Adaptive Execution: Exploration and Learning of Price Impact

Beomsoo Park
Electrical Engineering
Stanford University
beomsoo@stanford.edu

Benjamin Van Roy
Management Science & Engineering
Electrical Engineering
Stanford University
bvr@stanford.edu

September 28, 2013

Abstract

We consider a model in which a trader aims to maximize expected risk-adjusted profit while trading a single security. In our model, each price change is a linear combination of observed factors, impact resulting from the trader's current and prior activity, and unpredictable random effects. The trader must learn coefficients of a price impact model while trading. We propose a new method for simultaneous execution and learning – the confidence-triggered regularized adaptive certainty equivalent (CTRACE) policy – and establish a poly-logarithmic finite-time expected regret bound. In addition, we demonstrate via Monte Carlo simulation that CTRACE outperforms the certainty equivalent policy and a recently proposed reinforcement learning algorithm that is designed to explore efficiently in linear-quadratic control problems.

Key words: adaptive execution, price impact, reinforcement learning, regret bound

1 Introduction

Trades move the market by either disturbing the balance between supply and demand or adjusting other market participants' valuations. This effect is called *price impact*. Because it is responsible for a large fraction of transaction costs, it is important to design trading strategies that effectively manage price impact. In light of this, academics and practitioners have devoted significant attention to the topic [Bertsimas and Lo [1998], Almgren and Chriss [2000], Kissell and Glantz [2003], Alfonsi et al. [2010], Moallemi et al. [2012], Obizhaeva and Wang [2013]].

The learning of a price impact model poses a challenging problem. Price impact represents an aggregation of numerous market participants' interpretations of and reactions to executed trades.

As such, learning requires "excitation" of the market, which can be induced by regular trading activity or trades deliberately designed to facilitate learning. The trader must balance the short term costs of accelerated learning against the long term benefits of an accurate model. Further, given the continual evolution of trading venues and population of market participants, price impact models require retuning over time. In this paper, we develop an algorithm that learns a price impact model while guiding trading decisions using the model being learned.

Our problem can be viewed as a special case of reinforcement learning. This topic more broadly addresses sequential decision problems in which unknown properties of an environment must be learned in the course of operation (see, e.g., Sutton and Barto [1998]). Research in this area has established how judicious investments in decisions that explore the environment at the expense of suboptimal short-term behavior can greatly improve longer-term performance. What we develop in this paper can be viewed as a reinforcement learning algorithm; the workings of price impact are unknown, and exploration facilitates learning.

In reinforcement learning, one seeks to optimize the balance between exploration and exploitation – the use of what has already been learned to maximize rewards without regard to further learning. Certainty equivalent control (CE) represents one extreme where at any time, current point estimates are assumed to be correct and actions are made accordingly. This is an instance of pure exploitation; though learning does progress with observations made as the system evolves, decisions are not deliberately oriented to enhance learning.

An important question is how aggressively a trader should explore to learn a price impact model. Unlike many other reinforcement learning problems, in ours a considerable degree of exploration is naturally induced by exploitative decisions. This is because a trader excites the market through regular trading activity regardless of whether or not she aims to learn a price impact model. This activity could, for example, be triggered by return-predictive factors, and given sufficiently large factor variability, the induced exploration might adequately resolve uncertainties about price impact. Results of this paper demonstrate that executing trades to explore beyond what would naturally occur through exploitation can yield significant benefit.

Our work is constructive: we propose the *confidence-triggered regularized adaptive certainty equivalent* policy (CTRACE), pronounced "see-trace," a new method that explores and learns a price impact model alongside trading. CTRACE can be viewed as a generalization of CE, which at

each point in time estimates coefficients of a price impact model via least-squares regression using available data and makes decisions that optimize trading under an assumption that the estimated model is correct and will be used to guide all future decisions. CTRACE deviates in two ways: (1) ℓ_2 regularization is applied in least-squares regression and (2) coefficients are only updated when a certain measure of confidence exceeds a pre-specified threshold and a minimum inter-update time has elapsed. Note that CTRACE reduces to CE as the regularization penalty, the threshold, and the minimum inter-update time vanish.

We demonstrate through Monte Carlo simulation that CTRACE outperforms CE. Further, we establish a finite-time regret bound for CTRACE; no such bound is available for CE. *Regret* is defined here to be the difference between realized risk-adjusted profit of a policy in question and one that is optimal with respect to the true price impact model. Our bound exhibits a poly-logarithmic dependence on time. Among other things, this regret bound implies that CTRACE is *efficient* in the sense that the ϵ-convergence time is bounded by a polynomial function of $1/\epsilon$ and $\log(1/\delta)$ with probability at least $1-\delta$. We define the ϵ-convergence time to be the first time when an estimate and all the future estimates following it are within an ϵ-neighborhood of a true value. Let us provide here some intuition for why CTRACE outperforms CE. First, regularization enhances exploration in a critical manner. Without regularization, we are more likely to obtain overestimates of price impact. Such an outcome abates trading and thus exploration, making it difficult to escape from the predicament. Regularization reduces the chances of obtaining overestimates, and further, tends to yield underestimates that encourage active exploration. Second, requiring a high degree of confidence reduces the chances of occasionally producing erratic estimates, which regularly arise with application of CE. Such estimates can result in undesirable trades and/or reductions in the degree of exploration.

It is also worth comparing CTRACE to a reinforcement learning algorithm recently proposed in Abbasi-Yadkori and Szepesvàri [2010] which appears well-suited for our problem. This algorithm was designed to explore efficiently in a broader class of linear-quadratic control problems, and is based on the *principle of optimism in the face of uncertainty*. Abbasi-Yadkori and Szepesvàri [2010] establish an $O(\sqrt{T\log(1/\delta)})$ regret bound that holds with probability at least $1-\delta$, where T denotes time and some logarithmic terms are hidden. Our bound for CTRACE is on expected regret and exhibits a dependence on T of $O(\log^2 T)$. We also demonstrate via Monte Carlo simulation that

CTRACE dramatically outperforms this algorithm.

To summarize, the primary contributions of this paper include:

(a) We propose a new method for simultaneous execution and learning – the confidence-triggered regularized adaptive certainty equivalent (CTRACE) policy.

(b) We establish a finite-time expected regret bound for CTRACE that exhibits a poly-logarithmic dependence on time. This bound implies that CTRACE is *efficient* in the sense that, with probability at least $1 - \delta$, the ϵ-convergence time is bounded by a polynomial function of $1/\epsilon$ and $\log(1/\delta)$.

(c) We demonstrate via Monte Carlo simulation that CTRACE outperforms the certainty equivalent policy and a reinforcement learning algorithm recently proposed by Abbasi-Yadkori and Szepesvàri [2010] which is designed to explore efficiently in linear-quadratic control problems. This computational analysis serves to illustrate and quantify potential economic benefit of CTRACE.

The organization of the rest of this paper is as follows: Section 2 presents our problem formulation, establishes existence of an optimal solution to our problem, and defines performance measures that can be used to evaluate policies. In Section 3, we propose CTRACE and establish a finite-time expected regret bound along with two important properties of the algorithm: inter-temporal consistency and efficiency. Section 4 is devoted to Monte Carlo simulation in which the performance of CTRACE is compared to that of two benchmark policies. Finally, we conclude this paper in Section 5. Proofs of selected key results are presented in the appendix. Due to space constraints, other proofs are provided in an e-companion.

2 Problem Formulation

2.1 Model Description

Decision Variable and Security Position We consider a trader who trades a single risky security over an infinite time horizon. She submits a market buy or sell order at the beginning of each period of equal length. $u_t \in \mathbb{R}$ represents the number of shares of the security to buy or sell at period t

and a positive (negative) value of u_t denotes a buy (sell) order. Let $x_{t-1} \in \mathbb{R}$ denote the trader's pre-trade security position before placing an order u_t at period t. Therefore, $x_t = x_{t-1} + u_t,\ t \geq 1$.

Price Dynamics The absolute return of the security is given by

$$\begin{aligned} \Delta p_t = p_t - p_{t-1} &= g^\top f_{t-1} + \lambda^* u_t + \sum_{m=1}^{M} \gamma_m^* (d_{m,t} - d_{m,t-1}) + \epsilon_t \\ d_{m,t} \triangleq \sum_{i=1}^{t} r_m^{t-i} u_i &= r_m d_{m,t-1} + u_t, \quad d_t \triangleq [d_{1,t} \ \cdots \ d_{M,t}]^\top. \end{aligned} \tag{1}$$

We will explain each term in detail as we progress. The superscript $*$ denotes unknown true parameters that the trader wants to learn. This can be viewed as a first-order Taylor expansion of a geometric model

$$\log\left(\frac{p_t}{p_{t-1}}\right) = (g^{\text{geo}})^\top f_{t-1} + \lambda^{\text{geo}} u_t + \sum_{m=1}^{M} \gamma_m^{\text{geo}} (d_{m,t} - d_{m,t-1}) + \epsilon_t^{\text{geo}}$$

over a certain period of time, say, a few weeks in calendar time, which makes this approximation reasonably accurate for practical purposes. Although it is unrealistic that the security price can be negative with positive probability, our model nevertheless serves its practical purpose for the following reasons: Our numerical experiments conducted in Section 4 show that price changes after a few weeks from now have ignorable impacts on a current optimal action. In other words, optimal actions for our infinite-horizon control problem appear to be quite close to those for a finite-horizon counterpart on a few week time scale. Furthermore, it turns out that in simulation we could learn a unknown price impact model fast enough to take actions that are close to optimal actions within a few weeks. Thus, learning based on our price dynamics model could also be justified. We discuss these notions in detail in the appendix.

