathcal{Y}}_t$ the complementary sub-windows $\overline{\mathcal{Y}}_t = \{y_s\}_{s=t-p+1}^{t}$, and $\underline{\mathcal{Y}}_t = \{y_s\}_{s=\hat{b}(t)+1}^{t-p}$. Define analogously the sub-windows of tuples $\overline{\mathcal{W}}_t$

and $\underline{\mathcal{W}}_t$ and their associated sets $\overline{\mathbf{W}}_t$ and $\underline{\mathbf{W}}_t$.

MTGP-LR tests whether a change has occurred between the sub-windows $\underline{\mathcal{Y}}_t$ and $\overline{\mathcal{Y}}_t$ (equivalently, between $\underline{\mathcal{W}}_t$ and $\overline{\mathcal{W}}_t$). We consider the hypotheses:

- H_0: the reward function governing $\overline{\mathcal{Y}}_t$ is the same as that governing $\underline{\mathcal{Y}}_t$, and the noise variance is σ_0^2.
- H_1: the rewards in $\overline{\mathcal{Y}}_t$ are generated by a new MTGP draw with zero-mean prior, kernel k (fixed), and noise variance σ_1^2.

Let $\underline{\mathcal{D}}_t = (\underline{\boldsymbol{y}}_t, \underline{\mathbf{W}}_t)$, and $\overline{\mathcal{D}}_t = (\overline{\boldsymbol{y}}_t, \overline{\mathbf{W}}_t)$, denote the datasets from the sub-windows $\underline{\mathcal{W}}_t$ and $\overline{\mathcal{W}}_t$, respectively, where $\underline{\boldsymbol{y}}_t$ (resp. $\overline{\boldsymbol{y}}_t$) stacks the rewards in $\underline{\mathcal{Y}}_t$ (resp. $\overline{\mathcal{Y}}_t$) *in chronological order* (and the corresponding tuples in $\underline{\mathbf{W}}_t$ and $\overline{\mathbf{W}}_t$ are ordered consistently). The LR for the kernel function k is

$$\mathcal{R}_t = 2\log\frac{p(\overline{\boldsymbol{y}}_t|H_1)}{p(\overline{\boldsymbol{y}}_t|H_0)} = 2\log\frac{p(\overline{\boldsymbol{y}}_t|\overline{\mathbf{W}}_t)}{p(\overline{\boldsymbol{y}}_t|\overline{\mathbf{W}}_t, \underline{\mathcal{D}}_t)}\,. \tag{9}$$

Next, we derive the analytical formula of the LR statistic in (9). Let $\overline{\boldsymbol{K}}_t = \{k(\overline{\mathbf{w}}, \overline{\mathbf{w}})\}_{\overline{\mathbf{w}}\in\overline{\mathbf{W}}_t}$, $\boldsymbol{K}_t = \{k(\underline{\mathbf{w}}, \overline{\mathbf{w}})\}_{\underline{\mathbf{w}}\in\underline{\mathbf{W}}_t, \overline{\mathbf{w}}\in\overline{\mathbf{W}}_t}$ and $\underline{\boldsymbol{K}}_t = \{k(\underline{\mathbf{w}}, \underline{\mathbf{w}})\}_{\underline{\mathbf{w}}\in\underline{\mathbf{W}}_t}$. Write the likelihoods of both hypotheses as

$$\begin{aligned}\log p(\overline{\boldsymbol{y}}_t \mid \overline{\mathbf{W}}_t, H_0) &= \log\int_f p(\overline{\boldsymbol{y}}_t|f)\, p(f|\overline{\mathbf{W}}_t, \underline{\mathcal{D}}_t)\, df \\ &= -\frac{1}{2}(\overline{\boldsymbol{y}}_t - \widetilde{\boldsymbol{\mu}}_t)^\top \left(\widetilde{\overline{\boldsymbol{K}}}_t + \sigma_0^2\, \boldsymbol{I}\right)^{-1} (\overline{\boldsymbol{y}}_t - \widetilde{\boldsymbol{\mu}}_t) - \frac{1}{2}\log\left|\widetilde{\overline{\boldsymbol{K}}}_t + \sigma_0^2\, \boldsymbol{I}\right| - \frac{p}{2}\log(2\pi)\end{aligned} \tag{10}$$

and

$$\begin{aligned}\log p(\overline{\boldsymbol{y}}_t \mid \overline{\mathbf{W}}_t, H_1) &= \log\int_f p(\overline{\boldsymbol{y}}_t|f)\, p(f|\overline{\boldsymbol{W}}_t)\, df \\ &= -\frac{1}{2}\,\overline{\boldsymbol{y}}_t^\top \left(\overline{\boldsymbol{K}}_t + \sigma_1^2\, \boldsymbol{I}\right)^{-1} \overline{\boldsymbol{y}}_t - \frac{1}{2}\log\left|\overline{\boldsymbol{K}}_t + \sigma_1^2\, \boldsymbol{I}\right| - \frac{p}{2}\log(2\pi)\,.\end{aligned} \tag{11}$$

where

$$\widetilde{\overline{\boldsymbol{K}}}_t = \overline{\boldsymbol{K}}_t - \boldsymbol{K}_t^\top \left(\underline{\boldsymbol{K}}_t + \sigma_0^2\, \boldsymbol{I}\right)^{-1} \boldsymbol{K}_t, \quad \text{and} \quad \widetilde{\boldsymbol{\mu}}_t = \boldsymbol{K}_t^\top \left(\underline{\boldsymbol{K}}_t + \sigma_0^2\, \boldsymbol{I}\right)^{-1} \underline{\boldsymbol{y}}_t\,.$$

Next, use (10) and (11) to write the LR statistic at time t as

$$\mathcal{R}_t = -\overline{\boldsymbol{y}}_t^\top\,(\overline{\boldsymbol{K}}_t + \sigma_1^2\,\boldsymbol{I})^{-1}\overline{\boldsymbol{y}}_t - \log|\overline{\boldsymbol{K}}_t + \sigma_1^2\,\boldsymbol{I}| + (\overline{\boldsymbol{y}}_t - \widetilde{\boldsymbol{\mu}}_t)^\top\,(\widetilde{\overline{\boldsymbol{K}}}_t + \sigma_0^2\,\boldsymbol{I})^{-1}(\overline{\boldsymbol{y}}_t - \widetilde{\boldsymbol{\mu}}_t) + \log|\widetilde{\overline{\boldsymbol{K}}}_t + \sigma_0^2\,\boldsymbol{I}|\,, \tag{12}$$

where $\overline{\boldsymbol{K}}_t \in \mathbb{S}_p(\mathbb{R})$, $\underline{\boldsymbol{K}}_t \in \mathbb{S}_{P_t-p}(\mathbb{R})$, and $\boldsymbol{K}_t \in \mathbb{M}_{P_t-p,p}(\mathbb{R})$ [11].

The change point detection test compares the statistic (12) against a fixed threshold $\mathcal{C}$. When $\mathcal{R}_t > \mathcal{C}$, the test rejects the null and detects a change point. MTGP-LR resets the reward function f_t based on recent observation data following a change point detection; the tactical layer adapts to the new detected market regime.

4.2 Inference error in nonstationary environments

Two types of incorrect inference arise: error of type I, which corresponds to wrong detection of a regime change, and error of type II, which corresponds to overlooking regime changes. Here, we show how the investor fixes the value of the test's threshold $\mathcal{C}$ to control either the probability of error of type I or the probability of error of type II. A low value of the threshold increases the probability of type I error, i.e., the probability of wrong detection of a regime change. On the other hand, a high value of the threshold increases the probability of type II error, i.e., the probability of missing regime changes.

The next results provide thresholds that guarantee a bound on type I and type II error probabilities. For proofs, see Appendix B.2. For convenience, define

$$\widetilde{\boldsymbol{V}}_{H_0,t} = \widetilde{\boldsymbol{K}}_t + \sigma_0^2\,\boldsymbol{I}, \quad \boldsymbol{V}_{H_1,t} = \overline{\boldsymbol{K}}_t + \sigma_1^2\,\boldsymbol{I}\,, \quad \text{and} \quad \boldsymbol{\Lambda}_t = \boldsymbol{V}_{H_1,t}^{-1} - \widetilde{\boldsymbol{V}}_{H_0,t}^{-1}\,. \tag{13}$$

Proposition 1. *Let $\delta_I, \delta_{II} \in (0,1)$. For $j \in \{0,1\}$, λ_{i,H_j} are the eigenvalues of the matrix $\widetilde{\boldsymbol{V}}_{H_j,t}^{1/2}\boldsymbol{\Lambda}_t\widetilde{\boldsymbol{V}}_{H_j,t}^{1/2}$ and let $\mu_{\boldsymbol{H}_j,t} = \mathbb{E}[\mathcal{R}_t|H_j]$. The following threshold guarantees a type I error probability of at most δ_I:*

$$\mathcal{C}_{I,t} = \mu_{H_0,t} + \max\left\{\sqrt{8\log\frac{1}{\delta_I}\left(\sum_i \lambda_{i,H_0}^2 + \widetilde{\boldsymbol{\mu}}_t^\top\,\boldsymbol{V}_{H_1,t}^{-1}\,\boldsymbol{V}_{H_0,t}\,\boldsymbol{V}_{H_1,t}^{-1}\,\widetilde{\boldsymbol{\mu}}_t\right)} - 8\log\delta_I \max_i\{|\lambda_{i,H_0}|\}\right\}, \tag{14}$$

The following threshold guarantees a type II error probability of at most δ_{II}

$$\mathcal{C}_{II,t} = \mu_{H_1,t} - \max\left\{\sqrt{8\,\log\frac{1}{\delta_{II}}\left(\sum_i \lambda_{i,H_1}^2 + \widetilde{\boldsymbol{\mu}}_t^\top\,\widetilde{\boldsymbol{V}}_{H_0,t}^{-1}\,\boldsymbol{V}_{H_1,t}\widetilde{\boldsymbol{V}}_{H_0,t}^{-1}\widetilde{\boldsymbol{\mu}}_t\right)} - 8\log\delta_{II}\max_i\{|\lambda_{i,H_1}|\}\right\}. \tag{15}$$

[11] For $m \in \mathbb{N}$, let $\mathbb{M}_m(\mathbb{R})$ denote the set of all real $m \times m$ matrices, and let $\mathbb{S}_m(\mathbb{R})$ denote the set of all real symmetric $m \times m$ matrices.

Proposition 1 establishes two limits: (i) $\lim_{\delta_{\mathrm{I}}\to 0^+} \mathcal{C}_{\mathrm{I},t} = +\infty$, which drives the probability of false detection to zero; and (ii) $\lim_{\delta_{\mathrm{II}}\to 0^+} \mathcal{C}_{\mathrm{II},t} = -\infty$, which yields perfect sensitivity (no missed change points). Both formulae (14) and (15) define thresholds that guarantee type I and II error probability, so their use is not suited for error probabilities that approach one. In practice and in our experiments, we compute the LR thresholds with

$$\mu_{H_1,t} = \mathbb{E}[\mathcal{R}_t|H_1] = \log|\widetilde{\boldsymbol{V}}_{H_0,t}| - \log|\boldsymbol{V}_{H_1,t}| - \mathrm{Tr}(\boldsymbol{I}) + \mathrm{Tr}(\widetilde{\boldsymbol{V}}_{H_0,t}^{-1}\,\boldsymbol{V}_{H_1,t}) + \tilde{\boldsymbol{\mu}}_t^\top\,\widetilde{\boldsymbol{V}}_{H_0,t}^{-1}\,\tilde{\boldsymbol{\mu}}_t\,,$$
$$\mu_{H_0,t} = \mathbb{E}[\mathcal{R}_t|H_0] = \log|\widetilde{\boldsymbol{V}}_{H_0,t}| - \log|\boldsymbol{V}_{H_1,t}| + \mathrm{Tr}(\boldsymbol{I}) - \mathrm{Tr}(\boldsymbol{V}_{H_1,t}^{-1}\,\widetilde{\boldsymbol{V}}_{H_0,t}) - \tilde{\boldsymbol{\mu}}_t^\top\,\boldsymbol{V}_{H_1,t}^{-1}\,\tilde{\boldsymbol{\mu}}_t\,,$$

and

$$\sum_i \lambda_{i,H_0}^2 = \mathrm{Tr}[\widetilde{\boldsymbol{V}}_{H_0,t}\,\boldsymbol{\Lambda}_t\,\widetilde{\boldsymbol{V}}_{H_0,t}\,\boldsymbol{\Lambda}_t]\,, \quad \text{and} \quad \sum_i \lambda_{i,H_1}^2 = \mathrm{Tr}[\boldsymbol{V}_{H_1,t}\,\boldsymbol{\Lambda}_t\,\boldsymbol{V}_{H_1,t}\,\boldsymbol{\Lambda}_t]\,,$$

where $\boldsymbol{\Lambda}_t$, $\widetilde{\boldsymbol{V}}_{H_0,t}$, and $\boldsymbol{V}_{H_1,t}$ are in (13).[12] We use the method in Arnoldi (1951) to evaluate $\max_i\{\lambda_{i,\boldsymbol{H}_k}\}$.

4.3 Detection speed after a regime change

The previous section shows how to control inference error of the LR test at a fixed time t by constructing dynamic thresholds that bound the conditional type-I and type-II error probabilities. Here, we study the detection delay of the LR test. Specifically, following a change point b_j, we bound the tail probability that the test has not yet triggered after L additional observations. [13]

Recall that the LR $\mathcal{R}_t$ in (9) compares the marginal likelihood of the most recent p observations under H_1 to their predictive likelihood under H_0. The conditional mean of the LR $\mathcal{R}_t$ under H_1 is the Gaussian Kullback–Leibler divergence

$$\mu_{H_1,t} = \mathbb{E}[\mathcal{R}_t \mid H_1, \mathcal{H}_{t-1}] = 2\,\mathrm{KL}\Big(\mathcal{N}(\mathbf{0}, \boldsymbol{V}_{H_1,t}) \,\Big\|\, \mathcal{N}(\tilde{\boldsymbol{\mu}}_t, \widetilde{\boldsymbol{V}}_{H_0,t})\Big) \;\geq\; 0.$$

[12] An implementation of the algorithm is here.

[13] Note that the action–context pair $(a_{t+1}, \mathbf{x}_{t+1})$ is selected from past observations. Thus, we work on the filtration defined, for $t \geq 1$, as $\mathcal{H}_t = \sigma\big(\{(\mathbf{x}_1, a_1, y_1), \ldots, (\mathbf{x}_t, a_t, y_t), \mathbf{x}_{t+1}, a_{t+1}\}\big) = \sigma\big(\{(\mathbf{w}_s, y_s)\}_{s\leq t},\, \mathbf{w}_{t+1}\big)$, where $\mathbf{w}_s = (a_s, \mathbf{x}_s)$. In particular, $\mathcal{R}_t$ is $\mathcal{H}_t$–measurable. Moreover, conditional on $\mathcal{H}_{t-1}$, the quantities $\boldsymbol{V}_{H_1,t}$, $\widetilde{\boldsymbol{V}}_{H_0,t}$, and $\tilde{\boldsymbol{\mu}}_t$ defined in (13), which enter the LR $\mathcal{R}_t$ in (9), are deterministic.