Price Impact The term $\lambda^* u_t$ represents "permanent price impact" on the security price of a current trade. The permanent price impact is endogenously derived in Kyle [1985] from informational asymmetry between an informed trader and uninformed competitive market makers, and in Rosu [2009] from equilibrium of a limit order market where fully strategic liquidity traders dynamically choose limit and market orders. Huberman and Stanzl [2004] prove that the linearity of a time-independent permanent price impact function is a necessary and sufficient condition for the absence of "price manipulation" and "quasi-arbitrage" under some regularity conditions.

The term $\sum_{m=1}^{M} \gamma_m^* d_{m,t}$ indicates "transient price impact" that models other traders' responses to non-informative orders. For example, suppose that a large market buy order has arrived and other traders monitoring the market somehow realize that there is no definitive evidence for abrupt change in the fundamental value of the security. Then, they naturally infer that the large buy order came merely for some liquidity reason, and gradually "correct" the perturbed price into what they believe it is supposed to be by submitting counteracting selling orders. We verify this notion in Section 4.1 using vector autoregressive models fitted to real transactions data from Dufour and Engle [2000]. The dynamics of $d_{m,t}$ in (1) indicates that the impact of a current trade on the security price decays exponentially over time, which is considered in Obizhaeva and Wang [2013] that incorporate the dynamics of supply and demand in a limit order market to optimal execution strategies. In Gatheral [2010], it is shown that the exponentially decaying transient price impact is compatible only with a linear instantaneous price impact function in the absence of "dynamic arbitrage."

Observable Return-Predictive Factors We assume that there are multiple observable return-predictive factors that affect the absolute return of the security as in Garleanu and Pedersen [2012]. Examples for short-term return-predictive factors include:

(a) *Order Imbalance*: A total amount of orders outstanding at the best bid and ask price could provide useful information about short-term mid-price movement, especially when the "queue length" on one side is much longer than on the other side as investigated in Cont et al. [2012]. For example, a much greater number of sell limit orders on the best ask price might represent high "sell-pressure" from liquidity suppliers followed by short-term downward movement of mid-price.

(b) *Trade Sign Autocorrelation*: Many empirical studies found that trade signs are positively correlated, at least over a short time scale, as discussed in Hasbrouck [1991], Dufour and Engle [2000] and Bouchaud et al. [2004]. In other words, buy orders tend to follow buy orders and similarly for sell orders. It implies a certain degree of predictability for trade direction and potentially for short-term price movement. For example, one can model logit transformation of the proportion of buy market order over 5-minute interval as an autoregressive process.

(c) *Cointegration*: The price of the security traded by the trader could be cointegrated with that

of similar securities or sector/market indices. Specifically, the difference between the two prices might be mean-reverting such that large deviations from a long-term mean value are likely to be followed by movement reverting to it.

In our price dynamics model, $f_t \in \mathbb{R}^K$ denotes these factors and $g \in \mathbb{R}^K$ denotes factor loadings. The term $g^\top f_{t-1}$ represents predictable excess return or "alpha." We assume that f_t is a first-order vector autoregressive process $f_t = \Phi f_{t-1} + \omega_t$ where $\Phi \in \mathbb{R}^{K\times K}$ is a stable matrix that has all eigenvalues inside a unit disk and $\omega_t \in \mathbb{R}^K$ is a martingale difference sequence adapted to the filtration $\{\mathcal{F}_t \triangleq \sigma(\{x_0, d_0, f_0, \omega_1, \ldots, \omega_t, \epsilon_1, \ldots, \epsilon_t\})\}$. We further assume that ω_t is bounded almost surely, i.e. $\|\omega_t\| \le C_\omega \ a.s.$ for all $t \ge 1$ for some deterministic constant C_ω, and $\mathrm{Cov}[\omega_t|\mathcal{F}_{t-1}] = \Omega \in \mathbb{R}^{K\times K}$ being positive definite and independent of t.

Unpredictable Noise The term ϵ_t represents random fluctuations that cannot be accounted for by price impact and observable return-predictive factors. We assume that ϵ_t is a martingale difference sequence adapted to the filtration $\{\mathcal{F}_t\}$, and independent of x_0, d_0, f_0 and ω_τ for any $\tau \ge 1$. Also, $\mathsf{E}[\epsilon_t^2|\mathcal{F}_{t-1}] = \Sigma_\epsilon \in \mathbb{R}$ being independent of t. Finally, each ϵ_t is assumed to be sub-Gaussian, i.e., $\mathsf{E}[\exp(a\epsilon_t)|\mathcal{F}_{t-1}] \le \exp(C_\epsilon^2 a^2/2),\ \forall t \ge 1,\ \forall a \in \mathbb{R}$ for some $C_\epsilon > 0$.

Policy A policy is defined as a sequence $\pi = \{\pi_1, \pi_2, \ldots\}$ of functions where π_t maps the trader's information set at the beginning of period t into an action u_t. The trader observes f_{t-1} and p_{t-1} at the end of period $t-1$ and thus her information set at the beginning of period t is given by $\mathcal{I}_{t-1} = \{x_0, d_0, f_0, \ldots, f_{t-1}, p_0, \ldots, p_{t-1}, u_1, \ldots, u_{t-1}\}$. A policy π is *admissible* if $z_t \triangleq [x_t \ \ d_t^\top \ \ f_t^\top]^\top$ generated by $u_t = \pi_t(\mathcal{I}_{t-1})$ satisfies $\lim_{T\to\infty} \|z_T\|^2/T = 0$. A set of admissible policies is denoted by Π.

Objective Function The trader's objective is to maximize over admissible policies $\pi \in \Pi$ expected average "risk-adjusted" profit, i.e.,

$$\max_{\pi\in\Pi} \ \liminf_{T\to\infty} \ \mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\{\mathsf{E}[\Delta p_t x_{t-1}|\mathcal{I}_{t-1}, u_t = \pi(\mathcal{I}_{t-1})] - \rho\mathrm{Var}[p_t x_t|\mathcal{I}_{t-1}, u_t = \pi(\mathcal{I}_{t-1})]\right\}\right]. \qquad (2)$$

The first term in the summand indicates expected profit from trading u_t shares conditioned on her information set at the moment. The second term corresponds to conditional variance of the value of her post-trade position marked to market price p_t, which captures her risk aversion. ρ is a risk-aversion coefficient that quantifies the extent to which the trader is risk-averse. It is straightforward

to see that $\mathrm{Var}\left[p_t x_t | \mathcal{I}_{t-1}, u_t\right] = \Sigma_\epsilon x_t^2$ and thus (2) is equivalent to

$$\max_{\pi \in \Pi} \liminf_{T \to \infty} \mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\Delta p_t x_{t-1} - \rho \Sigma_\epsilon x_t^2\right)\right].$$

Assumptions The following is a list of assumptions on which our analysis is based throughout this paper. Let $\theta^* \triangleq [\lambda^* \ \gamma_1^* \ \ldots \ \gamma_M^*]^\top \in \mathbb{R}^{M+1}$ denote a vector of true price impact coefficients.

Assumption 1. *(a) θ^* is unknown to the trader.*

(b) The factor loadings g are known to the trader.

(c) The decaying rates $r \triangleq [r_1, \ \ldots, \ r_M]^\top \in [0,1)^M$ of the transient price impact are known to the trader and all the elements are distinct.

(d) θ^* is in the parameter set $\Theta \triangleq \{\theta \in \mathbb{R}^{M+1} : 0 \le \theta \le \theta_{max}, \ \mathbf{1}^\top \theta \ge \beta\}$ for some $\theta_{max} > 0$ component-wise and some $\beta > 0$.

Note that price impact coefficients can be learned only through executed trades whereas the factor loadings can be learned by observing prices without any transaction. In practice, the decay rates are definitely not known a priori. However, it can be handled effectively for practical purposes by using a sufficiently dense r with a large M so that potential bias from modeling mismatch can be greatly reduced at the expense of increased variance, which can be reduced by regularization. We will illustrate this approach in Section 4.3. Finally, the constraint $\mathbf{1}^\top \theta \ge \beta$ is imposed to capture non-zero execution costs in practice. Note that Θ is compact and convex. In addition to these assumptions, we will make two more assumptions in Section 2.2. Assumption 1 will be taken to hold for all results stated on the remainder of the paper.

Notations $\|\cdot\|$ and $\|\cdot\|_F$ denote the ℓ_2-norm and the Frobenius norm of a matrix, respectively. $a \vee b$ and $a \wedge b$ denote $\max\{a,b\}$ and $\min\{a,b\}$, respectively. For a symmetric matrix A, $A \succ 0$ means that A is positive definite and $A \succeq 0$ means that A is positive semidefinite. $\lambda_{\min}(A)$ indicates the smallest eigenvalue of A and $\lambda_{\max}(A)$ the largest eigenvalue of A. $(A)_{ij}$ of a matrix A indicates the entry of A in the ith row and in the jth column. $(v)_i$ of a vector v indicates the ith entry of v. $\mathrm{diag}(v)$ of a vector v denotes a diagonal matrix whose ith diagonal entry is $(v)_i$. $A_{*,j}$ denotes the jth column of A and $A_{i:j,k}$ indicates a segment of the kth column of A from the ith entry to the jth entry. $\mathbf{1}\{\mathcal{B}\}$ denotes an indicator function on the event $\mathcal{B}$.

2.2 Existence of Optimal Solution

Now, we will show that there exists an optimal policy among admissible policies that maximizes expected average risk-adjusted profit. For convenience, we will consider the following minimization problem that is equivalent to (2):

$$\min_{\pi\in\Pi} \limsup_{T\to\infty} \mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\rho\Sigma_\epsilon x_t^2 - \Delta p_t x_{t-1}\right)\right]$$

We call the negative of average risk-adjusted profit "average cost." This problem can be expressed as a discrete-time linear quadratic control problem

$$\min_{\pi\in\Pi} \limsup_{T\to\infty} \mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} z_{t-1}^\top & u_t \end{bmatrix}\begin{bmatrix} Q & S \\ S^\top & R \end{bmatrix}\begin{bmatrix} z_{t-1} \\ u_t \end{bmatrix}\right] \quad \text{s.t.} \quad z_t = Az_{t-1}+Bu_t+W_t, \quad u_t = \pi_t(\mathcal{I}_{t-1})$$

where $z_t = [x_t \;\; d_t^\top \;\; f_t^\top]^\top$, $v = [0 \;\; \gamma^{*\top}(\mathrm{diag}(r) - I) \;\; g^\top]^\top$, $\gamma^* = [\gamma_1^* \;\; \cdots \;\; \gamma_M^*]^\top$, $e_1 = [1 \;\; 0 \;\; \cdots \;\; 0]^\top$,

$$Q = \rho\Sigma_\epsilon e_1 e_1^\top - \frac{1}{2}(ve_1^\top + e_1 v^\top), \quad S = \rho\Sigma_\epsilon e_1 - \frac{1}{2}(\lambda^* + \gamma^{*\top}\mathbf{1})e_1, \quad R = \rho\Sigma_\epsilon,$$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \mathrm{diag}(r) & 0 \\ 0 & 0 & \Phi \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ \mathbf{1} \\ 0 \end{bmatrix}, \quad W_t = \begin{bmatrix} 0 \\ 0 \\ \omega_t \end{bmatrix}, \quad \tilde{\Omega} \triangleq \mathrm{Cov}[W_t] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Omega \end{bmatrix}.$$

Note that R is strictly positive but Q is not necessarily positive semidefinite. Therefore, special care should be taken in order to prove existence of an optimal policy. We start with a well-known Bellman equation for average-cost linear quadratic control problems

$$H(z_{t-1}) + h = \min_{u_t} \mathsf{E}\left[\rho\Sigma_\epsilon(x_{t-1}+u_t)^2 - \Delta p_t x_{t-1} + H(z_t)\right] \tag{3}$$

where $H(\cdot)$ denotes a differential value function and h denotes minimum average cost. It is natural to conjecture $H(z_t) = z_t^\top P z_t$. Plugging it into (3), we can obtain a discrete-time Riccati algebraic equation

$$P = A^\top PA + Q - (S^\top + B^\top PA)^\top (R + B^\top PB)^{-1}(S^\top + B^\top PA) \tag{4}$$

with a second-order optimality condition $R + B^\top PB > 0$. The following theorem characterizes an optimal policy among admissible policies that minimizes expected average cost, and proves existence of such an optimal policy.

Theorem 1. *For any $\theta^* \in \Theta$, there exists a unique symmetric solution P to (4) that satisfies $R+B^\top PB > 0$ and $\rho_{sr}(A+BL) < 1$ where $L = -(R+B^\top PB)^{-1}(S^\top + B^\top PA)$ and $\rho_{sr}(\cdot)$ denotes a spectral radius. Moreover, a policy $\pi = (\pi_1, \pi_2, \ldots)$ with $\pi_t(\mathcal{I}_{t-1}) = Lz_{t-1}$ is an optimal policy among admissible policies that attains minimum expected average cost $tr(P\tilde{\Omega})$.*

For ease of exposition, we define some notations: $P(\theta)$ denotes a unique symmetric stabilizing solution to (4) with $\theta^* = \theta$. $L(\theta) \triangleq -(R+B^\top P(\theta)B)^{-1}(S(\theta)^\top + B^\top P(\theta)A)$ denotes a gain matrix for an optimal policy with $\theta^* = \theta$, $G(\theta) \triangleq A + BL(\theta)$ denotes a closed-loop system matrix with $\theta^* = \theta$, and $U(\theta) \triangleq \mathbf{1}L(\theta) + [A - I \;\; O]$ denotes a linear mapping from z_{t-1} to a regressor ψ_t used in least-squares regression for learning price impact, i.e., $\psi_t = U(\theta)z_{t-1}$. Having these notations, we make two assumptions about $L(\theta)$ as follows:

Assumption 2. *(a) $(L(\theta))_1 \neq 0$ and $(L(\theta))_{M+2} \neq 0$ for any $\theta \in \Theta$*

(b) There exists $C_L > 0$ such that $\|L(\theta_1) - L(\theta_2)\| \leq C_L\|\theta_1 - \theta_2\|$ for any $\theta_1, \theta_2 \in \Theta$.

In some special cases, e.g., a case only with permanent price impact that will be discussed in Section 2.3, we can prove the above two assumptions to hold. Together with Assumption 1, Assumption 2 will be taken to hold for all results stated on the remainder of the paper.

2.3 Closed-Form Solution: A Single Factor and Permanent Impact Only

When we consider only the permanent price impact and a single observable factor, we can derive an exact closed-form P and L as follows.

$$P_{xx} = \frac{\lambda^* - \rho\Sigma_\epsilon + \sqrt{2\lambda^*\rho\Sigma_\epsilon + (\rho\Sigma_\epsilon)^2}}{2}$$

$$P_{xf} = \frac{-g\lambda^*}{(1-\Phi)\lambda^* - \Phi\rho\Sigma_\epsilon + \Phi\sqrt{2\lambda^*\rho\Sigma_\epsilon + (\rho\Sigma_\epsilon)^2}}$$

$$P_{ff} = \frac{-g^2\Phi^2}{2(1-\Phi^2)\left((1-\Phi)^2\lambda^* + (1+\Phi^2)\rho\Sigma_\epsilon + (1-\Phi^2)\sqrt{2\lambda^*\rho\Sigma_\epsilon + (\rho\Sigma_\epsilon)^2}\right)}$$

$$L_x = \frac{-2\rho\Sigma_\epsilon}{\rho\Sigma_\epsilon + \sqrt{2\lambda^*\rho\Sigma_\epsilon + (\rho\Sigma_\epsilon)^2}}$$

$$L_f = \frac{g\Phi}{(1-\Phi)\lambda^* + \rho\Sigma_\epsilon + \sqrt{2\lambda^*\rho\Sigma_\epsilon + (\rho\Sigma_\epsilon)^2}}$$

Although this is a special case of our general setting, we can get useful insights into the effect of permanent price impact coefficient λ^* on various quantities. Here are some examples:

- $|L_x|$ and $|L_f|$ are strictly decreasing in λ^*.
- $\lim_{\lambda^*\to 0} L_x = -1$, $\lim_{\lambda^*\to\infty} L_x = 0$.
- $\lim_{\lambda^*\to 0} L_f = g\Phi/(2\rho\Sigma_\epsilon)$, $\lim_{\lambda^*\to\infty} L_f = 0$.
- The expected average risk-adjusted profit $-P_{ff}\Omega$ is strictly decreasing in λ^*.
- $\lim_{\lambda^*\to 0}(-P_{ff}\Omega) = g^2\Phi^2\Omega/(4(1-\Phi^2)\rho\Sigma_\epsilon)$, $\lim_{\lambda^*\to\infty}(-P_{ff}\Omega) = 0$.