Similarly, the conditional mean under H_0 is

$$\mu_{H_0,t} = \mathbb{E}[\mathcal{R}_t \mid H_0, \mathcal{H}_{t-1}] = -2\,\mathrm{KL}\Big(\mathcal{N}(\widetilde{\boldsymbol{\mu}}_t, \widetilde{\boldsymbol{V}}_{H_0,t}) \,\Big\|\, \mathcal{N}(\mathbf{0}, \boldsymbol{V}_{H_1,t})\Big) \;\le\; 0.$$

Thus, under H_1 the statistic has a nonnegative conditional drift, whereas under H_0 it has a nonpositive conditional drift. To derive guarantees for the detection delay, we assume a uniform positive lower bound on the drift $\mu_{H_1,t}$ of the LR statistic following a change point.

Assumption 2. *If no further regime change occurs after a change point b_j on the interval $\{b_j, \dots, t\}$, then there exists a constant $\mu_{\mathrm{KL}} > 0$ such that*

$$\mu_{H_1,t} = \mathbb{E}[\mathcal{R}_t \mid H_1, \mathcal{H}_{t-1}] \;\ge\; \mu_{\mathrm{KL}} \qquad \textit{a.s.}$$

Deviation envelopes under H_1 Next, we introduce quantities controlling the deviations of the LR $\mathcal{R}_t$ around its conditional mean $\mu_{H_1,t}$ under H_1. Define the precision gap matrix

$$\boldsymbol{G}_t = \boldsymbol{V}_{H_1,t}^{1/2}\, \boldsymbol{\Lambda}_t\, \boldsymbol{V}_{H_1,t}^{1/2}, \qquad \boldsymbol{\Lambda}_t = \boldsymbol{V}_{H_1,t}^{-1} - \widetilde{\boldsymbol{V}}_{H_0,t}^{-1},$$

and let $\{\lambda_{i,t}\}_{i=1}^{p}$ be the eigenvalues of $\boldsymbol{G}_t$. We summarise the spectrum of $\boldsymbol{G}_t$ using

$$\nu_t^2 = 2\sum_{i=1}^{p} \lambda_{i,t}^2 = 2\|\boldsymbol{G}_t\|_F^2, \qquad \zeta_t = 2\max_{1\le i\le p} |\lambda_{i,t}| = 2\|\boldsymbol{G}_t\|, \tag{16}$$

where ν_t^2 and ζ_t serve, respectively, as a proxy for the variance and a scale parameter in sub-exponential tail bounds for centered Gaussian quadratic forms.[14]

We also define the uniform envelopes over all post-change times and all change points,

$$\nu_\star^2 = \sup_{j\in\{1,\dots,\ell-1\}} \sup_{t\ge b_j} \nu_t^2, \qquad \zeta_\star = \sup_{j\in\{1,\dots,\ell-1\}} \sup_{t\ge b_j} \zeta_t, \tag{17}$$

which we assume to be finite.

[14] In practice, $\max_i |\lambda_{i,t}|$ can be approximated using Krylov-Arnoldi iterations, similar the procedure used to compute the eigenvalue bounds in Proposition 1.

Detection-delay parameters and budgets Fix a threshold $\mathcal{C} < \mu_{\mathrm{KL}}$. The gap $\mu_{\mathrm{KL}} - \mathcal{C} > 0$ together with the envelopes $(\nu_\star, \zeta_\star)$ determine the rate parameter

$$\alpha = \min\left\{\frac{(\mu_{\mathrm{KL}} - \mathcal{C})^2}{2\nu_\star^2}, \frac{\mu_{\mathrm{KL}} - \mathcal{C}}{2\zeta_\star}\right\} > 0. \tag{18}$$

For each change point b_j, define the first observation time $\tau_j(\mathcal{C})$ after b_j at which the LR statistic $\mathcal{R}_t$ crosses the threshold $\mathcal{C}$ as

$$\tau_j(\mathcal{C}) = \inf\{t \geq b_j : \mathcal{R}_t \geq \mathcal{C}\}, \qquad \inf \varnothing = +\infty,$$

and define the associated detection delay

$$D_j = \tau_j(\mathcal{C}) - b_j \in \{0, 1, 2, \dots\} \cup \{+\infty\}. \tag{19}$$

Fix $\delta_{\mathrm{det}} \in (0, 1)$ and a horizon $T \geq 1$, define the deterministic delay budget

$$L_{\mathrm{det}}(T) = \left\lceil \frac{1}{\alpha} \log\left(\frac{T^2(T+1)}{2\,\delta_{\mathrm{det}}}\right)\right\rceil - 1, \tag{20}$$

and the event that all changes up to time T are detected within this budget:

$$\mathcal{E}_{\mathrm{det}}(T) = \left\{\max_{1 \leq j \leq \ell(T)-1} D_j \leq L_{\mathrm{det}}(T)\right\}. \tag{21}$$

Theorem 2. *Let Assumption 2 hold. Fix $\mathcal{C} < \mu_{\mathrm{KL}}$ and define α as in (18). Define the detection delay D_j as in (19). First, the detection delay after any fixed change point has an exponential tail. Specifically, for every $j \in \{1, \dots, \ell - 1\}$ and $L \in \mathbb{N}$, the following holds:*

$$\mathbb{P}[D_j > L \mid H_1 \text{ holds on } \{b_j, b_j + 1, \dots, b_j + L\}] \;\leq\; \exp\big(-(L+1)\alpha\big). \tag{22}$$

Second, this tail bound yields a uniform-in-T guarantee over all change points up to time T. Specifically, fix $\delta_{\mathrm{det}} \in (0, 1)$, define $L_{\mathrm{det}}(T)$ as in (20), and define $\mathcal{E}_{\mathrm{det}}(T)$ as in (21). Then, for every $T \geq 1$,

$$\mathbb{P}\big[\mathcal{E}_{\mathrm{det}}(T)\big] \;\geq\; 1 - \frac{2\,\delta_{\mathrm{det}}}{T(T+1)}. \tag{23}$$

Consequently, the uniform event $\mathcal{E}_{\mathrm{det}} = \bigcap_{T\geq 1} \mathcal{E}_{\mathrm{det}}(T)$ *satisfies*

$$\mathbb{P}\big[\mathcal{E}_{\mathrm{det}}\big] \;\geq\; 1 - \delta_{\mathrm{det}}. \tag{24}$$

A proof of Theorem 2 is in Appendix B.3. Theorem 2 establishes an exponential tail bound on the detection delay following a change point. Combined with the dynamic thresholds of Proposition 1, this result yields a uniform-in-T guarantee across all change points for MTGP-LR.

4.4 Type-I error in stationary segments

Proposition 1 controls the type-I error of the LR test at a fixed time t by constructing a threshold $\mathcal{C}$ such that $\mathbb{P}\left[\mathcal{R}_t \geq \mathcal{C} \mid H_0\right]$ is below a prescribed level. MTGP-LR evaluates the LR statistic sequentially and therefore may trigger at any time. To analyse the regret of our algorithm, this section studies the probability that MTGP-LR incorrectly detects a regime change, at least once, within a stationary segment.

Define the stopping time[15]

$$\tau(\{\mathcal{C}_t\}) = \inf\{t \geq 1 :\; \mathcal{R}_t \geq \mathcal{C}_t\}, \qquad \inf \varnothing = +\infty. \tag{25}$$

The next result shows that the one-step type-I error guarantee established in Proposition 1 extends to a bound on the probability of at least one false detection over an entire stationary segment. The result follows from a union bound and yields a uniform-in-T guarantee.

Theorem 3. *Fix* $T \geq 1$ *and let* $\{\delta_t\}_{t=1}^{T} \subset (0,1)$. *Assume that under* H_0 *(no regime change on* $\{1,\dots,T\}$*) the LR statistics* $\{\mathcal{R}_t\}_{t=1}^{T}$ *are well defined,* $\mathcal{G}_t$*–measurable, and that there exists a* $\mathcal{G}_t$*–measurable threshold sequence* $\{\mathcal{C}_t(\delta_t)\}_{t=1}^{T}$ *such that, for every* $t \in \{1,\dots,T\}$*,*

$$\mathbb{P}\big[\mathcal{R}_t \geq \mathcal{C}_t(\delta_t) \;\big|\; H_0\big] \;\leq\; \delta_t. \tag{26}$$

Let $\tau = \tau(\{\mathcal{C}_t(\delta_t)\})$ *be the stopping time defined in* (25)*. First, over the finite horizon* T*, the probability that*

[15] Here, we work on the filtration $\{\mathcal{G}_t\}_{t\geq 1}$,

$$\mathcal{G}_t = \sigma\big(\{(\mathbf{w}_1, y_1), \dots, (\mathbf{w}_t, y_t)\}\big), \qquad \mathbf{w}_s = (a_s, \mathbf{x}_s),$$

and with the threshold sequence $\{\mathcal{C}_t\}_{t\geq 1}$ such that each $\mathcal{C}_t$ is $\mathcal{G}_t$–measurable.

the sequential test triggers at least once under H_0 satisfies

$$\mathbb{P}\left[\tau \leq T \mid H_0\right] \;\leq\; \sum_{t=1}^{T} \delta_t. \tag{27}$$

Second, this bound can be made uniform in T by choosing a specific schedule. Fix $\delta_{\mathrm{fa}} \in (0,1)$ and set

$$\delta_t = \frac{6\,\delta_{\mathrm{fa}}}{\pi^2 t^2}, \qquad t \geq 1. \tag{28}$$

Choose $\mathcal{C}_t = \mathcal{C}_t(\delta_t)$ satisfying (26) for all $t \geq 1$. Then, for every $T \geq 1$,

$$\mathbb{P}\left[\tau \leq T \mid H_0\right] \;\leq\; \delta_{\mathrm{fa}}. \tag{29}$$

In particular, the event

$$\mathcal{E}_{\mathrm{fa}} = \{\tau = +\infty\} \tag{30}$$

satisfies $\mathbb{P}(\mathcal{E}_{\mathrm{fa}} \mid H_0) \geq 1 - \delta_{\mathrm{fa}}$.

Theorem 3 shows that, for any fixed horizon T, the probability that the sequential test triggers at least once under H_0 is bounded by the sum of the levels δ_t used to calibrate the thresholds. Moreover, it provides an explicit summable choice (28) $\sum_{t\geq 1} \delta_t = \delta_{\mathrm{fa}}$, which yields a uniform-in-T bound on $\mathbb{P}\left[\tau \leq T \mid H_0\right]$. Consequently, under H_0, this delivers a high-probability guarantee that no false alarm occurs at any time. A proof of Theorem 3 is in Appendix B.4. In MTGP-LR, (26) is ensured by the threshold sequence in Proposition 1, and Theorem 3 generalises it to a sequential guarantee.

4.5 Regret in regime-switching environments

Here, we derive a high-probability regret bound for MTGP-LR in regime-switching markets. Our method combines three key ingredients: (i) regret guarantees within stationary segments (Theorem 1); (ii) a uniform-in-T bound on the detection delay following a change point (Theorem 2); and (iii) a uniform-in-T bound on the number of false detections within stationary segments (Theorem 3).

Regime-switching and regret Fix a horizon $T \in \mathbb{N}$ and recall that $0 = b_0 < b_1 < \cdots < b_{\ell(T)} = T$ is the (unknown) change point sequence up to time T, so that the latent reward function is constant within each

regime:

$$f_t \equiv f^{(j)} \qquad \text{for all } t \in \{b_{j-1}+1, \dots, b_j\}, \quad j = 1, \dots, \ell(T).$$

At each round $t \in \{1, \dots, T\}$, the agent observes a context $\mathbf{x}_t \in \mathcal{X}$, selects an action $a_t \in \mathcal{A}$, and receives

$$y_t = f_t(a_t, \mathbf{x}_t) + \varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Let $a_t^\star \in \arg\max_{a \in \mathcal{A}} f_t(a, \mathbf{x}_t)$ be an optimal action, and define instantaneous and cumulative regret, as in (1), by

$$r_t = f_t(a_t^\star, \mathbf{x}_t) - f_t(a_t, \mathbf{x}_t), \qquad R(T) = \sum_{t=1}^{T} r_t.$$

After a change point, MTGP-LR may continue to rely on the pre-change posterior until the LR test triggers and the GP is reset. To control the regret accumulated during these post-change delays, we impose a standard assumption of uniformly bounded latent rewards.

Assumption 3. *There exists $B > 0$ such that $|f_t(a, \mathbf{x})| \le B$ for all $t \in \{1, \dots, T\}$ and $(a, \mathbf{x}) \in \mathcal{A} \times \mathcal{X}$. Equivalently, $0 \le r_t \le 2B$ for all $t \in [T]$.*

Recall that $\hat{b}_k$ denotes the $k-$th (random) reset times generated by the LR test. Let $\hat{\ell}(T) + 1$ denote the number of reset times up to time T, so that $\hat{b}_{\hat{\ell}(T)} = T$. The k-th segment is then given by

$$I_k(T) = \{\hat{b}_{k-1} + 1, \dots, \hat{b}_k\}, \qquad k = 1, \dots, \hat{\ell}(T).$$

For each change point b_j, with $j \le \ell(T) - 1$, recall from Section 4.3 the detection time

$$\tau_j(\mathcal{C}) = \inf\{\, t \ge b_j : \mathcal{R}_t \ge \mathcal{C} \,\},$$

and define the corresponding detection delay by $D_j = \tau_j(\mathcal{C}) - b_j$. The next result establishes a regret bound that holds uniformly over all horizons $T \ge 1$.

Theorem 4. *Let Assumption 3 hold. Fix $\delta \in (0, 1)$ and choose $\delta_{\text{gp}}, \delta_{\text{fa}}, \delta_{\text{det}} \in (0, 1)$ such that $\delta_{\text{gp}} + \delta_{\text{fa}} + \delta_{\text{det}} = \delta$. Run MTGP-LR with parameters (C_1, β_t, γ_t) as in Theorem 1, using confidence level δ_{gp} in the definition of β_t. Choose the LR thresholds such that*

- *the uniform false-alarm event $\mathcal{E}_{\text{fa}}$ of Theorem 3 satisfies $\mathbb{P}[\mathcal{E}_{\text{fa}}] \ge 1 - \delta_{\text{fa}}$,*

- *the uniform detection-delay event $\mathcal{E}_{\text{det}}$ from Theorem 2 satisfies $\mathbb{P}[\mathcal{E}_{\text{det}}] \geq 1 - \delta_{\text{det}}$ with delay budget $L_{\text{det}}(T)$.*

Then, with probability at least $1 - \delta$, the following holds simultaneously for all $T \geq 1$:

$$R(T) \;\leq\; \sqrt{C_1\, \ell(T)\, T\, \beta_T\, \gamma_T} \;+\; 2\,\ell(T) \;+\; 2B\,(\ell(T) - 1)\, L_{\text{det}}(T). \tag{31}$$

A proof of Theorem 4 is in Appendix B.5. The first term on the RHS of (31) corresponds to the stationary GP-UCB regret bound in Theorem 1, scaled by the factor $\ell(T)$ to account for $\ell(T)$ stationary regimes. The second term quantifies the additive constants incurred each time MTGP-LR resets. The last term corresponds to the regret accrued during detection delays.

There are multiple methodological choices to reset the reward function following a change point. In environments with low noise, a change in the reward function is quickly detected due to abrupt increase in the LR statistic. On the other hand, in noisy environments, a change in the reward function is not immediately reflected in the LR statistic, and the delay in change point detection is usually longer. The execution of child orders in the tactical layer is carried out in noisy high-frequency markets so the LR test requires a sufficient number of observations from a new regime to detect a change. Thus, MTGP-LR resets the reward function with observation data in the sub-window $\overline{\mathcal{Y}}$ to retrain the kernel hyper-parameters after detecting a change point; see (5). A description of MTGP-LR for the tactical layer is in Algorithm 1.[16]

Computational efficiency GPs need small samples of data to learn the reward functions efficiently. This property is especially useful for algorithmic trading after regime changes because the new reward functions are learned quickly. However, naive MTGP inference scales as $\mathcal{O}(n^3)$ in the number of observations n. This makes an MTGP bandit challenging to deploy when the feature or action spaces are high-dimensional. A variety of scalable GP techniques (e.g., inducing-point and structure-exploiting methods) address this regime, but a full discussion is beyond the scope of this work.

In our numerical experiments, to mitigate these costs by (i) we use the inducing-point method KISS-GP (Kernel Interpolation for Scalable Structured GPs) Wilson and Nickisch (2015), Wilson et al. (2015) to compute predictive means efficiently on large datasets; (ii) we use the low-rank method LOVE (Lanczos Variance Estimates) Pleiss et al. (2018) to approximate the predictive covariance of the context kernel $k^{\mathcal{X}}$ rapidly; and (iii) we restrict updates to a trailing window of reward observations.

[16] The symbol $\#$ denotes the cardinality of a set.

Algorithm 1 MTGP-LR

Input:

Sequence $(\beta_t)_{t\in\mathcal{T}} > 0$ in (8), kernel $k_{\boldsymbol{\theta}}$, number of actions N, number of contextual features m, $P \in \mathbb{N}^\star$, $p \leq P$, time horizon $T > 1$, probability bound $\delta_{\mathrm{I}} \in (0,1)$ or probability bound $\delta_{\mathrm{II}} \in (0,1)$.

Initialise:

$\mathcal{D} \leftarrow \{\}$ is the data set for the current stationary segment; $f \sim \mathrm{MTGP}(\mathbf{0}, \boldsymbol{K}_{\boldsymbol{\theta}})$; $\tilde{\boldsymbol{\beta}}_{1:T} \leftarrow \boldsymbol{\beta}_{1:T}$;

while $t \leq T$ **do**

- Observe the vector of contexts $\mathbf{x}_t \in \mathcal{X}$;
- Compute the vector **UCB** $\leftarrow \left\{\vartheta_{t-1}(a,\mathbf{x}_t) + \tilde{\beta}_t^{1/2}\,\varphi_{t-1}(a,\mathbf{x}_t)\right\}_{a\in\mathcal{A}}$;
- Select action $a_t \leftarrow \arg\max_{a\in\mathcal{A}} \mathbf{UCB}_a$; Observe reward y_t;
- Add $\{y_t, \mathbf{x}_t, a_t\}$ to data set $\mathcal{D}$
- **if** $\#\mathcal{D} \geq P$ **then**
  - Compute LR statistic $\mathcal{R}$ in (12) with data sets $\underline{\mathcal{D}}$ and $\overline{\mathcal{D}}$;
  - Compute threshold $\mathcal{C} = \mathcal{C}_{\mathrm{I}}$ in (14) with the bound δ_{I} for the probability of type I error. Alternatively, compute $\mathcal{C} = \mathcal{C}_{\mathrm{II}}$ in (15) with δ_{II} for the probability of type II error;
  - **if** $\mathcal{R} \geq \mathcal{C}$ **then**
    - $\mathcal{D} \leftarrow \overline{\mathcal{D}}$; $\tilde{\boldsymbol{\beta}}_{t+1:T} \leftarrow \boldsymbol{\beta}_{1:T-t}$;
- Retrain the hyper-parameters $\boldsymbol{\theta}$ with data set $\overline{\mathcal{D}}$ and log marginal likelihood in (see (5)).

5 Numerical experiments

Consider an investor who wishes to execute a large TWAP order of size $\mathcal{Q}$ within the interval $[0,T]$. She tracks the execution schedule prescribed by the strategic layer ρ. The tactical layer optimises child order execution. We consider two practical setups: (i) regime-switching short-term price signals for market order timing, and (ii) regime-switching liquidity for limit order placement.

5.1 Market order timing

In this experiment, the tactical layer implements the TWAP parent order by sequentially executing market orders. MTGP-LR uses price signals to time the execution of these orders, as described below. We use stochastic control tools to derive the optimal tactical layer in closed-form under our simulation setup. Our results demonstrate the superior performance of MTGP-LR compared to control-based strategies in realistic scenarios where brokers misestimate market parameters, a common issue in practice. It also highlights a key practical advantage of our framework: MTGP-LR does not rely on assumptions about the dynamics of

market features, the number of regimes, or the transition probabilities between them.

Simulation setup Consider the problem described in Section 2.2. Here, an investor uses MOs in a LOB to track the TWAP ρ, which we assume independent from the processes defined below. Let $(I_t)_{t\in[0,T]} = (I_t^1, \dots, I_t^N)_{t\in[0,T]}^\intercal$ denote N signals driving short-term changes in the fundamental price $(S_t)_{t\in[0,T]}$. Asset prices transition through regimes such as momentum, mean-reversion, and random walk, and we assume that each signal is tailored to one of these regimes. Thus, all signals provide information about future prices, but at any time t, only one signal is the most relevant. Regime-switching is modeled by a continuous-time Markov chain (CTMC) $(\boldsymbol{\alpha}_t)_{t\in[0,T]}$ with N states. The CTMC describes the probability of the signal driving the price transitioning from I^i to I^j for $(i,j) \in \{1,\dots,N\}^2$. Formally, we define $(\boldsymbol{\alpha}_t)_{t\in[0,T]}$ on a finite state space $\mathcal{E} = \{\boldsymbol{e}^1, \dots, \boldsymbol{e}^N\}$ where $(\boldsymbol{e}^1, \dots, \boldsymbol{e}^N)$ denotes the canonical basis of $\mathbb{R}^N$. The transition rate matrix of the CTMC is $\Theta = (\lambda_{ij})_{i,j\in\{1,\dots,N\}^2}$, where $\lambda_{ii} = -\sum_{i\neq j} \lambda_{ij}$, and for all $i \neq j$, $\lambda_{ij} \geq 0$. For every $i,j \in \{1,\dots,N\}^2$, the value of λ_{ij} is the instantaneous intensity of the CTMC transitioning from state e_i to state e_j. The dynamics of the fundamental price are given by

$$\mathrm{d}S_t = \boldsymbol{\alpha}_t^\top I_t \, dt + \sigma \, dW_t \,,$$

where $(W_t)_{t\in[0,T]}$ is a standard Brownian motion, and $\sigma > 0$ is the volatility parameter. We assume that the CTMC is independent of W_t.

In our simulations, the price impact of investor's MOs is linear and temporary due to her orders walking the book. The execution prices are $\hat{S}_t^\nu = S_t - \kappa\nu_t$, where $\hat{S}_0^\nu = S_0$ and where $\kappa > 0$ is the temporary price impact parameter.

Oracle Next, we derive the oracle tactical layer $(\nu_t)_{t\in[0,T]}$ which assumes the investor correctly specifies all model parameters. The broker's inventory $(q_t)_{t\in[0,T]}$ follows the dynamics:

$$dq_t^\nu = -\nu_t \, dt \,, \quad q_0^\nu = \mathcal{Q} \,.$$

The cash process of the investor $X^\nu = (X_t)_{t\in[0,T]}$ has dynamics

$$dX_t^\nu = \nu_t \, \tilde{S}_t^\nu \, dt \,, \quad X_0^\nu = 0 \,.$$

The investor maximises her expected final wealth subject to a tracking penalty. We define the stopping time $\tau = T \wedge \min\{t : \, q_t^\nu = 0\}$. The performance criterion of the investor using a strategy $\nu \in \mathcal{A}$ is

$$\mathbb{E}\left[X_\tau^\nu + q_\tau^\nu \, (S_\tau - \Phi \, q_\tau^\nu) - \phi \int_0^\tau (\nu_s - \rho_s)^2 \, ds \right] . \tag{32}$$

In the performance criterion (32), the first two terms represent the terminal wealth, composed of the terminal cash X_T and the proceeds from liquidating the remaining inventory at the penalised price $S_T - \Phi \, q_T$. The third term is a running penalty that quantifies the cost of deviating from the target execution trajectory ρ.[17] We use stochastic control tools to solve the problem rigorously in Appendix C. We show that the unique maximiser of (32), for general processes ρ and I, is the tactical layer

$$\nu_t^\star = \frac{q_t}{T - t + \tilde{\phi}} - \frac{1}{2(\phi + \kappa)(T - t + \tilde{\phi})} \mathbb{E}_{t, I, \boldsymbol{\alpha}_t = \boldsymbol{e}^i} \left[\int_t^T (T - s + \tilde{\phi}) \, \boldsymbol{\alpha}_s^\intercal I_s \, ds \right] + \frac{\phi}{(\phi + \kappa)(T - t + \tilde{\phi})} \int_t^T \mathbb{E}_{t,\rho}[\rho_s] \, ds + \frac{\phi}{\phi + \kappa} \rho_t . \tag{33}$$

where $\tilde{\phi} = \frac{\phi + \kappa}{\Phi}$. The first term is TWAP. The second term is a speculative component based on signal predictions. The last two terms track the strategic layer.

Experimental results In our experiment, we set $N = 2$ so the midprice's dynamics are

$$dS_t = \boldsymbol{\alpha}_t^\intercal \begin{pmatrix} I_t^{(1)} \\ I_t^{(2)} \end{pmatrix} dt + \sigma \, dW_t \, . \tag{34}$$

We assume that each signal $I_t^{(i)}$, for $i \in \{1, 2\}$, evolves according to an Ornstein–Uhlenbeck (OU) process

$$dI_t^{(i)} = -\gamma^{(i)} \left(I_t^{(i)} - \theta^{(i)} \right) dt + \sigma^{(i)} \, dW_t^{I^{(i)}} \, , \tag{35}$$

where $\gamma^{(i)} > 0$ is the mean-reversion rate, $\theta^{(i)}$ is the long-term mean level, and $\sigma^{(i)} > 0$ is the volatility coefficient. The processes $W_t^{I^{(1)}}$ and $W_t^{I^{(2)}}$ are independent standard Brownian motions, each also independent

[17] Higher values of the penalty parameter ϕ enforce closer tracking. When the predictive horizon of signals is comparable to T, large deviations may be justified. When it is shorter, as in our experiments, one adjusts the value of ϕ accordingly.

from the Brownian motion W_t^S driving the midprice process.

The consider execution problem consists in liquidating a position $Q_0 = 1000$ shares (selling program). We simulate a trading day ($T = 1$) with timestep $\Delta t = 1/1000$. The midprice follows (34) with $\sigma = 0.1$, and $S_0 = 100$. The CTMC's transition rate matrix is $\Theta = \left[\begin{smallmatrix} -8 & 8 \\ 2 & -2 \end{smallmatrix}\right]$ and the initial regime is $\boldsymbol{\alpha}_t = (1, 0)$. Each predictive signal $I^{(i)}$ evolves as a high-volatility OU process in (35) with $(\gamma^{(1)}, \theta^{(1)}, \sigma^{(1)}) = (0.5, 4, 2.8)$ and $(\gamma^{(2)}, \theta^{(2)}, \sigma^{(2)}) = (0.8, -3, 1.5)$. The temporary impact parameter is $\kappa = 10^{-5}$. In this setting, MTGP-LR employs two arms: executing the child market order either at the beginning or at the end of each time slice.

We compare the following methods: (i) MTGP-LR, (ii) MTGP without change point detection, (iii) a random strategy, (iv) the control-based tactical layer (33) with well-specified model parameters (oracle), and (v) the control-based tactical layer (33) with misspecified model parameters. Misspecification stands for using a wrong transition rate matrix $\tilde{\Theta} = \left[\begin{smallmatrix} -2 & 2 \\ 2 & -2 \end{smallmatrix}\right]$.

In our experiments, we use the squared exponential (SE) kernel

$$k^{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right\}$$

for GPs which defines, a continuous, stationary, and monotonic covariance function, see (6). This kernel imposes a relatively high degree of smoothness on the function space, controlled by the length-scale parameter $\ell > 0$ (see Rasmussen and Ghahramani (2000)). Furthermore, to control the false detection rate, we simulate substantial shifts in the reward function following regime changes (as typically observed in practice) and set a sufficiently large value for the threshold $\mathcal{C}_I$ accordingly.

Method	**P&L**	**Std**
Oracle	6.456	5.378
MTGP-LR	3.443	5.855
MTGP	2.274	6.465
Misspecified	2.657	5.482
Random	−0.743	2.465

Table 1: Mean and standard deviation of P&L improvement to the TWAP across tactical layers over $N = 300$ simulations.

Table 1 reports the mean and standard deviation of the realised P&L over 300 runs for each tactical layer. The oracle tactical layer outperforms all others, because it is the optimal strategy in our simulation setup. Among the remaining strategies, MTGP-LR achieves the best performance. It outperforms MTGP, demonstrating that the change point detection test significantly enhances adaptability to changing market conditions. It also outperforms the control-based tactical layer in scenarios where the broker misestimates model parameters.

5.2 Limit order placement

In this experiment, the tactical layer implements the TWAP parent order by sequentially posting limit orders. MTGP-LR uses price and liquidity signals to optimise the placement of each limit order. Our results demonstrate that MTGP-LR performs well in intractable and high-dimensional settings.

Simulation setup Consider the problem described in Section 2.3. Define a stochastic drift vector $\boldsymbol{\phi}_t = (\phi_{1,t}, \dots, \phi_{N,t})^\intercal$ and an imbalance vector $\boldsymbol{\gamma}_t = (\gamma_{1,t}, \dots, \gamma_{N,t})^\intercal$. In regime i, the ratio of the buy order flow to the sell order flow is $\gamma_{i,t} > 0$, inducing a drift $\phi_{i,t}$ in the midprice. More precisely, if buying pressure dominates, i.e., if $\gamma_{i,t} > 1$, then $\phi_{i,t} > 0$. Conversely, if selling pressure dominates, i.e., if $\gamma_{i,t} < 1$, then $\phi_{i,t} < 0$. Finally, in a balanced market we set $\gamma_i = 1$ and $\phi_i = 0$. The midprice dynamics are given by

$$dS_t = \boldsymbol{\alpha}_t^\intercal \, \boldsymbol{\phi}_t \, dt + \sigma \, dW_t \, .$$

We model each component of the drift and imbalance processes as mean-reverting OU processes.

$$d\phi_{i,t} = \theta_{\phi_i} \left(\bar{\phi}_i - \phi_{i,t} \right) dt + \sigma_{\phi_i} dW_t^{\phi_i}, \quad d\gamma_{i,t} = \theta_{\gamma_i} \left(\bar{\gamma}_i - \gamma_{i,t} \right) dt + \sigma_{\gamma_i} dW_t^{\gamma_i}$$

where $\theta_{\phi_i}, \theta_{\gamma_i} > 0$ are mean-reversion speeds, $\bar{\phi}_i, \bar{\gamma}_i$ denote long-term means, $\sigma_{\phi_i}, \sigma_{\gamma_i}$ are volatility parameters, and W^{ϕ_i}, W^{γ_i} are independent Brownian motions [18].