2.4 Performance Measure: Regret

In this subsection, we define performance measures that can be used to evaluate policies. For notational simplicity, let $L^* = L(\theta^*)$, $G^* = G(\theta^*)$ and $P^* = P(\theta^*)$. Using (4), we can show that

$$\begin{aligned}
J_T^\pi(z_0|\mathcal{F}_T) &\triangleq \sum_{t=1}^T \left\{\rho\Sigma_\epsilon(x_{t-1} + \pi_t(\mathcal{I}_{t-1}))^2 - \Delta p_t x_{t-1}\right\} \\
&= z_0^\top P^* z_0 - z_T^\top P^* z_T + 2\sum_{t=1}^T (Az_{t-1} + B\pi_t(\mathcal{I}_{t-1}))^\top P^* W_t + \sum_{t=1}^T W_t^\top P^* W_t - \sum_{t=1}^T x_{t-1}\epsilon_t \\
&\quad + \sum_{t=1}^T (\pi_t(\mathcal{I}_t) - L^* z_{t-1})^\top (R + B^\top P^* B)(\pi_t(\mathcal{I}_{t-1}) - L^* z_{t-1}) \quad \text{for any policy } \pi.
\end{aligned}$$

First, we define *pathwise regret* $R_T^\pi(z_0|\mathcal{F}_T)$ of a policy π at period T as $J_T^\pi(z_0|\mathcal{F}_T) - J_T^{\pi^*}(z_0|\mathcal{F}_T)$ where $\pi_t^*(\mathcal{I}_{t-1}) = L^* z_{t-1}^*$ and $z_t^* = G^* z_{t-1}^* + W_t$ with $z_0^* = z_0$. In other words, the pathwise regret of a policy π at period T amounts to excess costs accumulated over T periods when applying π relative to when applying the optimal policy π^*. By definition of π^*, the pathwise regret of a policy

π at period T can be expressed as

$$R_T^\pi(z_0|\mathcal{F}_T) = z_T^{*\top} P^* z_T^* - z_T^\top P^* z_T + \sum_{t=1}^{T} (\pi_t(\mathcal{I}_{t-1}) - L^* z_{t-1})^\top (R + B^\top P^* B)(\pi_t(\mathcal{I}_{t-1}) - L^* z_{t-1})$$
$$+ 2\sum_{t=1}^{T} ((Az_{t-1} + B\pi_t(\mathcal{I}_{t-1})) - (A + BL^*) z_{t-1}^*)^\top P^* W_t + \sum_{t=1}^{T} (x_{t-1}^* - x_{t-1})\epsilon_t.$$

Second, we define *expected regret* $\bar{R}_T^\pi(z_0)$ of a policy π at period T as $\mathsf{E}[R_T^\pi(z_0|\mathcal{F}_T)]$. Taking expectation of pathwise regret, we can obtain a more concise expression for expected regret because the last two terms vanish by the law of total expectation. Hence, we have

$$\bar{R}_T^\pi(z_0) = \mathsf{E}[z_T^{*\top} P^* z_T^* - z_T^\top P^* z_T] + \mathsf{E}\left[\sum_{t=1}^{T} (\pi_t(\mathcal{I}_{t-1}) - L^* z_{t-1})^\top (R + B^\top P^* B)(\pi_t(\mathcal{I}_{t-1}) - L^* z_{t-1})\right].$$

Finally, we define *relative regret* $\tilde{R}_T^\pi(z_0)$ of a policy π at period T as $\bar{R}_T^\pi(z_0)/|\mathrm{tr}(P^*\tilde{\Omega})|$ where $\mathrm{tr}(P^*\tilde{\Omega})$ is minimum expected average cost for θ^*. In the remainder of this paper, we will study algorithms through the lenses of expected regret and relative regret.

3 Confidence-Triggered Regularized Adaptive Certainty Equivalent Policy

Our problem can be viewed as a special case of reinforcement learning, which focuses on sequential decision-making problems in which unknown properties of an environment must be learned in the course of taking actions. It is often emphasized in reinforcement learning that longer-term performance can be greatly improved by making decisions that explore the environment efficiently at the expense of suboptimal short-term behavior. In our problem, a price impact model is unknown, and submission of large orders can be considered exploratory actions that facilitate learning.

Certainty equivalent control (CE) represents one extreme where at any time, current point estimates are assumed to be correct and actions are made accordingly. Although learning is carried out with observations made as the system evolves, no decisions are designed to enhance learning. Thus, this is an instance of pure exploitation of current knowledge. In our problem, CE estimates the unknown price impact coefficients θ^* at each period via least-squares regression using available data, and makes decisions that maximize expected average risk-adjusted profit under an assumption

Algorithm 1 CTRACE

Input: θ_0, x_0, d_0, r, g, κ, C_v, τ, $L(\cdot)$, $\theta_{\max}$, $\{p_t\}_{t=0}^{\infty}$, $\{f_t\}_{t=0}^{\infty}$
Output: $\{u_t\}_{t=1}^{\infty}$

$V_0 \leftarrow \kappa I$, $t_0 \leftarrow 0$, $i \leftarrow 1$
for $t = 1, 2, \ldots$ **do**
  $u_t \leftarrow L(\theta_{t-1}) z_{t-1}$, $x_t \leftarrow x_{t-1} + u_t$, $d_t \leftarrow \mathrm{diag}(r) d_{t-1} + \mathbf{1} u_t$
  $\psi_t \leftarrow [u_t \ \ (d_t - d_{t-1})^\top]^\top$, $V_t \leftarrow V_{t-1} + \psi_t \psi_t^\top$
  if $\lambda_{\min}(V_t) \geq \kappa + C_v t$ and $t \geq t_{i-1} + \tau$ **then**
    $\theta_t \leftarrow \mathrm{argmin}_{\theta \in \Theta} \sum_{i=1}^{t} \left((\Delta p_i - g^\top f_{i-1}) - \psi_i^\top \theta \right)^2 + \kappa \|\theta\|^2$, $t_i \leftarrow t$, $i \leftarrow i + 1$
  else
    $\theta_t \leftarrow \theta_{t-1}$
  end if
end for

that the estimated model is correct. That is, an action u_t for CE is given by $u_t = L(\theta_{t-1}^{\mathrm{CE}}) z_{t-1}$ where $\theta_{t-1}^{\mathrm{CE}} = \mathrm{argmin}_{\theta \in \Theta} \sum_{i=1}^{t-1} \left((\Delta p_i - g^\top f_{i-1}) - \psi_i^\top \theta \right)^2$ with a regressor $\psi_i = [u_i \ \ (d_i - d_{i-1})^\top]^\top$.

An important question is how aggressively the trader should explore to learn θ^*. Unlike many other reinforcement learning problems, a fairly large amount of exploration is naturally induced by exploitative decisions in our problem. That is, regular trading activity triggered by the return-predictive factors f_t excites the market regardless of whether or not she aims to learn price impact. Given sufficiently large factor variability, the induced exploration might adequately resolve uncertainties about price impact. However, we will demonstrate by taking a constructive approach that executing trades to explore beyond what would naturally occur through the factor-driven exploitation can result in significant benefit. More precisely, we propose the *confidence-triggered regularized adaptive certainty equivalent* policy (CTRACE) as presented in Algorithm 1 and derive its poly-logarithmic finite-time expected regret bound of $O(\log^2 T)$ in Theorem 4.

Now, let us first focus on important observation that exploitative actions triggered by the return-predictive factors induce a large degree of exploration that could yield strong consistency of least-squares estimates. It is worth noting that pure exploitation is not sufficient for strong consistency in other problems such as Lai and Wei [1986] and Chen and Guo [1986].

Lemma 1. *For any $\theta \in \Theta$, let $u_t = L(\theta) z_{t-1}$, $z_t = G(\theta) z_{t-1} + W_t$ and $\psi_t^\top = \left[u_t \ \ (d_t - d_{t-1})^\top \right] = (U(\theta) z_{t-1})^\top$. Also, let $\Pi_{zz}(\theta)$ denote a unique solution to $\Pi_{zz}(\theta) = G(\theta) \Pi_{zz}(\theta) G(\theta)^\top + \tilde{\Omega}$. Then,*

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \psi_t \psi_t^\top = U(\theta) \Pi_{zz}(\theta) U(\theta)^\top \succ 0 \ \ a.s. \tag{5}$$

In words, if a fixed θ is used constantly over time when computing optimal actions, the corresponding least-squares estimate $\hat{\theta}_t$ is strongly consistent because Lai and Wei [1982] show $\|\hat{\theta}_T - \theta^*\| \leq O\left(\left(\log \lambda_{\max}(V_T)/\lambda_{\min}(V_T)\right)^{1/2}\right)$ a.s. where $V_T \triangleq \sum_{t=1}^{T} \psi_t \psi_t^\top$ and Lemma 1 implies $O\left(\left(\log \lambda_{\max}(V_T)/\lambda_{\min}(V_T)\right)^{1/2}\right) = O\left(\left(\log T/T\right)^{1/2}\right)$. Note that this cannot be applied to CE since it uses an updated estimate at every period when computing an optimal action.