The tactical layer posts bid and ask limit orders at depths δ^a and δ^b, respectively, to track the TWAP but also to speculate on order-flow imbalances. We model buy (resp. sell) market orders that fill the sell

[18] To align drift with order-flow asymmetry in regime i, we assume $\operatorname{sign}(\bar{\mu}_i) = \operatorname{sign}(\bar{\gamma}_i - 1)$. Thus, when the long-run order-flow ratio is above one (buy pressure), the long-run drift is positive, and vice versa; when $\bar{\gamma}_i = 1$, we have $\bar{\mu}_i = 0$. This economic link is captured in the means; the driving noises are kept independent for parsimony.

(resp. buy) LOs of the market maker with counting processes N^a (resp. N^b). The arrival intensities of N^a and N^b are

$$\Lambda^b(\delta^b) = c\, e^{-\kappa\delta^b} \quad \text{and} \quad \Lambda^a(\delta^a) = c\, e^{-\kappa\,(\boldsymbol{\alpha}_t^\top \boldsymbol{\gamma}_t)\delta^a},$$

where $c > 0$ is a baseline intensity and $\kappa > 0$ models LOB liquidity. In the numerical experiment proposed here, we specialise to a strict agent that places only ask limit orders at depth δ^a to follow ρ.

Oracle We consider $N = 3$ market regimes: sell-heavy, balanced, buy-heavy. Thus, the problem, using stochastic control tools, leads to a Hamilton–Jacobi–Bellman (HJB) equation of dimension 11: six dimensions correspond to the stochastic imbalance and drift signals, one to the mid-price, one to the regime indicator, one to the agent's inventory, one to her cash, and one to time. This high-dimensional PDE is challenging to solve with numerical schemes. Moreover, implementing the feedback control requires estimating the parameters of the OU processes, inferring the prevailing regime accurately, and estimating the transition matrix of the CTMC. Our approach is specifically designed to overcome these challenges.

Experimental results The size of the TWAP parent order is $\mathcal{Q} = 1000$. At the start of each slice, the tactical layer posts LOs, and at the end of each slice, it submits a MO to execute any remaining quantity in the child order. The trading window $[0, T]$, where $T = 1$, is discretised into 250 time steps. The initial mid-price is $S_0 = 100$, and the volatility parameter is $\sigma = 0.1$.

The stochastic drift and imbalance processes are regime-specific OU processes with long-run means

$$\bar{\boldsymbol{\phi}} = (-0.25, 0, 0.25), \qquad \bar{\boldsymbol{\gamma}} = (0.75, 1, 1.25),$$

identical volatility parameters $\sigma_{\phi_i} = \sigma_{\gamma_i} = 0.1$, and identical mean-reversion speeds $\theta_{\phi_i} = \theta_{\gamma_i} = 3$. The generator of the CTMC is $\Theta = \begin{bmatrix} -4.5 & 3.0 & 1.5 \\ 1.5 & -4.5 & 3.0 \\ 3.0 & 1.5 & -4.5 \end{bmatrix}$.

The contextual information used by MTGP-LR includes the previous price change between two consecutive time slices and the quantity executed with the limit order in the previous slice. The action space consists of three arms corresponding to quoting depths $\mathcal{D} = \{0.05, 0.10, 0.20\}$,.

Method	Mean	Std
MTGP-LR	197.018	14.168
MTGP	181.474	20.490
LO-fixed-depth	136.962	4.110

Table 2: Mean and standard deviation of TWAP improvement across tactical layers over $N = 300$ simulations agents.

Table 2 reports the mean and standard deviation of the excess realised P&L relative to a TWAP benchmark, computed over $N = 300$ runs of the tactical layers. The results highlight the benefits of the change point detection mechanism in MTGP-LR, which outperforms the classical MTGP. Furthermore, both bandit-based methods outperform the fixed-depth LO baseline, where "fixed" refers to always posting at depth ($1/\kappa$, with $\kappa = 10$), i.e., the depth that would be optimal in a market with zero volume imbalance.

6 Conclusion

This paper introduced MTGP-LR as a data-driven and model-free tactical layer for brokers to implement trading schedules. The method uses multi-task Gaussian Processes to map market features to the performance of trading decisions, and a LR test to adapt to changing environments. While the experimental focus of this paper is on tactical layers in execution, future work will extend the theoretical framework to be applicable to a wider range of financial decision problems.

A Preliminary results

Here, we state preliminary results which we use in the proofs in B.1 and B.2. The next proposition describes tails of sub-exponential and Gaussian random variables; see Wainwright (2019).

Proposition 2. *A centered random variable X is sub-exponential with parameters (ν, α), where $\nu > 0$ and $\alpha > 0$, if*

$$\mathbb{E}\left[\exp \lambda X\right] \leq \exp\left(\frac{\lambda^2 \nu^2}{2}\right), \qquad \forall \lambda \ \ s.t. \ \ |\lambda| < \frac{1}{\alpha}.$$

Let X be a sub-exponential random variable with parameters (ν, α), and $t > 0$, then

$$\mathbb{P}\left[X \leq -t\right] \leq \exp\left(-\frac{1}{2}\min\left(\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right)\right).$$

Let X be a normally distributed $X \sim \mathcal{N}(0,1)$, and $t \geq 0$, then

$$\mathbb{P}\left[X \geq t\right] \leq \frac{1}{2}\exp\left(-\frac{t^2}{2}\right).$$

Next, we prove the following Bernstein-type inequality for weighted χ^2 fluctuations.

Lemma 1 (Bernstein bound for centered Gaussian quadratic forms)**.** *Let $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and let $\boldsymbol{G} \in \mathbb{R}^{p\times p}$ be symmetric with eigenvalues $\{\lambda_i\}_{i=1}^p$. Define*

$$\nu^2 = 2\sum_{i=1}^p \lambda_i^2 \ = \ 2\|\boldsymbol{G}\|_F^2 \qquad \textit{and} \qquad \zeta = 2\max_{1\leq i\leq p}|\lambda_i| \ = \ 2\|\boldsymbol{G}\|.$$

Then for every $\varepsilon > 0$,

$$\mathbb{P}\big(\boldsymbol{g}^\top \boldsymbol{G}\,\boldsymbol{g} - \mathrm{Tr}(\boldsymbol{G}) \leq -\varepsilon\big) \ \leq \ \exp\left(-\min\left\{\frac{\varepsilon^2}{2\nu^2}, \frac{\varepsilon}{2\zeta}\right\}\right). \tag{36}$$

Proof of Lemma 1. Let $\boldsymbol{G} = \boldsymbol{U}\mathrm{diag}(\lambda_1, \ldots, \lambda_p)\boldsymbol{U}^\top$ be an eigendecomposition and set $\boldsymbol{z} = \boldsymbol{U}^\top \boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. Then

$$\boldsymbol{g}^\top \boldsymbol{G}\,\boldsymbol{g} - \mathrm{Tr}(\boldsymbol{G}) = \sum_{i=1}^p \lambda_i(z_i^2 - 1), \qquad z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Fix $t \in \big(0, 1/\zeta\big)$ (so that $|2t\lambda_i| < 1$ for all i). By Chernoff's method,

$$\mathbb{P}\left(\sum_{i=1}^{p} \lambda_i(z_i^2 - 1) \le -\varepsilon\right) = \mathbb{P}\left(\exp\Big(-t\sum_{i=1}^{p} \lambda_i(z_i^2 - 1)\Big) \ge e^{t\varepsilon}\right) \le e^{-t\varepsilon}\,\mathbb{E}\left[\exp\Big(-t\sum_{i=1}^{p} \lambda_i(z_i^2 - 1)\Big)\right].$$

Independence gives a product. For each i,

$$\mathbb{E}\Big[e^{-t\lambda_i(z_i^2-1)}\Big] = e^{t\lambda_i}\mathbb{E}[e^{-t\lambda_i z_i^2}] = e^{t\lambda_i}(1 + 2t\lambda_i)^{-1/2},$$

valid since $1 + 2t\lambda_i > 0$. Hence

$$\log \mathbb{E}\Big[e^{-t\sum_i \lambda_i(z_i^2-1)}\Big] = \sum_{i=1}^{p}\Big(t\lambda_i - \tfrac{1}{2}\log(1 + 2t\lambda_i)\Big).$$

For $|u| < 1$, the inequality $-\log(1 + u) \le -u + u^2$ holds. Applying it with $u = 2t\lambda_i$ (so $|u| < 1$),

$$t\lambda_i - \tfrac{1}{2}\log(1 + 2t\lambda_i) \le t\lambda_i - \tfrac{1}{2}\Big(2t\lambda_i - (2t\lambda_i)^2\Big) = 2t^2\lambda_i^2.$$

Summing yields

$$\log \mathbb{E}\Big[e^{-t\sum_i \lambda_i(z_i^2-1)}\Big] \le 2t^2\sum_{i=1}^{p}\lambda_i^2 = t^2\nu^2.$$

Therefore, for all $t \in (0, 1/\zeta)$,

$$\mathbb{P}\big(\boldsymbol{g}^\top \boldsymbol{G}\,\boldsymbol{g} - \mathrm{Tr}(\boldsymbol{G}) \le -\varepsilon\big) \le \exp\big(-t\varepsilon + t^2\nu^2\big).$$

Optimizing the RHS over admissible t gives the Bernstein minimum: choose $t = \varepsilon/(2\nu^2)$ if $\varepsilon/(2\nu^2) \le 1/\zeta$, and otherwise $t = 1/\zeta$. This yields (36). □

B Proofs

B.1 Proof of Theorem 1

Assume that the reward function does not change and denote the latent reward function of the bandit by $f_t = f$ for all $t \in \mathcal{T}$. Let $\vartheta_{t-1}(a, \mathbf{x}_t)$ and $\varphi_{t-1}(a, \mathbf{x}_t)$ be the posterior mean and variance for action a conditionally on observations up to time $t-1$, and let $\mathcal{D}_t = \{y_i, \mathbf{x}_i, a_i\}_{i \leq t}$ be the set of reward observations, context observations, and selected actions by the bandit up to time t.

The proof uses similar arguments to those in Krause and Ong (2011) and Dani et al. (2007), which we adapt for multi-task GPs with discrete action space and continuous context space. The proof uses the tail properties of the Gaussian distribution in Proposition 2 to bound the regret r_t of the bandit at time t in terms of the posterior variance, where $r_t = f(a_t^\star, \mathbf{x}_t) - f(a_t, \mathbf{x}_t)$, a_t is the action selected at time t, and $a_t^\star$ is the optimal action at time t; see (1). Next, we show that the regret can be bounded by the information gain, by showing that the posterior variance can be written in terms of the information gain. Finally, we use the result in Srinivas et al. (2009) to show that the information gain associated with the MTGP is bounded, which results in a sub-linear regret for our bandit algorithm.

Regret bounded in terms of posterior variance Let $\delta \in (0, 1)$ be the target probability of the regret bound in (7). Here, we prove that the regret r_t at time t is bounded in terms of the posterior variance.

Let $(\varrho_t)_{t \in \mathcal{T}}$ be a set of values such that

$$\sum_{t \in \mathcal{T}} \varrho_t^{-1} = 1\,, \quad N\varrho_t/\delta > 1\,, \text{ and } N = \#\mathcal{A}\,,$$

let $\beta_t = 2\log(N\varrho_t/\delta)$ for all $t \in \mathcal{T}$ in the UCB formula (2), and let $\mathbf{x}_t \in \mathcal{X}$ be the context at time t. By definition, for $a \in \mathcal{A}$,

$$f(a, \mathbf{x}_t) \sim \mathcal{N}\left(\vartheta_{t-1}(a, \mathbf{x}_t), \varphi_{t-1}^2(a, \mathbf{x}_t)\right),$$

where the mean and the variance are conditioned on $\mathcal{D}_{t-1}$. Use the tail properties of the Gaussian distribution in Proposition 2 to write, for $a \in \mathcal{A}$,

$$\mathbb{P}\left[\left|f(a, \mathbf{x}_t) - \vartheta_{t-1}(a, \mathbf{x}_t)\right| \geq \beta_t^{1/2}\, \varphi_{t-1}(a, \mathbf{x}_t)\right] \leq e^{-\beta_t/2}\,.$$

Next, notice that $\mathcal{A}$ is countable and use the union bound to obtain for all $t \in \mathcal{T}$,

$$\mathbb{P}\left[\left|f\left(a, \mathbf{x}_t\right) - \vartheta_{t-1}\left(a, \mathbf{x}_t\right)\right| \leq \beta_t^{1/2}\, \varphi_{t-1}\left(a, \mathbf{x}_t\right), \quad \forall a \in \mathcal{A}\right] \geq 1 - N e^{-\beta_t/2}\,.$$

Use the values $\varrho_t = \pi^2\, t^2/6$ and the union bound ($\mathcal{T}$ is countable) to obtain

$$\mathbb{P}\left[\left|f\left(a, \mathbf{x}_t\right) - \vartheta_{t-1}\left(a, \mathbf{x}_t\right)\right| \leq \beta_t^{1/2}\, \varphi_{t-1}\left(a, \mathbf{x}_t\right), \quad \forall a \in \mathcal{A},\ \forall t \in \mathcal{T}\right] \geq 1 - \delta\,. \tag{37}$$

The choice for the values of ϱ results in $\beta_t = 2\, \log(N\, t^2\, \pi^2/3\, \delta)$ in the UCB formula in Theorem 1.

Finally, by definition,

$$\vartheta_{t-1}\left(a, \mathbf{x}_t\right) + \beta_t^{1/2}\, \varphi_{t-1}\left(a, \mathbf{x}_t\right) \geq \vartheta_{t-1}\left(a^*, \mathbf{x}_t\right) + \beta_t^{1/2}\, \varphi_{t-1}\left(a^*, \mathbf{x}_t\right) \geq f\left(a^*, \mathbf{x}_t\right),$$

and

$$r_t = f\left(a^*, \mathbf{x}_t\right) - f\left(a_t, \mathbf{x}_t\right) \leq \vartheta_{t-1}\left(a_t, \mathbf{x}_t\right) + \beta_t^{1/2}\, \varphi_{t-1}\left(a_t, \mathbf{x}_t\right) - f\left(a_t, \mathbf{x}_t\right)\,.$$

Thus, use (37) to obtain

$$\begin{aligned} &\mathbb{P}\left[r_t \leq 2\, \beta_t^{1/2}\, \varphi_{t-1}\left(a_t, \mathbf{x}_t\right), \quad \forall t \in \mathcal{T}\right] \\ &= \mathbb{P}\left[\left|f\left(a, \mathbf{x}_t\right) - \vartheta_{t-1}\left(a, \mathbf{x}_t\right)\right| \leq \beta_t^{1/2}\, \varphi_{t-1}\left(a, \mathbf{x}_t\right), \quad \forall a \in \mathcal{A}\,,\ \forall t \in \mathcal{T}\right] \\ &\geq 1 - \delta\,. \end{aligned} \tag{38}$$

Regret bound and information gain Here, we show that the bound in terms of the posterior variance for the regret can be written in terms of the information gain. Let $I\left(\boldsymbol{y}_T; \mathbf{f}_T\right)$ be the information gain at time T from observing $\boldsymbol{y}_T = \{y_1, \ldots, y_T\}$ and $\boldsymbol{f}_T = \{f(a_1, \mathbf{x}_1), \ldots, f(a_T, \mathbf{x}_T)\}$. By definition, write $I\left(\boldsymbol{y}_T; \mathbf{f}_T\right) = \mathrm{H}\left(\boldsymbol{y}_T\right) - \frac{1}{2} \log\left|2\, \pi\, e\, \sigma^2\, \boldsymbol{I}\right|$ where

$$\mathrm{H}\left(\boldsymbol{y}_T\right) = \mathrm{H}\left(\boldsymbol{y}_{T-1}\right) + \mathrm{H}\left(y_T | \boldsymbol{y}_{T-1}\right) = \mathrm{H}\left(\boldsymbol{y}_{T-1}\right) + \frac{1}{2} \log\left(2\, \pi\, e\left(\sigma^2 + \varphi_{t-1}^2\left(\left(a_t, \mathbf{x}_t\right)\right)\right)\right).$$

Thus, by induction, write the information gain in terms of the posterior variance as

$$I\left(\boldsymbol{y}_T; \mathbf{f}_T\right) = \frac{1}{2} \sum_{t=1}^{T} \log\left(1 + \sigma^{-2}\, \varphi_{t-1}^2\left(a_t, \mathbf{x}_t\right)\right) .$$

Next, note that β_t is non-decreasing and use (38) to write

$$\begin{aligned} 4\, \beta_t\, \varphi_{t-1}^2\left(a_t, \mathbf{x}_t\right) &\leq 4\, \beta_T\, \sigma^2 \left(\sigma^{-2}\, \varphi_{t-1}^2\left(a_t, \mathbf{x}_t\right)\right) \\ &\leq 4\, \beta_T\, \sigma^2 C_2 \log\left(1 + \sigma^{-2}\, \varphi_{t-1}^2\left(a_t, \mathbf{x}_t\right)\right) , \end{aligned}$$

where $C_2 = \sigma^{-2} \big/ \log\left(1 + \sigma^{-2}\right) \geq 1$ because $s^2 \leq C_2 \log\left(1 + s^2\right)$ for $s \in [0, \sigma^{-2}]$, and $\sigma^{-2} \varphi_{t-1}^2\left(a_t, \mathbf{x}_t\right) \leq \sigma^{-2} k\left(\left(a_t, \mathbf{x}_t\right), \left(a_t, \mathbf{x}_t\right)\right) \leq \sigma^{-2}$. Finally, use $C_1 = 8\, \sigma^2\, C_2$ to obtain

$$\mathbb{P}\left[\sum_{t=1}^{T} r_t^2 \leq \beta_T\, C_1\, I\left(\boldsymbol{y}_T; \mathbf{f}_T\right) \leq C_1\, \beta_T\, \gamma_{T, \mathbf{x}_{1:T}} , \quad \forall T \geq 1\right] \geq 1 - \delta\,. \tag{39}$$

The first result of Theorem 1 in (7) is obtained with an application of the Cauchy–Schwarz inequality to the sum of squared regrets.

Bounding the information gain To obtain the sub-linear regret bound for the regret, we show that the information gain in the probability in (39) is bounded. First, note that

$$\gamma_T = \max_{A \subset \mathcal{A} : \# A = T} I\left(\boldsymbol{y}_{A, \mathbf{x}_{1:T}}; \mathbf{f}_{A, \mathbf{x}_{1:T}}\right) = \max_{\substack{A \subset \mathcal{A} : \# A = T \\ \tilde{\mathbf{x}}_{1:T} \subset \mathcal{X} \\ \tilde{\mathbf{x}}_{1:T} = \mathbf{x}_{1:T}}} I\left(\boldsymbol{y}_{A, \tilde{\mathbf{x}}_{1:T}}; \mathbf{f}_{A, \tilde{\mathbf{x}}_{1:T}}\right) \leq \max_{\substack{A \subset \mathcal{A} : \# A = T \\ \tilde{\mathbf{x}}_{1:T} \subset \mathcal{X}}} I\left(\boldsymbol{y}_{A, \tilde{\mathbf{x}}_{1:T}}; \mathbf{f}_{A, \tilde{\mathbf{x}}_{1:T}}\right) = \gamma_T^{\mathcal{A} \times \mathcal{X}},$$

where $\gamma_T^{\mathcal{A} \times \mathcal{X}}$ is the information gain associated with an MTGP that is not conditioned on the context observations $\mathbf{x}_{1:T}$.

Next, note that the kernel $k^{\mathcal{A}}$ is of rank at most N because $\# \mathcal{A} = N$ and use the results in Krause and Ong (2011) for the composite kernel $k = k^{\mathcal{A}} \times k^{\mathcal{X}}$ to write

$$\gamma_T^{\mathcal{A} \times \mathcal{X}} \leq N \gamma_T^{\mathcal{X}} + N \log(T)\,, \tag{40}$$

where

$$\gamma_T^{\mathcal{X}} = \max_{\mathbf{x}\subset\mathcal{X}:\#\mathbf{x}=T} I\left(\tilde{\boldsymbol{y}}_{\mathbf{x}}; \tilde{\mathbf{f}}_{\mathbf{x}}\right) \quad \text{and} \quad \tilde{f} \sim \mathcal{GP}\left(0, k^{\mathcal{X}}\right).$$

The inequality in (40) provides a bound for the information gain $\gamma_T^{\mathcal{A}\times\mathcal{X}}$ of the MTGP's kernel conditioned on the context $\mathbf{x}_{1:T}$ in terms of the information gain $\gamma_T^{\mathcal{X}}$ of the kernel $k^{\mathcal{X}}$ for the context domain. The value of $\gamma_T^{\mathcal{X}}$ is thoroughly studied in the literature. Srinivas et al. (2009) provides bounds for $\gamma_T^{\mathcal{X}}$ for the usual kernels, i.e., the linear, the squared exponential, and the Matérn kernels. These bounds provide a sub-linear bound for the regret in Theorem 1. □

B.2 Proof of Proposition 1

We prove Proposition 1 using arguments similar to those in Caldarelli et al. (2022) which we adapt to our specific hypothesis test. Here, we focus on the threshold (14), the threshold (15) follows similar arguments. First, we show that bounding the LR statistic (9) is equivalent to bounding a sub-exponential random variable. Next, we use the tail properties of sub-exponential random variables in Proposition 2 to obtain the bound on the probability of the error of type I (wrong detection).

Let $\mathcal{T}_I = \mathbb{E}\left[\mathcal{R}|H_0\right] + t$ and note that the test $\mathcal{R} \geq \mathcal{T}_I$ is equivalent to the test

$$\begin{aligned}(\overline{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}})^\top \left[\left(\overline{\boldsymbol{K}} + \sigma_{H_1}^2 \boldsymbol{I}\right)^{-1} - \left(\widetilde{\boldsymbol{K}} + \sigma_{H_0}^2 \boldsymbol{I}\right)^{-1}\right](\overline{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}}) + 2\,\widetilde{\boldsymbol{\mu}}^\top \left(\overline{\boldsymbol{K}} + \sigma_{H_1}^2 \boldsymbol{I}\right)^{-1}(\overline{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}}) \\ \leq \mathbb{E}\left[(\overline{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}})^\top \left[\left(\overline{\boldsymbol{K}} + \sigma_{H_1}^2 \boldsymbol{I}\right)^{-1} - \left(\widetilde{\boldsymbol{K}} + \sigma_{H_0}^2 \boldsymbol{I}\right)^{-1}\right](\overline{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}})\right] - t\,.\end{aligned} \tag{41}$$

Define the random variable Z as

$$Z = (\overline{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}})^\top \left[\left(\overline{\boldsymbol{K}} + \sigma_{H_1}^2 \boldsymbol{I}\right)^{-1} - \left(\widetilde{\boldsymbol{K}} + \sigma_{H_0}^2 \boldsymbol{I}\right)^{-1}\right](\overline{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}}) + 2\,\widetilde{\boldsymbol{\mu}}^\top \left(\overline{\boldsymbol{K}} + \sigma_{H_1}^2 \boldsymbol{I}\right)^{-1}(\overline{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}})\,, \tag{42}$$

and use (41) to obtain that $\mathcal{R} \geq \mathcal{T}_I$ is equivalent to $Z - \mathbb{E}\left[Z|H_0\right] \leq -t$.

Next, we prove that Z is sub-exponential. The null hypothesis H_0 assumes that the rewards $\overline{\boldsymbol{y}}$ from the

sub-window $\overline{\mathcal{Y}}$ are Gaussian and we write $\overline{\boldsymbol{y}} \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\overline{\boldsymbol{K}}} + \sigma_{H_0}^2 \boldsymbol{I}\right)$. Define the following matrices

$$\begin{cases} \boldsymbol{\Lambda} & = \left(\overline{\boldsymbol{K}} + \sigma_{H_1}^2 \boldsymbol{I}\right)^{-1} - \left(\widetilde{\overline{\boldsymbol{K}}} + \sigma_{H_0}^2 \boldsymbol{I}\right)^{-1} , \\ \widetilde{\boldsymbol{V}}_{H_0,t} & = \widetilde{\overline{\boldsymbol{K}}} + \sigma_{H_0}^2 \boldsymbol{I} , \\ \boldsymbol{V}_{H_1,t} & = \overline{\boldsymbol{K}} + \sigma_{H_1}^2 \boldsymbol{I} , \end{cases}$$

and let $\boldsymbol{y} = \widetilde{\boldsymbol{V}}_{H_0,t}^{-\frac{1}{2}} (\overline{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}})$. Thus, write the random variable Z in (42) as

$$Z = \boldsymbol{y}\, \widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}}\, \boldsymbol{\Lambda}\, \widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}}\, \boldsymbol{y} + 2\, \widetilde{\boldsymbol{\mu}}^\top\, \boldsymbol{V}_{H_1,t}^{-1}\, \boldsymbol{V}_{H_0}^{\frac{1}{2}}\, \boldsymbol{y} . \tag{43}$$

Write the eigen-decomposition of $\widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}}\, \boldsymbol{\Lambda}\, \widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}}$ as

$$\boldsymbol{y}\, \widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}}\, \boldsymbol{\Lambda}\, \widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}}\, \boldsymbol{y} = \sum_i \lambda_{i,H_0} u_i ,$$

where $\{\lambda_{i,H_0}\}_{i\in\{1,\dots,\#\overline{\mathcal{Y}}\}}$ are the eigenvalues of $\widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}} \boldsymbol{\Lambda} \widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}}$ and $\{u_i\}_{i\in\{1,\dots,\#\overline{\mathcal{Y}}\}}$ are independent χ_1^2 variables. Use the result in Caldarelli et al. (2022) to obtain that, for $i \in \{1, \dots, \#\overline{\mathcal{Y}}\}$, the random variable $\lambda_{i,H_0} u_i - \mathbb{E}\left[\lambda_{i,H_0} u_i\right]$ is sub-exponential with parameters $(2\,|\lambda_{i,H_0}|\,,\, 4|\lambda_{i,H_0}|)$. Furthermore, note that

$$2\, \widetilde{\boldsymbol{\mu}}^\top\, \boldsymbol{V}_{H_1,t}^{-1}\, \widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}}\, \boldsymbol{y} \sim \mathcal{N}\left(0, 4\, \widetilde{\boldsymbol{\mu}}^\top\, \boldsymbol{V}_{H_1,t}^{-1}\, \widetilde{\boldsymbol{V}}_{H_0,t}\, \boldsymbol{V}_{H_1,t}^{-1}\, \widetilde{\boldsymbol{\mu}}\right) ,$$

so $2\, \widetilde{\boldsymbol{\mu}}^\top\, \boldsymbol{V}_{H_1,t}^{-1}\, \widetilde{\boldsymbol{V}}_{H_0,t}^{\frac{1}{2}}\, \boldsymbol{y}$ is also sub-exponential with parameters $\left(2\sqrt{\widetilde{\boldsymbol{\mu}}^\top\, \boldsymbol{V}_{H_1,t}^{-1}\, \widetilde{\boldsymbol{V}}_{H_0,t}\, \boldsymbol{V}_{H_1,t}^{-1}\, \widetilde{\boldsymbol{\mu}}}\,,\, 0\right)$. Conclude that the random variable Z in (43) is sub-exponential with parameters[19]

$$\left(2\sqrt{\sum_{i\in\{1,\dots,\#\overline{\mathcal{Y}}\}} \lambda_{i,H_0}^2 + \widetilde{\boldsymbol{\mu}}^\top\, \boldsymbol{V}_{H_1,t}^{-1}\, \widetilde{\boldsymbol{V}}_{H_0,t}\, \boldsymbol{V}_{H_1,t}^{-1}\, \widetilde{\boldsymbol{\mu}}}\,,\, \max_{i\in\{1,\dots,\#\overline{\mathcal{Y}}\}} \{4\,|\lambda_{i,H_0}|\}\right) .$$

[19] In the remainder of the proof, we drop the notation $i \in \{1, \dots, \#\overline{\mathcal{Y}}\}$.

Next, use Proposition 2 to write for all $t > 0$

$$\mathbb{P}\left(Z - \mathbb{E}\left[Z|H_0\right] \leq -t\right) \leq$$
$$\exp\left(-\frac{1}{2}\min\left(\frac{t^2}{4\left(\sum_i \lambda_{i,H_0}^2 + \widetilde{\boldsymbol{\mu}}^\top \boldsymbol{V}_{H_1,t}^{-1}\widetilde{\boldsymbol{V}}_{H_0,t}\boldsymbol{V}_{H_1,t}^{-1}\widetilde{\boldsymbol{\mu}}\right)}, \frac{t}{\max_i\{4|\lambda_{i,H_0}|\}}\right)\right). \tag{44}$$

Conclude that the probability that the term on the right-hand side of the inequality (44) is less than δ when

$$t \geq \max\left\{\sqrt{8\log\left(\frac{1}{\delta}\right)\left(\sum_i \lambda_{i,H_0}^2 + \widetilde{\boldsymbol{\mu}}^\top \boldsymbol{V}_{H_1,t}^{-1}\widetilde{\boldsymbol{V}}_{H_0,t}\boldsymbol{V}_{H_1,t}^{-1}\widetilde{\boldsymbol{\mu}}\right)}, 8\log\left(\frac{1}{\delta}\right)\max_i\{|\lambda_{i,H_0}|\}\right\}.$$

□

B.3 Proof of Theorem 2

Throughout, fix a threshold $\mathcal{C} < \mu_{\mathrm{KL}}$. We now prove the two claims of Theorem B.3 in turn.