Moreover, we can show that $\Pi_{zz}(\theta)$ is continuous on Θ by proving uniform convergence of $\mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T} z_{t-1} z_{t-1}^\top\right]$ to $\Pi_{zz}(\theta)$ on Θ. Continuity leads to $\underline{\lambda}^*_{\psi\psi} \triangleq \inf_{\theta\in\Theta} \lambda_{\min}\left(U(\theta)\Pi_{zz}(\theta)U(\theta)^\top\right) > 0$ which will be used later.

Corollary 1. *$\Pi_{zz}(\theta)$ is continuous on Θ and $\underline{\lambda}^*_{\psi\psi} \triangleq \inf_{\theta\in\Theta} \lambda_{min}\left(U(\theta)\Pi_{zz}(\theta)U(\theta)^\top\right) > 0$.*

Lemma 1 implies that $\lambda_{\min}\left(\sum_{t=1}^{T} \psi_t \psi_t^\top\right)$ increases linearly in time T a.s. asymptotically. In addition, we can obtain a similar result for a finite-sample case: There exists a finite, deterministic constant $T_1(\theta, \delta)$ such that $\lambda_{\min}\left(\sum_{t=1}^{T} \psi_t \psi_t^\top\right)$ grows linearly in time T for all $T \geq T_1(\theta, \delta)$ with probability at least $1-\delta$. This is a crucial result that will be used for bounding above "ϵ-convergence time" later. It is formally stated in the following lemma.

Lemma 2. *For any $\theta \in \Theta$, let $u_t = L(\theta) z_{t-1}$, $z_t = G(\theta) z_{t-1} + W_t$ and $\psi_t^\top = \left[u_t \ (d_t - d_{t-1})^\top\right] = (U(\theta) z_{t-1})^\top$. Then, there exists an event $\mathcal{B}(\delta)$ such that on $\mathcal{B}(\delta)$ with $Pr(\mathcal{B}(\delta)) \geq 1-\delta$*

$$\frac{7}{8} U(\theta)\Pi_{zz}(\theta)U(\theta)^\top \preceq \frac{1}{T}\sum_{t=1}^{T} \psi_t \psi_t^\top \preceq \frac{17}{16} U(\theta)\Pi_{zz}(\theta)U(\theta)^\top \quad \textit{for all } T \geq T_1(\theta, \delta)$$

where $T_1(\theta, \delta)$ is a finite, deterministic constant of which explicit expression can be found in e-companion.

Furthermore, we can extend Lemma 1 in such a way that $\lambda_{\min}\left(\sum_{t=1}^{T} \psi_t \psi_t^\top\right)$ still increases to infinity linearly in time T for time-varying $\{\theta_t\}$ adapted to $\{\sigma(\mathcal{I}_t)\}$ as long as θ_t remains sufficiently close to a fixed $\theta \in \Theta$ for all $t \geq 0$. Here, $\sigma(\mathcal{I}_t)$ denotes a σ-algebra generated by $\mathcal{I}_t$ and θ_t is $\sigma(\mathcal{I}_t)$-measurable for each t.

Lemma 3. *Consider any $\theta \in \Theta$ and $\{\theta_t \in \Theta\}$ adapted to $\{\sigma(\mathcal{I}_t)\}$ such that $\|\theta_t - \theta\| \leq \frac{\eta}{\sqrt{M+1} C_L}$ a.s. for all $t \geq 0$ where η is a sufficiently small number of which explicit expression can be found in e-companion. Let $u_t = L(\theta_{t-1}) z_{t-1}$, $z_t = G(\theta_{t-1}) z_{t-1} + W_t$ and $\psi_t^\top = \left[u_t \ (d_t - d_{t-1})^\top\right] =$*

$(U(\theta_{t-1})z_{t-1})^\top$. *Then,*

$$\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\psi_t\psi_t^\top \succeq \frac{\lambda_{min}(U(\theta)\Pi_{zz}(\theta)U(\theta)^\top)}{2}I \ \ a.s. \tag{6}$$

Similarly to Lemma 2, we can obtain a finite-sample result for Lemma 3. This result will provide with a useful insight into how our new exploratory policy operates in the long term.

Lemma 4. *Consider* $\{\theta_t \in \Theta\}$ *defined in Lemma 3. Let* $u_t = L(\theta_{t-1})z_{t-1}$, $z_t = G(\theta_{t-1})z_{t-1} + W_t$ *and* $\psi_t^\top = \begin{bmatrix} u_t & (d_t - d_{t-1})^\top \end{bmatrix} = (U(\theta_{t-1})z_{t-1})^\top$. *Then, for any* $0 < \delta < 1$ *on the event* $\mathcal{B}(\delta)$ *in Lemma 2*

$$\lambda_{min}\left(\frac{1}{T}\sum_{t=1}^{T}\psi_t\psi_t^\top\right) \geq \frac{3}{8}\lambda_{min}(U(\theta)\Pi_{zz}(\theta)U(\theta)^\top) \quad \textit{for all } T \geq T_1(\theta,\delta) \vee \frac{3\|z_0\|(2C_\omega + \|z_0\|)}{C_\omega^2}.$$

It is challenging to guarantee that all estimates generated by CE are sufficiently close to one another uniformly over time so that Lemma 3 and Lemma 4 can be applied to CE. In particular, CE is subject to overestimation of price impact that could be considerably detrimental to trading performance. The reason is that overestimated price impact discourages submission of large orders and thus it might take a while for the trader to realize that price impact is overestimated due to reduced "signal-to-noise ratio." To address this issue, we propose the *confidence-triggered regularized adaptive certainty equivalent* policy (CTRACE) as presented in Algorithm 1. CTRACE can be viewed as a generalization of CE and deviates from CE in two ways: (1) ℓ_2 regularization is applied in least-squares regression, (2) coefficients are only updated when a certain measure of confidence exceeds a pre-specified threshold and a minimum inter-update time has elapsed. Note that CTRACE reduces to CE as the regularization penalty κ and the threshold C_v tend to zero, and the minimum inter-update time τ tends to one.

Regularization induces active exploration in our problem by penalizing the ℓ_2-norm of price impact coefficients as well as reduces the variance of an estimator. Without regularization, we are more likely to obtain overestimates of price impact. Such an outcome attenuates trading intensity and thereby makes it difficult to escape from the misjudged perspective on price impact. Regularization decreases the chances of obtaining overestimates by reducing the variance of an estimator and furthermore tends to yield underestimates that encourage active exploration.

Another source of improvement of CTRACE relative to CE is that updates are made based on a certain measure of confidence for estimates whereas CE updates at every period regardless of confidence. To be more precise on this confidence measure, we first present a high-probability confidence region for least-squares estimates from Abbasi-Yadkori et al. [2011].

Proposition 1 (Corollary 10 of Abbasi-Yadkori et al. [2011])**.**

$$Pr(\theta^* \in \mathcal{S}_t(\delta),\ \forall t \geq 1) \geq 1-\delta \quad where$$

$$V_t = \kappa I + \sum_{i=1}^{t} \psi_i \psi_i^\top, \quad \hat{\theta}_t = V_t^{-1}\left(\sum_{i=1}^{t} \psi_i \psi_i^\top \theta^* + \sum_{i=1}^{t} \psi_i \epsilon_i\right),$$

$$\mathcal{S}_t(\delta) \triangleq \left\{\theta \in \mathbb{R}^{M+1} : (\theta - \hat{\theta}_t)^\top V_t (\theta - \hat{\theta}_t) \leq \left(C_\epsilon \sqrt{2\, log\left(\frac{det(V_t)^{1/2} det(\kappa I)^{-1/2}}{\delta}\right)} + \kappa^{1/2}\|\theta_{max}\|\right)^2\right\}.$$

This implies that for any $\theta \in \mathcal{S}_t(\delta)$

$$\|\theta - \hat{\theta}_t\|^2 \leq \frac{1}{\lambda_{\min}(V_t)}\left(C_\epsilon \sqrt{2\,\log\left(\frac{\det(V_t)^{1/2}\det(\kappa I)^{-1/2}}{\delta}\right)} + \kappa^{1/2}\|\theta_{\max}\|\right)^2.$$

By definition, CTRACE updates only when $\lambda_{\min}(V_t) \geq \kappa + C_v\, t$. $\lambda_{\min}(V_t)$ typically dominates $\log(\det(V_t))$ for large t because it increases linearly in t, and is inversely proportional to the squared estimation error $\|\hat{\theta}_t - \theta^*\|^2$. That is, CTRACE updates only when confidence represented by $\lambda_{\min}(V_t)$ exceeds the specified level $\kappa + C_v\, t$. From now on, we refer to this updating scheme as confidence-triggered update. Confidence-triggered update makes a significant contribution to reducing the chances of obtaining overestimates of price impact by updating "carefully" only at the moments when an upper bound on the estimation error is guaranteed to decrease.

The minimum inter-update time $\tau \in \mathbb{N}$ in Algorithm 1 can guarantee that the closed-loop system $\{z_t\}$ from CTRACE is stable as long as τ is sufficiently large. Meanwhile, there is no such stability guarantee for CE. The following lemma provides with a specific uniform bound on $\|z_t\|$.

Lemma 5. *For all $t \geq 0$, under CTRACE with $\tau \geq N \log(2C_g/\xi)/\log(1/\xi)$*

$$\|z_t\| \leq \frac{(2C_g+1)C_g C_\omega}{\xi(1-\xi^{\frac{1}{N}})} \triangleq C_z^*\ \ a.s. \quad and \quad \|\psi_t\| \leq \frac{(C_g+1)(2C_g+1)C_g C_\omega}{\xi(1-\xi^{\frac{1}{N}})} \triangleq C_\psi\ \ a.s.$$

Confidence-triggered update yields a good property of CTRACE that CE lacks: CTRACE is *inter-temporally consistent* in the sense that estimation errors $\|\theta_t - \theta^*\|$ are bounded with high probability by monotonically nonincreasing upper bounds that converge to zero almost surely as time tends to infinity. The following theorem formally states this property.

Theorem 2 (Inter-temporal Consistency of CTRACE). *Let $\{\theta_t\}$ be estimates generated by CTRACE with $M \geq 2$, $\tau \geq N \log(2C_g/\xi)/\log(1/\xi)$ and $C_v < \underline{\lambda}^*_{\psi\psi}$. Then, the ith update time t_i in Algorithm 1 is finite a.s. Moreover, $\|\theta_t - \theta^*\| \leq b_t,\ \forall t \geq 0$ on the event $\{\theta^* \in \mathcal{S}_t(\delta),\ \forall t \geq 1\}$ where*

$$b_t = \begin{cases} \frac{2C_\epsilon\sqrt{(M+1)\log\left(C_\psi^2\, t/\kappa + M + 1\right) + 2\log(1/\delta)} + 2\kappa^{1/2}\|\theta_{max}\|}{\sqrt{C_v t}} & \text{if } t = t_i \text{ for some } i \\ b_{t-1} & \text{otherwise} \end{cases}, \quad b_0 = \|\theta_0 - \theta^*\|,$$

and $\{b_t\}$ is monotonically nonincreasing for all $t \geq 1$ with $\lim_{t\to\infty} b_t = 0$ a.s.

Moreover, we can show that CTRACE is *efficient* in the sense that its ϵ-convergence time is bounded above by a polynomial of $1/\epsilon$ and $\log(1/\delta)$ with probability at least $1 - \delta$. We define *ϵ-convergence time* to be the first time when an estimate and all the future estimates following it are within an ϵ-neighborhood of θ^*. If ϵ is sufficiently small, we can apply Lemma 3 and 4 to guarantee that $\lambda_{\min}(V_t)$ increases linearly in t with high probability after ϵ-convergence time and thereby confidence-triggered update occurs at every τ periods. This is a critical property that will be used for deriving a poly-logarithmic finite-time expected regret bound for CTRACE. By Theorem 2, it is easy to see that the ϵ-convergence time of CTRACE is bounded above by $t_{N(\epsilon,\delta,C_v)}$ with probability at least $1 - \delta$ where $N(\epsilon, \delta, C_v)$ is defined as $\inf\{i \in \mathbb{N} : b_{t_i} \leq \epsilon\}$. The following theorem presents the polynomial bound on the ϵ-convergence time of CTRACE.

Theorem 3 (Efficiency of CTRACE). *For any $\epsilon > 0$, $0 < \delta, \delta' < 1$, $\tau \geq N \log(2C_g/\xi)/\log(1/\xi)$ and $C_v < \frac{7}{8}\underline{\lambda}^*_{\psi\psi}$ on the event $\mathcal{B}(\delta')$ defined in Lemma 2,*

$$t_{N(\epsilon,\delta,C_v)} \leq T_1^*(\delta') \vee \tau + T_2(\epsilon, \delta, C_v)$$

where explicit expressions for $T_1^*(\delta')$ and $T_2(\epsilon, \delta, C_v)$ can be found in Appendix and e-companion.

Finally, we derive a finite-time expected regret bound for CTRACE that is quadratic in loga-

rithm of elapsed time using the efficiency of CTRACE and Lemma 4.

Theorem 4 (Finite-Time Expected Regret Bound of CTRACE). *If π is CTRACE with $M \geq 2$, $\tau \geq N \log(2C_g/\xi)/\log(1/\xi)$ and $C_v < \frac{7}{8}\underline{\lambda}^*_{\psi\psi}$, then for any $\nu \in (\xi, 1)$ and all $T \geq 2$,*

$$\bar{R}^{\pi}_T(z_0) = O(\log^2 T)$$

where explicit expression for the finite-time upper bound can be found in Appendix and e-companion.

4 Computational Analysis

In this section, we will compare via Monte Carlo simulation the performance of CTRACE to that of two benchmark policies: CE and a reinforcement learning algorithm recently proposed in Abbasi-Yadkori and Szepesvàri [2010], which is referred to as AS policy from now on. AS policy was designed to explore efficiently in a broader class of linear-quadratic control problems and appears well-suited for our problem. It updates an estimate only when the determinant of V_t is at least twice as large as the determinant evaluated at the last update, and selects an element from a high-probability confidence region that yields maximum average reward, that is, *optimism in the face of uncertainty*. In our problem, AS policy can translate to updating an estimate with $\theta_t = \text{argmin}_{\theta \in \mathcal{S}_t(\delta) \cap \Theta} \text{tr}(P(\theta)\tilde{\Omega})$ at each update time t. Intuitively, the smaller price impact, the larger average profit, equivalently, the smaller $\text{tr}(P(\theta)\tilde{\Omega})$ which is the negative of average profit. In light of this, we restrict our attention to solutions to $\min_{\theta \in \mathcal{S}_t(\delta) \cap \Theta} \text{tr}(P(\theta)\tilde{\Omega})$ of the form $\{\alpha_t \hat{\theta}_{con,t} \in \mathcal{S}_t(\delta) \cap \Theta : 0 \leq \alpha_t \leq 1\}$ where $\hat{\theta}_{con,t}$ denotes a constrained least-squares estimate to Θ with ℓ_2 regularization. The motivation is to reduce the amount of computation needed for AS policy otherwise it would be prohibitive. Indeed, the minimum appears to be attained always with the smallest α_t such that $\alpha_t \hat{\theta}_{con,t} \in \mathcal{S}_t(\delta) \cap \Theta$, which is provable in the special case considered in Subsection 2.3. Note that α_t can be viewed as a measure of aggressiveness of exploration: $\alpha_t = 1$ means no extra exploration and smaller α_t implies more active exploration.

Table 1: Monte Carlo simulation setting (1 trading day = 6.5 hours)

M	6	K	2
Trading interval	5 mins	Initial asset price	\$50
Half-life of r	[5, 7.5, 10, 15, 30, 45] mins	Half life of factor	[10, 40] mins
r	[0.50, 0.63, 0.71, 0.79, 0.89, 0.93]	Φ	diag([0.707, 0.917])
γ (\$/share)	$[0, 6, 0, 3, 7, 5] \times 10^{-8}$	λ (\$/share)	2×10^{-8}
Σ_ϵ	0.0013 (annualized vol. = 10%)	Ω	diag([1, 1])
ρ	1×10^{-6}	$\theta_{\max}$	$(5 \times 10^{-7})\mathbf{1}$
β	5×10^{-9}	g	[0.006, 0.002]
T	3000 ($\approx$ 38 trading days)	Sample paths	600

4.1 Simulation Setting

Table 1 summarizes numerical values used in our simulation. The signal-to-noise ratio (SNR), which is defined as $\mathsf{E}[(\lambda u_t + \sum_{m=1}^{M} \gamma_m (d_{m,t} - d_{m,t-1}))^2] / \mathsf{E}[\epsilon_t^2]$ under optimal trades $u_t = L(\theta^*) z_{t-1}$, is 0.058 and the optimal average profit is \$765.19 per period. ϵ_t and ω_t are sampled independently from Gaussian distribution even though ω_t is assumed to be bounded almost surely for the theoretical analysis. In fact, it turns out that the use of Gaussian distribution for ω_t does not make a noticeable difference from a bounded case.