Proof of Theorem 2(i) Fix $j \in \{1, \ldots, \ell - 1\}$ and $L \in \mathbb{N}$. Let

$$\mathsf{Post}(j, L) = \{H_1 \text{ holds on } \{b_j, b_j + 1, \ldots, b_j + L\}\},$$

i.e., the regime has switched at b_j and remains unchanged up to time $b_j + L$.

For each $t \geq b_j$, condition on $\mathcal{H}_{t-1}$. Under $\mathsf{Post}(j, L)$, the distribution of the most recent p-vector $\overline{\boldsymbol{y}}_t$ under H_1 satisfies

$$\overline{\boldsymbol{y}}_t \mid (H_1, \mathcal{H}_{t-1}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_{H_1,t}),$$

and $\boldsymbol{V}_{H_1,t}, \widetilde{\boldsymbol{V}}_{H_0,t}, \widetilde{\boldsymbol{\mu}}_t$ are $\mathcal{H}_{t-1}$-measurable, hence deterministic given $\mathcal{H}_{t-1}$. An expansion of the log-likelihood ratio shows that the random fluctuation of $\mathcal{R}_t$ around its conditional mean is a centered quadratic form:

$$\mathcal{R}_t - \mu_{H_1,t} = \overline{\boldsymbol{y}}_t^\top \boldsymbol{\Lambda}_t \overline{\boldsymbol{y}}_t - \mathbb{E}\left[\overline{\boldsymbol{y}}_t^\top \boldsymbol{\Lambda}_t \overline{\boldsymbol{y}}_t \mid H_1, \mathcal{H}_{t-1}\right], \qquad \boldsymbol{\Lambda}_t = \boldsymbol{V}_{H_1,t}^{-1} - \widetilde{\boldsymbol{V}}_{H_0,t}^{-1}. \tag{45}$$

Write $\overline{\boldsymbol{y}}_t = \boldsymbol{V}_{H_1,t}^{1/2}\boldsymbol{g}$ with $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ (conditionally on $\mathcal{H}_{t-1}$). Then (45) becomes

$$\mathcal{R}_t - \mu_{H_1,t} = \boldsymbol{g}^\top \boldsymbol{G}_t\, \boldsymbol{g} - \mathrm{Tr}(\boldsymbol{G}_t), \qquad \boldsymbol{G}_t = \boldsymbol{V}_{H_1,t}^{1/2} \boldsymbol{\Lambda}_t \boldsymbol{V}_{H_1,t}^{1/2},$$

where $\boldsymbol{G}_t$ is $\mathcal{H}_{t-1}$-measurable. Applying Lemma 1 conditionally on $\mathcal{H}_{t-1}$ gives, for every $\varepsilon > 0$,

$$\mathbb{P}(\mathcal{R}_t \le \mu_{H_1,t} - \varepsilon \,|\, H_1, \mathcal{H}_{t-1}) \le \exp\left(- \min\left\{ \frac{\varepsilon^2}{2\nu_t^2}, \frac{\varepsilon}{2\zeta_t} \right\} \right), \tag{46}$$

where ν_t^2, ζ_t are defined from the spectrum of $\boldsymbol{G}_t$ as in (16).

Now set $\varepsilon = \mu_{H_1,t} - \mathcal{C}$. On $\mathsf{Post}(j, L)$ and for all $t \in \{b_j, \dots, b_j + L\}$, Assumption 2 implies $\mu_{H_1,t} \ge \mu_{\mathrm{KL}}$ a.s., hence $\varepsilon \ge \mu_{\mathrm{KL}} - \mathcal{C} > 0$. Using also $\nu_t^2 \le \nu_\star^2$ and $\zeta_t \le \zeta_\star$ by definition of the envelopes (17), (46) yields the uniform one-step bound

$$\mathbb{P}(\mathcal{R}_t < \mathcal{C} \,|\, H_1, \mathcal{H}_{t-1}) \le \exp(-\alpha)\,, \qquad t \in \{b_j, \dots, b_j + L\}, \tag{47}$$

where α is defined in (18).

Define events $A_t = \{\mathcal{R}_t < \mathcal{C}\}$. The event $\{D_j > L\}$ is exactly

$$\{D_j > L\} = \bigcap_{t=b_j}^{b_j+L} A_t,$$

since $D_j > L$ means no crossing of the threshold occurs at any time $t \in \{b_j, \dots, b_j + L\}$. We now bound the probability of this intersection by iterated conditioning. Using that A_t is $\mathcal{H}_t$-measurable, hence $A_{b_j}, \dots, A_t$ are also $\mathcal{H}_t$-measurable, we have

$$\mathbb{P}\left(\bigcap_{t=b_j}^{b_j+L} A_t \,\middle|\, \mathsf{Post}(j, L) \right) = \mathbb{E}\left[\mathbf{1}_{\cap_{t=b_j}^{b_j+L-1} A_t}\, \mathbb{P}\big(A_{b_j+L} \mid \mathcal{H}_{b_j+L-1}, \mathsf{Post}(j, L)\big) \,\middle|\, \mathsf{Post}(j, L) \right].$$

On $\mathsf{Post}(j, L)$, the conditional bound (47) applies at time $b_j + L$, and conditioning on the additional event $\cap_{t=b_j}^{b_j+L-1} A_t$ does not change the applicability since this event is $\mathcal{H}_{b_j+L-1}$-measurable. Hence

$$\mathbb{P}\big(A_{b_j+L} \mid \mathcal{H}_{b_j+L-1}, \mathsf{Post}(j, L)\big) \le e^{-\alpha} \qquad \text{a.s. on } \mathsf{Post}(j, L),$$

and therefore

$$\mathbb{P}\left(\bigcap_{t=b_j}^{b_j+L} A_t \,\middle|\, \mathsf{Post}(j,L)\right) \le e^{-\alpha}\,\mathbb{P}\left(\bigcap_{t=b_j}^{b_j+L-1} A_t \,\middle|\, \mathsf{Post}(j,L)\right).$$

Iterating this argument $L+1$ times gives

$$\mathbb{P}\left(\bigcap_{t=b_j}^{b_j+L} A_t \,\middle|\, \mathsf{Post}(j,L)\right) \le e^{-(L+1)\alpha}.$$

Since $\{D_j > L\} = \cap_{t=b_j}^{b_j+L} A_t$, this proves (22).

Proof of Theorem 2(ii) Fix $T \ge 1$ and $\delta_{\mathrm{det}} \in (0,1)$, and let $L_{\mathrm{det}}(T)$ be as in (20). By construction of $L_{\mathrm{det}}(T)$,

$$(L_{\mathrm{det}}(T)+1)\alpha \;\ge\; \log\!\left(\frac{T^2(T+1)}{2\delta_{\mathrm{det}}}\right), \qquad \text{equivalently} \qquad e^{-(L_{\mathrm{det}}(T)+1)\alpha} \;\le\; \frac{2\delta_{\mathrm{det}}}{T^2(T+1)}. \tag{48}$$

For each $j \in \{1,\dots,\ell(T)-1\}$, apply part (i) with $L = L_{\mathrm{det}}(T)$ to obtain

$$\mathbb{P}(D_j > L_{\mathrm{det}}(T)) \;\le\; e^{-(L_{\mathrm{det}}(T)+1)\alpha} \;\le\; \frac{2\delta_{\mathrm{det}}}{T^2(T+1)},$$

where we used (48). Since there are at most $\ell(T)-1 \le T-1 < T$ change points up to time T, a union bound yields

$$\mathbb{P}(\mathcal{E}_{\mathrm{det}}(T)^c) = \mathbb{P}\left(\max_{1\le j\le \ell(T)-1} D_j > L_{\mathrm{det}}(T)\right) \le \sum_{j=1}^{\ell(T)-1} \mathbb{P}(D_j > L_{\mathrm{det}}(T)) \le T\cdot\frac{2\delta_{\mathrm{det}}}{T^2(T+1)} = \frac{2\delta_{\mathrm{det}}}{T(T+1)}.$$

This is (23).

Finally, for the uniform event $\mathcal{E}_{\mathrm{det}} = \cap_{T\ge1}\mathcal{E}_{\mathrm{det}}(T)$,

$$\mathbb{P}(\mathcal{E}_{\mathrm{det}}^c) = \mathbb{P}\left(\bigcup_{T\ge1}\mathcal{E}_{\mathrm{det}}(T)^c\right) \le \sum_{T\ge1}\mathbb{P}(\mathcal{E}_{\mathrm{det}}(T)^c) \le \sum_{T\ge1}\frac{2\delta_{\mathrm{det}}}{T(T+1)} = \delta_{\mathrm{det}},$$

using $\sum_{T\geq 1} \frac{2}{T(T+1)} = 1$. This proves (24). $\square$

B.4 Proof of Theorem 3

Throughout the proof we work under H_0, i.e., no regime change on the time interval under consideration. In this case, the LR statistics $\{\mathcal{R}_t\}$ are well defined and $\mathcal{G}_t$–measurable, and the thresholds $\{\mathcal{C}_t\}$ are $\mathcal{G}_t$–measurable by construction. We now prove the two claims of Theorem 3 in turn.

Proof of Theorem 3(i). Fix $T \geq 1$ and set $\mathcal{C}_t = \mathcal{C}_t(\delta_t)$. By definition of the stopping time $\tau = \inf\{t \geq 1 : \mathcal{R}_t \geq \mathcal{C}_t\}$, we have the set identity

$$\{\tau \leq T\} = \bigcup_{t=1}^{T} \{\tau = t\}.$$

Moreover, for each $t \in \{1, \ldots, T\}$, the event $\{\tau = t\}$ implies that the threshold is crossed at time t, i.e.

$$\{\tau = t\} \subseteq \{\mathcal{R}_t \geq \mathcal{C}_t\}.$$

Indeed, $\tau = t$ means (by definition of $\inf$) that $\mathcal{R}_t \geq \mathcal{C}_t$ and that no crossing occurred at earlier times. Therefore,

$$\{\tau \leq T\} = \bigcup_{t=1}^{T} \{\tau = t\} \subseteq \bigcup_{t=1}^{T} \{\mathcal{R}_t \geq \mathcal{C}_t\}.$$

Applying the union bound yields

$$\mathbb{P}(\tau \leq T \mid H_0) \leq \sum_{t=1}^{T} \mathbb{P}(\mathcal{R}_t \geq \mathcal{C}_t \mid H_0). \tag{49}$$

By the one-step calibration assumption (26), for each $t \leq T$,

$$\mathbb{P}(\mathcal{R}_t \geq \mathcal{C}_t(\delta_t) \mid H_0) \leq \delta_t.$$

Substituting into (49) gives

$$\mathbb{P}(\tau \leq T \mid H_0) \leq \sum_{t=1}^{T} \delta_t,$$

which is (27).

Proof of Theorem 3(ii). Fix $\delta_{\mathrm{fa}} \in (0,1)$ and let δ_t be as in (28). Then for every $T \geq 1$, Theorem 3(i) yields

$$\mathbb{P}(\tau \leq T \mid H_0) \leq \sum_{t=1}^{T} \delta_t \leq \sum_{t=1}^{\infty} \delta_t = \sum_{t=1}^{\infty} \frac{6\,\delta_{\mathrm{fa}}}{\pi^2 t^2} = \delta_{\mathrm{fa}},$$

where we used $\sum_{t=1}^{\infty} t^{-2} = \pi^2/6$.

Next, note that the events $\{\tau \leq T\}$ form an increasing sequence in T, and

$$\{\tau < \infty\} = \bigcup_{T=1}^{\infty} \{\tau \leq T\}.$$

By continuity of probability for increasing events,

$$\mathbb{P}(\tau < \infty \mid H_0) = \lim_{T \to \infty} \mathbb{P}(\tau \leq T \mid H_0) \leq \delta_{\mathrm{fa}}.$$

Equivalently, $\mathbb{P}(\tau = +\infty \mid H_0) \geq 1 - \delta_{\mathrm{fa}}$, which is (30), and hence also implies (29) for every finite T.

□

B.5 Proof of Theorem 4

Fix $\delta_{\mathrm{gp}}, \delta_{\mathrm{fa}}, \delta_{\mathrm{det}} \in (0,1)$ and set $\delta = \delta_{\mathrm{gp}} + \delta_{\mathrm{fa}} + \delta_{\mathrm{det}}$. We work on a probability space supporting the latent functions and the observation noise, and all events below are defined with respect to this joint law.

Step 1: notation. Let $\mathcal{E}_{\mathrm{gp}}$ denote the high-probability event from Theorem 1 with confidence level δ_{gp}, i.e., the event on which the stationary regret bound of Theorem 1 holds simultaneously for all horizons. By Theorem 1,

$$\mathbb{P}(\mathcal{E}_{\mathrm{gp}}) \;\geq\; 1 - \delta_{\mathrm{gp}}.$$

Let $\mathcal{E}_{\mathrm{fa}}$ be the uniform false-alarm event from Section 4.4 (Theorem 3), calibrated so that

$$\mathbb{P}(\mathcal{E}_{\mathrm{fa}}) \;\geq\; 1 - \delta_{\mathrm{fa}}.$$

Let $\mathcal{E}_{\mathrm{det}}$ be the uniform detection-delay event from Section 4.3 (Theorem 2), calibrated so that

$$\mathbb{P}(\mathcal{E}_{\mathrm{det}}) \;\geq\; 1 - \delta_{\mathrm{det}}.$$

By a union bound,

$$\mathbb{P}(\mathcal{E}_{\mathrm{gp}} \cap \mathcal{E}_{\mathrm{fa}} \cap \mathcal{E}_{\mathrm{det}}) \;\geq\; 1 - \delta. \tag{50}$$

We prove that on $\mathcal{E}_{\mathrm{gp}} \cap \mathcal{E}_{\mathrm{fa}} \cap \mathcal{E}_{\mathrm{det}}$, the regret bound (31) holds simultaneously for all $T \geq 1$.

Fix an arbitrary horizon $T \geq 1$. Let $0 = b_0 < b_1 < \cdots < b_{\ell(T)} = T$ be the (unknown) change points up to time T and $\ell(T)$ the number of true regimes up to time T. Let $0 = \hat{b}_0 < \hat{b}_1 < \cdots < \hat{b}_{\hat{\ell}(T)} = T$ be the (random) reset times of MTGP-LR up to time T (i.e. times at which the LR test triggers and the GP posterior is reset), and define the segments

$$I_k(T) = \{\hat{b}_{k-1} + 1, \ldots, \hat{b}_k\}, \qquad k = 1, \ldots, \hat{\ell}(T).$$

Let $L_k(T) = |I_k(T)| = \hat{b}_k - \hat{b}_{k-1}$ so that $\sum_{k=1}^{\hat{\ell}(T)} L_k(T) = T$.