Since we do not have access to real transaction data for the purpose of academic research, we infer plausible parameter values from other empirical work on price impact estimation as an alternative. Firstly, Obizhaeva [2012] estimates permanent price impact in a bias-free way using proprietary portfolio transition data. A unique feature of this data allowing for bias-free estimation is that full execution schedule for portfolio transition is fixed prior to the beginning of execution and thus independent of future price movements and other market conditions. The author's estimation is based on the following price impact model:

$$\frac{p_t - p_{t-1}}{p_{t-1}} = 10^{-4} \cdot \frac{\sigma_{\text{daily}}}{0.02} \cdot \frac{\lambda_{\text{rel}}}{2} \cdot \frac{Q}{(0.01)ADV} + (\text{term related to bid-ask spread}) \tag{7}$$

where σ_{daily} denotes daily volatility, ADV average daily volume in dollar and λ_{rel} relative permanent price impact coefficient, which is estimated 0.3. We reconstruct values corresponding to λ in our model as shown in the last row of Table 2. Recall that we choose $\lambda = 2 \times 10^{-8}$ in our example, which lies between Group 1 and Group 2, and $\gamma = [0, 6, 0, 3, 7, 5] \times 10^{-8}$, each non-zero value of which is close to Group 1. Therefore, our choices of λ and γ are consistent with those of stocks

	Group 1	Group 2	Group 3	Group 4	Group 5
Daily Volatility (%)	2.04	2.00	1.92	1.95	1.88
Average Daily Volume ($MM)	1.22	5.14	9.97	15.92	23.92
λ ($\times 10^{-8}$)	6.27	1.46	0.72	0.46	0.29
	Group 6	Group 7	Group 8	Group 9	Group 10
Daily Volatility (%)	1.85	1.79	1.78	1.76	1.76
Average Daily Volume ($MM)	31.45	42.11	60.16	101.51	212.55
λ ($\times 10^{-8}$)	0.22	0.16	0.11	0.07	0.03

Table 2: Permanent price impact estimation by Obizhaeva [2012]: Estimated median daily volatility and median average daily volume from Obizhaeva [2012]. The last row corresponds to λ in our example obtained from the first two rows and initial price \$50. Each month, observations are split into 10 bins according to stocks' dollar trading volume in pre-transition month. The thresholds are 30th, 50th, 60th, 70th, 75th, 80th, 85th, 90th and 95th percentiles of dollar trading volume for common stocks listed on the NYSE. Group 1 (Group 10) contains orders in stocks with lowest (highest) dollar trading volume. The sample ranges from January 2001 to December 2005.

with low dollar trading volume.

Secondly, we infer plausible values for decay rates of transient price impact from the vector autoregressive model of returns and trades in Dufour and Engle [2000]. The authors define permanent price impact as the limiting value of impulse response to the update from private information which is gleaned from unexpected trades. Extending this definition, we define transient price impact as impulse response to the update to public information. Figure 1 shows simulated impulse responses for Fannie Mae and AT&T obtained similarly to Figure 1 and 2 in Dufour and Engle [2000]. Note that the two stocks have half-lives of 26.9 and 22.5 minutes, respectively, which are comparable to about 20 minutes in our example. It implies that our choices of decay rates and γ in Table 1 produce transient price impact having a half-life that is consistent with the empirical study based on real transaction data.

Finally, we choose regularization coefficient κ, confidence-triggered update threshold C_v and significance level δ via cross-validation. For simplicity, we set minimum inter-update time τ to be one. A challenge for cross-validation in our case is that joint distribution of response and predictor variables is not independent of the parameters κ, C_v and δ as opposed to typical supervised learning problems with independent, identically distributed data. Suppose that a trader has historical transactions data $\{p_t^{\text{hist}}, u_t^{\text{hist}}\}_{t=1}^T$ obtained using $C_v = 100$ and $\kappa = 10^{10}$. Difficulty arises with the fact that optimal trades corresponding to $C_v = 200$ and $\kappa = 10^{10}$ would be different from $\{u_t^{\text{hist}}\}$ and thereby the resulting transaction prices would be different from $\{p_t^{\text{hist}}\}$. However, these new

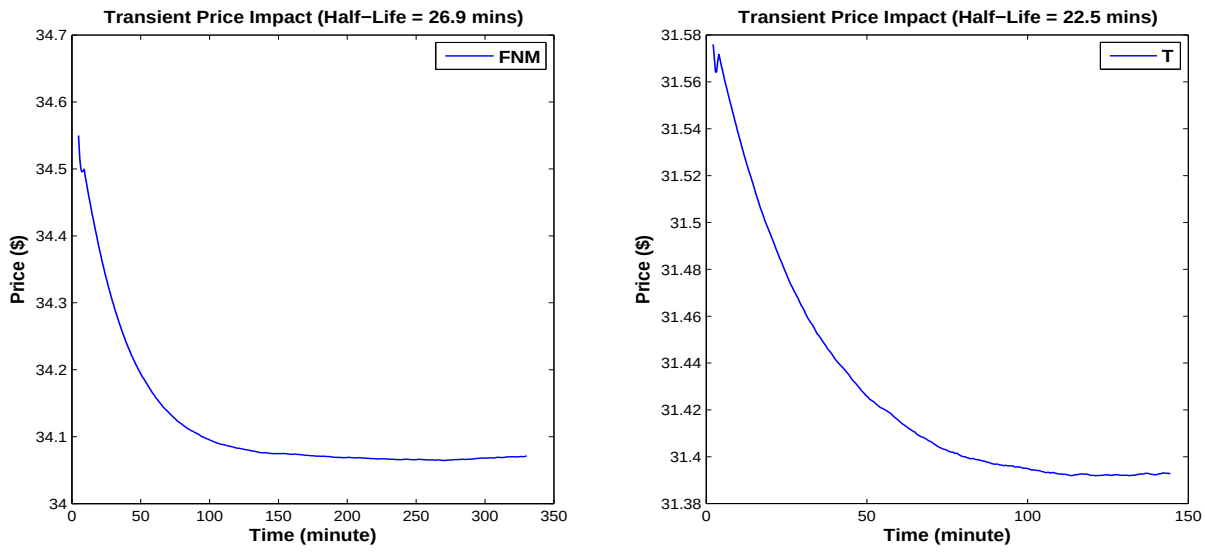

Figure 1: Reconstructed transient price impact for (a) Fannie Mae (half-life = 26.9 mins) and (b) AT&T (half-life = 22.5 mins) using the vector autoregressive model in Dufour and Engle [2000]

prices cannot be computed without knowing true price dynamics.

To circumvent this, we consider another plausible regime with different $\gamma = [1, 5, 1, 4, 8, 2] \times 10^{-8}$ and carry out cross-validation as if the estimated price dynamics were true. This approach would work only if optimal κ, C_v and δ are not very sensitive to different choices of γ. We confirm that this is indeed the case and Figure 5 in Section 4.3 can be viewed as supporting examples. Consequently, $\kappa = 1 \times 10^{11}$ and $C_v = 600$ are selected for CTRACE and $\kappa = 1 \times 10^{8}$ and $\delta = 0.99$ for AS policy. The reason for smaller κ and δ being close to 1 for AS policy is to keep the radius of confidence regions small because the exploration done by AS policy tends to be more than necessary and thus costly.

4.2 Simulation Results

Figure 2(a) illustrates improvement of relative regret due to regularization. It shows the relative regret of CTRACE with varying κ and fixed $C_v = 0$, i.e. no confidence-triggered update. The vertical bars indicate two standard errors in each direction. It is clear that the relative regret is reduced as CTRACE regularizes more, and improvement from no regularization to $\kappa = 1 \times 10^{11}$ is statistically significant at 95% confidence level. Figure 2(b) shows improvement achieved by confidence-triggered update with varying C_v but fixed $\kappa = 1 \times 10^{11}$. As you can see, update based on confidence makes a substantial contribution to reducing relative regret further. Improvement

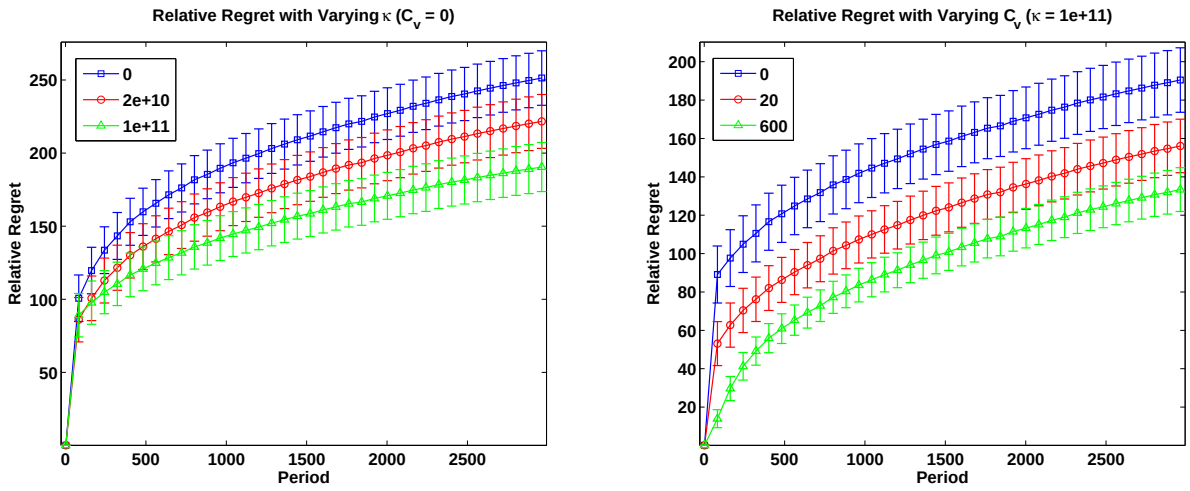

Figure 2: Relative regret with varying κ and C_v: (a) Varying $\kappa \in \{0, 2 \times 10^{10}, 1 \times 10^{11}\}$ with fixed $C_v = 0$. (b) Varying $C_v \in \{0, 20, 600\}$ with fixed $\kappa = 1 \times 10^{11}$.

from $C_v = 0$ to $C_v = 600$ is also statistically significant at 95% confidence level.