Step 2: regret decomposition. For each segment $I_k(T)$, define the index of the first change point strictly after $\hat{b}_{k-1}$ by

$$j_k = \min\{j \in \{1, \ldots, \ell(T) - 1\} : b_j > \hat{b}_{k-1}\},$$

with the convention $j_k = +\infty$ if the set is empty. Define the *stationary prefix length* of segment k by

$$S_k(T) = \begin{cases} \min\{\hat{b}_k, b_{j_k}\} - \hat{b}_{k-1}, & \text{if } j_k < +\infty, \\ \hat{b}_k - \hat{b}_{k-1}, & \text{if } j_k = +\infty. \end{cases}$$

Then $0 \leq S_k(T) \leq L_k(T)$ and, by construction, no true change occurs on the time set

$$J_k(T) = \{\hat{b}_{k-1} + 1, \ldots, \hat{b}_{k-1} + S_k(T)\} \subseteq I_k(T),$$

so that $(f_t)_{t \in J_k(T)}$ is constant (equal to some $f^{(j)}$). If $S_k(T) < L_k(T)$, then a true change occurs at time $b_{j_k} = \hat{b}_{k-1} + S_k(T)$ and the remaining times in $I_k(T)$ (after b_{j_k}) come after the change point but before reset.

Define the set of delay times up to T by

$$\mathcal{M}(T) = \bigcup_{k=1}^{\hat{\ell}(T)} \left(I_k(T) \setminus J_k(T) \right).$$

This is the set of rounds that occur after a change point but before the subsequent LR-trigger reset. Using the decomposition $[T] = \bigcup_{k=1}^{\hat{\ell}(T)} I_k(T)$ (disjoint union), we obtain the regret decomposition

$$R(T) = \sum_{t=1}^{T} r_t = \sum_{k=1}^{\hat{\ell}(T)} \sum_{t \in J_k(T)} r_t \; + \sum_{t \in \mathcal{M}(T)} r_t. \tag{51}$$

Step 3: Bounding delay regret. By Assumption 3, $0 \le r_t \le 2B$ for all t, hence

$$\sum_{t \in \mathcal{M}(T)} r_t \;\le\; 2B \, |\mathcal{M}(T)|. \tag{52}$$

We now relate $|\mathcal{M}(T)|$ to the detection delays D_j defined in Section 4.3. For each change point b_j ($j = 1, \dots, \ell(T) - 1$), let $\tau_j(\mathcal{C}) = \inf\{t \ge b_j : \mathcal{R}_t \ge \mathcal{C}\}$ and $D_j = \tau_j(\mathcal{C}) - b_j$. By definition of $\tau_j(\mathcal{C})$, MTGP-LR can make at most D_j decisions after time b_j before the first detection occurs. Therefore,

$$|\mathcal{M}(T)| \;\le\; \sum_{j=1}^{\ell(T)-1} D_j.$$

On the event $\mathcal{E}_{\text{det}}$, Theorem 2 yields $\max_{1 \le j \le \ell(T)-1} D_j \le L_{\text{det}}(T)$, hence

$$\sum_{j=1}^{\ell(T)-1} D_j \;\le\; (\ell(T) - 1) \, L_{\text{det}}(T). \tag{53}$$

Combining (52)–(53) gives, on $\mathcal{E}_{\text{det}}$,

$$\sum_{t \in \mathcal{M}(T)} r_t \;\le\; 2B \, (\ell(T) - 1) \, L_{\text{det}}(T). \tag{54}$$

Step 4: regret in stationary segments. Fix $k \in \{1, \dots, \hat{\ell}(T)\}$. On the index set $J_k(T)$ the environment is stationary, and MTGP-LR runs GP-UCB using only the post-reset data (from times $\hat{b}_{k-1}+1$ onward). Thus, conditional on the history up to $\hat{b}_{k-1}$, this is exactly a stationary segment of length $S_k(T)$. On the event $\mathcal{E}_{\mathrm{gp}}$, applying Theorem 1 yields

$$\sum_{t \in J_k(T)} r_t \;\le\; \sqrt{C_1\, S_k(T)\, \beta_{S_k(T)}\, \gamma_{S_k(T)}} \;+\; 2. \tag{55}$$

Since $S_k(T) \le L_k(T) \le T$ and β_t, γ_t are nondecreasing in t,

$$\sqrt{C_1\, S_k(T)\, \beta_{S_k(T)}\, \gamma_{S_k(T)}} \;\le\; \sqrt{C_1\, L_k(T)\, \beta_T\, \gamma_T}. \tag{56}$$

Combining (55)–(56) gives, on $\mathcal{E}_{\mathrm{gp}}$,

$$\sum_{t \in J_k(T)} r_t \;\le\; \sqrt{C_1\, L_k(T)\, \beta_T\, \gamma_T} \;+\; 2. \tag{57}$$

Summing (57) over $k = 1, \dots, \hat{\ell}(T)$ and using $\sum_k L_k(T) = T$ yields

$$\sum_{k=1}^{\hat{\ell}(T)} \sum_{t \in J_k(T)} r_t \;\le\; \sqrt{C_1\, \beta_T\, \gamma_T} \sum_{k=1}^{\hat{\ell}(T)} \sqrt{L_k(T)} \;+\; 2\,\hat{\ell}(T).$$

By Cauchy–Schwarz,

$$\sum_{k=1}^{\hat{\ell}(T)} \sqrt{L_k(T)} \le \sqrt{\hat{\ell}(T) \sum_{k=1}^{\hat{\ell}(T)} L_k(T)} = \sqrt{\hat{\ell}(T)\, T},$$

Thus we obtain

$$\sum_{k=1}^{\hat{\ell}(T)} \sum_{t \in J_k(T)} r_t \;\le\; \sqrt{C_1\, \hat{\ell}(T)\, T\, \beta_T\, \gamma_T} \;+\; 2\,\hat{\ell}(T). \tag{58}$$

Step 5: wrong detection. On the event $\mathcal{E}_{\mathrm{fa}}$, resets are not triggered in the absence of a new change. In particular, up to time T, each reset corresponds to a distinct change point, so that

$$\hat{\ell}(T) \;\le\; \ell(T). \tag{59}$$

Substituting (59) into (58) gives, on $\mathcal{E}_{\mathrm{gp}} \cap \mathcal{E}_{\mathrm{fa}}$,

$$\sum_{k=1}^{\hat{\ell}(T)} \sum_{t \in J_k(T)} r_t \;\le\; \sqrt{C_1\, \ell(T)\, T\, \beta_T\, \gamma_T} \;+\; 2\,\ell(T). \tag{60}$$

Step 6: combine bounds. Combining the decomposition (51), the stationary-prefix bound (60), and the delay bound (54), we obtain that on $\mathcal{E}_{\mathrm{gp}} \cap \mathcal{E}_{\mathrm{fa}} \cap \mathcal{E}_{\mathrm{det}}$,

$$R(T) \;\le\; \sqrt{C_1\, \ell(T)\, T\, \beta_T\, \gamma_T} \;+\; 2\,\ell(T) \;+\; 2B\,(\ell(T)-1)\,L_{\mathrm{det}}(T).$$

Since $\mathcal{E}_{\mathrm{gp}}$, $\mathcal{E}_{\mathrm{fa}}$, and $\mathcal{E}_{\mathrm{det}}$ are each uniform-in-T events by construction, the same inequality holds simultaneously for all $T \ge 1$ on their intersection. Finally, (50) implies that this intersection has probability at least $1-\delta$. This proves (31). □

C Model-based benchmarks

This section solves the model of Section 5. Fix a horizon $T > 0$. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ support: (i) a continuous-time Markov chain $\boldsymbol{\alpha}_t \in \mathcal{E} = \{\boldsymbol{e}^1, \dots, \boldsymbol{e}^N\}$ with generator $\Lambda = (\lambda_{ij})_{i,j=1}^N$, (ii) any factor processes $(I_t)_{t\in[0,T]}$ and $(\rho_t)_{t\in[0,T]}$ with infinitesimal generators $\mathcal{L}^I$ and $\mathcal{L}^\rho$ (possibly multidimensional), and (iii) a Brownian motion driving the midprice with volatility $\sigma > 0$. The tactical trading speed ν is progressively measurable and square integrable; denote by $\mathcal{A}_t$ the set of admissible controls on $[t, T]$. The state dynamics are as in Section 5:

$$dX_s^\nu = (S_s - \kappa\nu_s)\nu_s\, ds, \qquad dq_s^\nu = -\nu_s\, ds,$$

and the midprice satisfies

$$dS_s = (\boldsymbol{\alpha}_s)^\top I_s\, ds + \sigma\, dW_s,$$

with $(I_s, \rho_s, \boldsymbol{\alpha}_s)$ evolving autonomously (independently of ν).

For $t \in [0, T]$ and state $(x, q, I, S, \rho, \boldsymbol{e}^i)$, define the value function

$$u(t, x, q, I, S, \rho, \boldsymbol{e}^i) = \sup_{\nu\in\mathcal{A}_t} \mathbb{E}_{t,x,q,I,S,\rho,\alpha_t=\boldsymbol{e}^i}\left[X_T^\nu + q_T^\nu S_T - \Phi(q_T^\nu)^2 - \phi\int_t^T (\nu_s - \rho_s)^2\, ds\right], \tag{61}$$

where $\mathbb{E}_t$ denotes expectation conditional on time-t values of $(x, q, I, S, \rho, \boldsymbol{e}^i)$. We assume conditions ensuring the dynamic programming principle holds and that u is sufficiently regular for the computations below (standard in LQ problems; see Pham (2009)).

Because α is a Markov chain, the HJB is *coupled* across regimes. For each $i \in \{1, \dots, N\}$, let $u^i(t, x, q, I, S, \rho) = u(t, x, q, I, S, \rho, \boldsymbol{e}^i)$. The dynamic programming principle holds, so the functions $(u^i)_{i=1}^N$ satisfy the Hamilton-Jacobi-Bellman (HJB) equation

$$
\begin{aligned}
0 = \Big(\partial_t + \mathcal{L}^I + \mathcal{L}^\rho + (\boldsymbol{e}^i)^\top I\, \partial_S + \tfrac{1}{2}\sigma^2 \partial_{SS}\Big) u^i + \sum_{j=1}^N \lambda_{ij}\big(u^j - u^i\big) \\
+ \sup_{\nu \in \mathbb{R}} \left\{ (S - \kappa\nu)\nu\, \partial_x u^i - \nu\, \partial_q u^i - \phi(\nu - \rho)^2 \right\} ,
\end{aligned} \tag{62}
$$

with terminal condition

$$
u^i(T, x, q, I, S, \rho) = x + qS - \Phi q^2 \qquad i = 1, \dots, N\,. \tag{63}
$$

Proposition 3. *Let* $\tilde{\phi} = \dfrac{\phi + \kappa}{\Phi}$. *Define*

$$
C(t) = -\frac{\phi + \kappa}{T - t + \tilde{\phi}}\,, \qquad t \in [0, T]\,.
$$

For $i \in \{1, \dots, N\}$, *define* B^i *by*

$$
B^i(t, I, \rho) = \frac{1}{T - t + \tilde{\phi}}\, \mathbb{E}_{t, I, \rho, \alpha_t = \boldsymbol{e}^i}\left[\int_t^T (T - s + \tilde{\phi})\, \boldsymbol{\alpha}_s^\top I_s\, ds - 2\phi \int_t^T \rho_s\, ds \right] , \tag{64}
$$

and define A^i *by*

$$
A^i(t, I, \rho) = \mathbb{E}_{t, I, \rho, \alpha_t = \boldsymbol{e}^i}\left[\int_t^T \left(\frac{\big(B^{\boldsymbol{\alpha}_s}(s, I_s, \rho_s) - 2\phi\rho_s\big)^2}{4(\phi + \kappa)} - \phi\rho_s^2 \right) ds \right] , \tag{65}
$$

where $B^{\boldsymbol{\alpha}_s}(s, I_s, \rho_s)$ *means* $B^j(s, I_s, \rho_s)$ *when* $\boldsymbol{\alpha}_s = \boldsymbol{e}^j$.

Then the coupled HJB system (62)–(63) admits the solution

$$u(t,x,q,I,S,\rho,\boldsymbol{e}^i) = x + qS + C(t)q^2 + B^i(t,I,\rho)\,q + A^i(t,I,\rho)\,. \tag{66}$$

Moreover, an optimal feedback control is

$$\nu_t^\star = \frac{2\phi\,\rho_t - \big(2C(t)q_t + B^{\alpha_t}(t,I_t,\rho_t)\big)}{2(\phi+\kappa)} = \frac{q_t}{T-t+\tilde{\phi}} - \frac{B^{\alpha_t}(t,I_t,\rho_t)}{2(\phi+\kappa)} + \frac{\phi}{\phi+\kappa}\rho_t\,. \tag{67}$$

Proof. Eliminate (x,S) and compute the pointwise maximiser. Motivated by the terminal condition, seek a solution of the form

$$u^i(t,x,q,I,S,\rho) = x + qS + \theta^i(t,q,I,\rho).$$

Then $\partial_x u^i \equiv 1$ and $\partial_S u^i \equiv q$, and (62) becomes

$$\begin{aligned} 0 = \big(\partial_t + \mathcal{L}^I + \mathcal{L}^\rho\big)\theta^i(t,q,I,\rho) + (\boldsymbol{e}^i)^\top I\,q + \sum_{j=1}^N \lambda_{ij}\big(\theta^j - \theta^i\big) \\ + \sup_{\nu\in\mathbb{R}}\big\{-\kappa\nu^2 - \nu\,\partial_q\theta^i(t,q,I,\rho) - \phi(\nu-\rho)^2\big\}\,. \end{aligned}$$

The supremum is over a concave quadratic in ν:

$$-\kappa\nu^2 - \phi(\nu-\rho)^2 - \nu\,\partial_q\theta^i = -(\phi+\kappa)\nu^2 + (2\phi\rho - \partial_q\theta^i)\nu - \phi\rho^2.$$

Hence the maximiser is

$$\nu^\star = \frac{2\phi\rho - \partial_q\theta^i}{2(\phi+\kappa)}, \tag{68}$$

and the value of the supremum equals

$$\frac{(2\phi\rho - \partial_q\theta^i)^2}{4(\phi+\kappa)} - \phi\rho^2.$$

Substituting back yields the coupled PDE for θ^i:

$$0 = \big(\partial_t + \mathcal{L}^I + \mathcal{L}^\rho\big)\theta^i + (\boldsymbol{e}^i)^\top I\, q + \sum_{j=1}^{N} \lambda_{ij}\big(\theta^j - \theta^i\big) + \frac{(2\phi\rho - \partial_q\theta^i)^2}{4(\phi+\kappa)} - \phi\rho^2, \tag{69}$$

with terminal condition $\theta^i(T, q, I, \rho) = -\Phi q^2$.

Quadratic ansatz and identification of coefficients. Seek θ^i in quadratic form:

$$\theta^i(t, q, I, \rho) = A^i(t, I, \rho) + B^i(t, I, \rho)\, q + C(t)\, q^2,$$

where, crucially, the quadratic coefficient is *regime-independent* (this will follow from the equations). Then $\partial_q\theta^i = B^i + 2Cq$. Substitute into (69) and match coefficients in powers of q.

(i) Quadratic term. The only q^2 contributions come from $\partial_t(Cq^2)$ and the square term:

$$0 = \left(C'(t) + \frac{C(t)^2}{\phi+\kappa}\right) q^2.$$

Hence

$$C'(t) + \frac{C(t)^2}{\phi+\kappa} = 0, \qquad C(T) = -\Phi,$$

whose unique solution is $C(t) = -(\phi+\kappa)/(T - t + \tilde{\phi})$ with $\tilde{\phi} = (\phi+\kappa)/\Phi$.

(ii) Linear term. Collecting the coefficients of q gives, for each i,

$$0 = (\partial_t + \mathcal{L}^I + \mathcal{L}^\rho)B^i(t, I, \rho) + (\boldsymbol{e}^i)^\top I + \sum_{j=1}^{N} \lambda_{ij}\big(B^j - B^i\big) - \frac{C(t)}{\phi+\kappa}\big(2\phi\rho - B^i(t, I, \rho)\big), \tag{70}$$

with terminal condition $B^i(T, \cdot) = 0$.

(iii) Constant term. Similarly,

$$0 = (\partial_t + \mathcal{L}^I + \mathcal{L}^\rho)A^i(t, I, \rho) + \sum_{j=1}^{N} \lambda_{ij}\big(A^j - A^i\big) + \frac{(2\phi\rho - B^i(t, I, \rho))^2}{4(\phi+\kappa)} - \phi\rho^2, \tag{71}$$

with terminal condition $A^i(T,\cdot)=0$.

Feynman–Kac representation and closed forms for B^i and A^i. Equation (70) is linear. Using the Markov property of (I_t,ρ_t,α_t) and Feynman–Kac representation for regime-switching systems yields

$$B^{\alpha_t}(t,I_t,\rho_t)=\mathbb{E}_t\left[\int_t^T \exp\left(\int_t^s \frac{C(u)}{\phi+\kappa}du\right)\left(\boldsymbol{\alpha}_s^\top I_s-\frac{2\phi\rho_s}{\phi+\kappa}C(s)\right)ds\right]. \tag{72}$$

Because $C(u)=-(\phi+\kappa)/(T-u+\tilde{\phi})$,

$$\exp\left(\int_t^s \frac{C(u)}{\phi+\kappa}du\right)=\exp\left(-\int_t^s \frac{du}{T-u+\tilde{\phi}}\right)=\frac{T-s+\tilde{\phi}}{T-t+\tilde{\phi}}.$$

Plugging this into (72) and simplifying yields (64).

Likewise, applying Feynman–Kac to (71) yields (65); note that the integrand must depend on the *current* regime $\boldsymbol{\alpha}_s$ through $B^{\boldsymbol{\alpha}_s}(s,I_s,\rho_s)$.

Feedback control and verification. With $\partial_q\theta^{\alpha_t}=B^{\alpha_t}(t,I_t,\rho_t)+2C(t)q_t$, the maximiser (68) gives (67). Finally, a standard verification theorem for regime-switching controlled diffusions (e.g. Pham (2009)) shows that u in (66) equals the value function (61) and that $\nu^\star$ is optimal. □

Closed-form oracle speed Combining (67) with (64) yields the explicit oracle speed

$$\nu_t^\star=\frac{q_t}{T-t+\tilde{\phi}}-\frac{1}{2(\phi+\kappa)(T-t+\tilde{\phi})}\mathbb{E}_{t,I,\alpha_t=\boldsymbol{e}^i}\left[\int_t^T(T-s+\tilde{\phi})\,(\alpha_s)^\intercal I_s\,ds\right]+$$
$$\frac{\phi}{(\phi+\kappa)(T-t+\tilde{\phi})}\int_t^T\mathbb{E}_{t,\rho}[\rho_s]\,ds+\frac{\phi}{\phi+\kappa}\rho_t.$$

References

Almgren, R., Chriss, N., 2000. Optimal execution of portfolio transactions. Journal of Risk 3, 5–39.

Arnoldi, W.E., 1951. The principle of minimized iterations in the solution of the matrix eigenvalue problem. Quarterly of Applied Mathematics 9, 17–29.

Arroyo, A., Cartea, A., Moreno-Pino, F., Zohren, S., 2024. Deep attentive survival analysis in limit order books: Estimating fill probabilities with convolutional-transformers. Quantitative Finance 24, 35–57.

Arroyo, A., Scalzo, B., Stanković, L., Mandic, D.P., 2022. Dynamic portfolio cuts: A spectral approach to graph-theoretic diversification, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 5468–5472.

Balata, A., Ludkovski, M., Maheshwari, A., Palczewski, J., 2021. Statistical learning for probability-constrained stochastic optimal control. European Journal of Operational Research 290, 640–656.

Bechler, K., Ludkovski, M., 2015. Optimal execution with dynamic order flow imbalance. SIAM Journal on Financial Mathematics 6, 1123–1151. doi:10.1137/140992254.

Belak, C., Muhle-Karbe, J., Ou, K., 2018. Optimal trading with general signals and liquidation in target zone models. arXiv preprint arXiv:1808.00515 .

Bellani, C., Brigo, D., Done, A., Neuman, E., 2021. Optimal trading: The importance of being adaptive. International Journal of Financial Engineering 8, 2050022.

Bergault, P., Drissi, F., Guéant, O., 2022. Multi-asset optimal execution and statistical arbitrage strategies under ornstein–uhlenbeck dynamics. SIAM Journal on Financial Mathematics 13, 353–390. doi:10.1137/21M1407756.

Besbes, O., Gur, Y., Zeevi, A., 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. Advances in neural information processing systems 27.

Bogunovic, I., Krause, A., 2021. Misspecified gaussian process bandit optimization. Advances in Neural Information Processing Systems 34, 3004–3015.

Bogunovic, I., Scarlett, J., Cevher, V., 2016. Time-varying gaussian process bandit optimization, in: Artificial Intelligence and Statistics, PMLR. pp. 314–323.

Bonilla, E.V., Chai, K., Williams, C., 2007. Multi-task gaussian process prediction. Advances in neural information processing systems 20.

Caldarelli, E., Wenk, P., Bauer, S., Krause, A., 2022. Adaptive gaussian process change point detection, in: International Conference on Machine Learning, PMLR. pp. 2542–2571.

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

Cartea, Á., Chang, P., Mroczka, M., Oomen, R., 2022. Ai-driven liquidity provision in otc financial markets. Quantitative Finance 22, 2171–2204.

Cartea, Á., Donnelly, R., Jaimungal, S., 2018. Enhancing trading strategies with order book signals. Applied Mathematical Finance 25, 1–35.

Cartea, Á., Drissi, F., Monga, M., 2023a. Execution and statistical arbitrage with signals in multiple automated market makers, in: 2023 IEEE 43rd International Conference on Distributed Computing Systems Workshops (ICDCSW), IEEE. pp. 37–42.

Cartea, Á., Drissi, F., Monga, M., 2024. Decentralized finance and automated market making: Predictable loss and optimal liquidity provision. SIAM Journal on Financial Mathematics 15, 931–959.

Cartea, Á., Drissi, F., Monga, M., 2025. Decentralised finance and automated market making: Execution and speculation. Journal of Economic Dynamics and Control , 105134.

Cartea, Á., Drissi, F., Osselin, P., 2023b. Bandits for algorithmic trading with signals. Available at SSRN 4484004 .

Cartea, Á., Jaimungal, S., 2016. A closed-form execution strategy to target volume weighted average price. SIAM Journal on Financial Mathematics 7, 760–785.

Cartea, Á., Jaimungal, S., Penalva, J., 2015. Algorithmic and High-Frequency Trading. Cambridge University Press.

Cartea, Á., Jaimungal, S., Sánchez-Betancourt, L., 2023c. Reinforcement learning for algorithmic trading, in: Lehalle, C.A., Capponi, A. (Eds.), Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices. Cambridge University Press.

Chen, T., Ludkovski, M., Voß, M., 2024. On parametric optimal execution and machine learning surrogates. Quantitative Finance 24, 15–34.

Chu, W., Li, L., Reyzin, L., Schapire, R., 2011. Contextual bandits with linear payoff functions, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings. pp. 208–214.

Cohen, S.N., Treetanthiploet, T., 2021. Correlated bandits for dynamic pricing via the arc algorithm. arXiv preprint arXiv:2102.04263 .

Dani, V., Kakade, S.M., Hayes, T., 2007. The price of bandit information for online optimization. Advances in Neural Information Processing Systems 20.

Dette, H., Gösmann, J., 2020. A likelihood ratio approach to sequential change point detection for a general class of parameters. Journal of the American Statistical Association 115, 1361–1377.

Ding, K., Li, J., Liu, H., 2019. Interactive anomaly detection on attributed networks, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, p. 357–365.

Donnelly, R., 2022. Optimal execution: A review. Applied Mathematical Finance , 1–32.

Drissi, F., 2022. Solvability of differential riccati equations and applications to algorithmic trading with signals. Applied Mathematical Finance 29, 457–493.

Duran-Martin, G., Kara, A., Murphy, K., 2022. Efficient online bayesian inference for neural bandits, in: International Conference on Artificial Intelligence and Statistics, PMLR. pp. 6002–6021.

Duran-Martin, G., Sánchez-Betancourt, L., Shestopaloff, A.Y., Murphy, K., 2024. A unifying framework for generalised bayesian online learning in non-stationary environments. arXiv preprint arXiv:2411.10153 .

Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G.D., Pineau, J., 2018. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis, in: Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., Wiens, J. (Eds.), Proceedings of the 3rd Machine Learning for Healthcare Conference, pp. 67–82.

Frei, C., Westray, N., 2015. Optimal execution of a vwap order: a stochastic control approach. Mathematical Finance 25, 612–639.

Garivier, A., Moulines, E., 2008. On upper-confidence bound policies for non-stationary bandit problems. arXiv preprint arXiv:0805.3415 .

Gonzalvez, J., Lezmi, E., Roncalli, T., Xu, J., 2019. Financial applications of gaussian processes and bayesian optimization. arXiv preprint arXiv:1903.04841 .

Guéant, O., 2012. Execution and block trade pricing with optimal constant rate of participation. arXiv preprint arXiv:1210.7608 .

Guéant, O., Manziuk, I., 2019. Deep reinforcement learning for market making in corporate bonds: beating the curse of dimensionality. Applied Mathematical Finance 26, 387–452.

Guéant, O., 2016. The Financial Mathematics of Market Liquidity: From Optimal Execution to Market Making. doi:10.1201/b21350.

Hollifield, B., Miller, R.A., Sandås, P., 2004. Empirical analysis of limit order markets. The Review of Economic Studies 71, 1027–1063.

Jafar, S.H., 2022. Financial applications of gaussian processes and bayesian optimization, in: Bayesian Reasoning and Gaussian Processes for Machine Learning Applications. Chapman and Hall/CRC, pp. 111–122.

Kovaleva, P., Iori, G., 2012. Optimal trading strategies in a limit order market with imperfect liquidity .

Krause, A., Ong, C., 2011. Contextual gaussian process bandit optimization. Advances in neural information processing systems 24.

Lehalle, C.A., 2015. Mathematical Models to Study and Control the Price Formation Process. Ph.D. thesis. Université Pierre et Marie Curie.

Lehalle, C.A., Laruelle, S., 2018. Market microstructure in practice. World Scientific.

Lehalle, C.A., Neuman, E., 2019. Incorporating signals into optimal trading. Finance and Stochastics 23, 275–311.

Liu, B., Yu, T., Lane, I., Mengshoel, O., 2018. Customized nonlinear bandits for online response selection in neural conversation models. Proceedings of the AAAI Conference on Artificial Intelligence 32.

Ludkovski, M., Saporito, Y., 2021. Krighedge: Gaussian process surrogates for delta hedging. Applied Mathematical Finance 28, 330–360.

Ludkovski, M., Zail, H., 2022. Gaussian process models for incremental loss ratios. Variance 15.

Lyu, X., Binois, M., Ludkovski, M., 2021. Evaluating gaussian process metamodels and sequential designs for noisy level set estimation. Statistics and Computing 31, 43.

Lyu, X., Ludkovski, M., 2022. Adaptive batching for gaussian process surrogates with application in noisy level set estimation. Statistical Analysis and Data Mining: The ASA Data Science Journal 15, 225–246.

Moallemi, C.C., Wang, M., 2022. A reinforcement learning approach to optimal execution. Quantitative Finance 22, 1051–1069.

Ning, B., Lin, F.H.T., Jaimungal, S., 2021. Double deep q-learning for optimal execution. Applied Mathematical Finance 28, 361–380.

Pham, H., 2009. Continuous-time stochastic control and optimization with financial applications. volume 61. Springer Science & Business Media.

Pleiss, G., Gardner, J., Weinberger, K., Wilson, A.G., 2018. Constant-time predictive distributions for gaussian processes, in: International Conference on Machine Learning, PMLR. pp. 4114–4123.

Rasmussen, C., Ghahramani, Z., 2000. Occam's razor. Advances in neural information processing systems 13.

Scalzo, B., Arroyo, A., Stanković, L., Mandic, D.P., 2021. Nonstationary portfolios: Diversification in the spectral domain, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 5155–5159.

Seldin, Y., Szepesvári, C., Auer, P., Abbasi-Yadkori, Y., 2013. Evaluation and analysis of the performance of the exp3 algorithm in stochastic environments, in: European Workshop on Reinforcement Learning, PMLR. pp. 103–116.

Srinivas, N., Krause, A., Kakade, S.M., Seeger, M., 2009. Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:0912.3995 .

Srivastava, V., Reverdy, P., Leonard, N.E., 2014. Surveillance in an abruptly changing world via multiarmed bandits, in: 53rd IEEE Conference on Decision and Control, IEEE. pp. 692–697.

Tapia, J.F., 2015. Modeling, optimization and estimation for the on-line control of trading algorithms in limit-order markets. Ph.D. thesis. Université Pierre et Marie Curie-Paris VI.

Wainwright, M.J., 2019. High-dimensional statistics: A non-asymptotic viewpoint. volume 48. Cambridge University Press.

Wald, J.K., Horrigan, H.T., 2005. Optimal limit order choice. The Journal of Business 78, 597–620.

Waldon, H., Drissi, F., Limmer, Y., Berdica, U., Foerster, J.N., Cartea, A., 2024. Dare: The deep adaptive regulator for control of uncertain continuous-time systems, in: ICML 2024 Workshop: Foundations of Reinforcement Learning and Control–Connections and Perspectives.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian processes for machine learning. volume 2. MIT press Cambridge, MA.

Wilson, A., Nickisch, H., 2015. Kernel interpolation for scalable structured gaussian processes (kiss-gp), in: International conference on machine learning, PMLR. pp. 1775–1784.

Wilson, A.G., Dann, C., Nickisch, H., 2015. Thoughts on massively scalable gaussian processes. arXiv preprint arXiv:1511.01870 .

Yam, S.C.P., Zhou, W., 2017. Optimal liquidation of child limit orders. Mathematics of Operations Research 42, 517–545.

Yingsaeree, C., 2012. Algorithmic trading: Model of execution probability and order placement strategy. Ph.D. thesis. UCL (University College London).

Zhou, Q., Zhang, X., Xu, J., Liang, B., 2017. Large-scale bandit approaches for recommender systems, in: Neural Information Processing - 24th International Conference, ICONIP 2017, pp. 811–821.