As shown in Figure 3(a), CTRACE clearly outperforms CE in terms of relative regret and the difference is statistically significant at 95% confidence level. The dominance stems from both regularization and confidence-triggered update as shown in Figure 2. Figure 3(b) displays an empirical distribution of profit difference at period 3000, approximately 38 trading days from the start, in percentage relative to the profit earned by CE. Note that the distribution is positively skewed and CTRACE appears to make greater profit than CE more frequently. Specifically, the average profit difference between CTRACE and CE amounts to 4.2% of the profit earned by CE. In comparison, the average profit difference in percentage between the "clairvoyant" optimal policy and CE is by 8.9%.

Finally, we compare performance of CTRACE to that of AS policy in Figure 4. Figure 4(a) shows that CTRACE outperforms AS policy even more drastically than CE in terms of relative regret, and the superiority is statistically significant at 95% confidence level. In Figure 4(b), you can see an empirical distribution of profit difference at period 3000 in percentage relative to the profit earned by AS policy. It is clear that CTRACE is more profitable than AS policy in most of the sample paths. The average profit difference between CTRACE and AS policy amounts to 96.0% of the profit earned by AS policy. There are two primary factors contributing to the significant performance difference we observe between CTRACE and AS policy. The first is that excessively aggressive exploration done by AS policy is too costly. It is attributed to the fact that

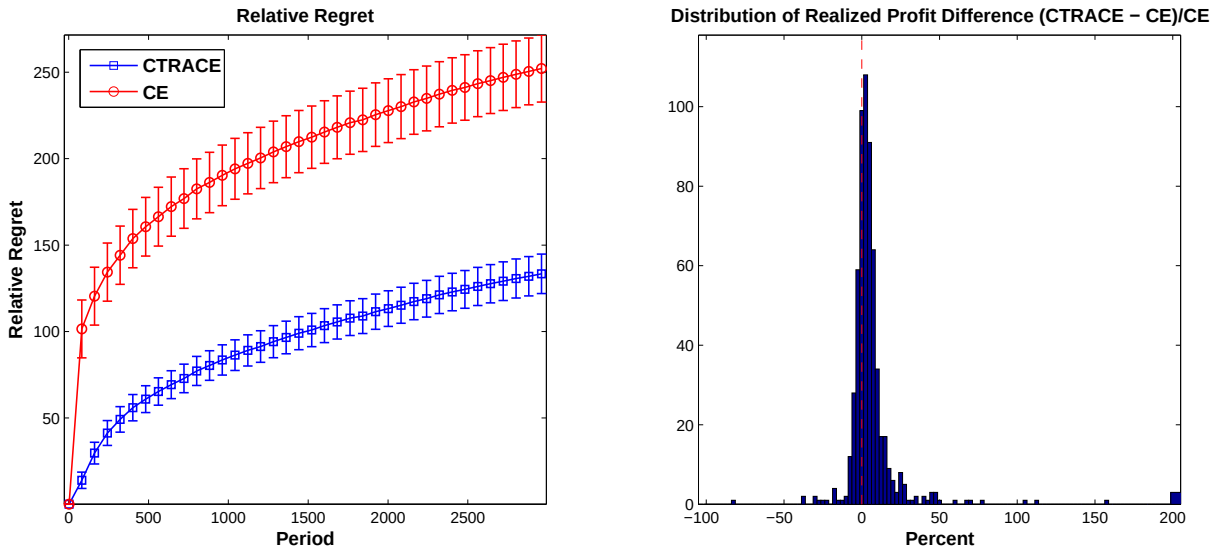

Figure 3: (a) Relative regret of CTRACE and CE. (b) Distribution of realized profit difference at period 3000 in percentage relative to the profit earned by CE.

AS policy is designed to explore actively in situations where pure exploitation performed by CE is not sufficient to identify a true model. In our problem, however, a great degree of exploration is naturally induced by observable return-predictive factors and thus aggressiveness of AS policy turns out to be even more than necessary. Meanwhile, CTRACE strikes a desired balance between exploration and exploitation by taking into account factor-driven natural exploration.

The second is that AS policy turns out to yield too few updates of price impact coefficients, which leads to much lower realized profit than that of CTRACE even if the tuning parameters are chosen via cross-validation. In Abbasi-Yadkori and Szepesvàri [2010], a critical step for establishing their regret bound is that AS policy updates an estimate at most $O(\log T)$ times. Note that the authors do not present any computational analysis to investigate practical implications of this issue. Meanwhile, CTRACE updates an estimate $\Theta(T)$ times. Therefore, we think that the observed underperformance of AS policy is not owing to the choice of the tuning parameters but to inherent structural difference regarding update frequency as well as the extent of exploration, at least in our problem.

4.3 Robustness against Mis-specified Decay Rates

So far, we assume that the trader knows exact decay rates r of transient price impact (Assumption 1-(c)). However, this may as well be too strong to hold in practice. Therefore, it is natural to ask

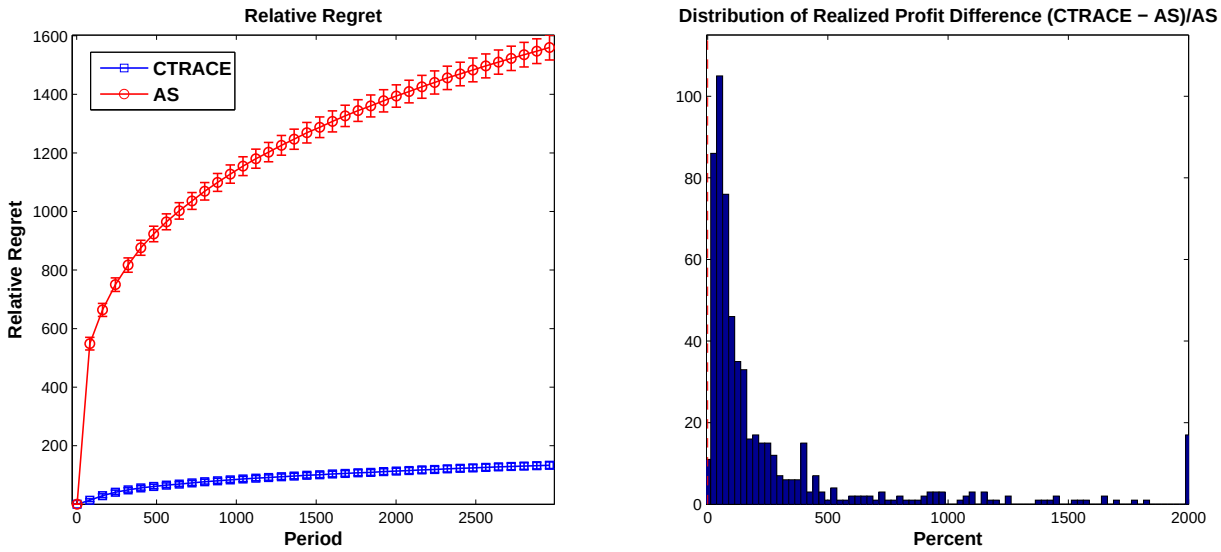

Figure 4: (a) Relative regret of CTRACE and AS policy. (b) Distribution of realized profit difference at period 3000 in percentage relative to the profit earned by AS policy.

how critical this assumption could be. We find that the performance of CTRACE is not so sensitive to using the exact decay rates. In order to illustrate this point, we perform additional computational analyses. In particular, we consider a setting where the trader does not know the exact half-lives of $[5, 7.5, 10, 15, 30, 45]$ minutes but instead uses one of the following two sets of seven half-lives: $[4, 9, 16, 25, 36, 49, 64]$ minutes (squares) and $[2, 4, 8, 16, 32, 64, 128]$ minutes (powers of 2). Note that we choose the two sets somewhat arbitrarily to assess sensitivity of CTRACE to mis-specification of decay rates.

As illustrated in Figure 5, CTRACE with mis-specified decay rates is competitive with CTRACE using exact decay rates as long as the confidence-triggered update level C_v is chosen appropriately, e.g., $C_v \in \{20, 40, 60\}$ for the first case and $C_v \in \{160, 320, 480\}$ for the second case. Indeed, our cross-validation with the mis-specified decay rates chose $C_v = 40$ for the first case and $C_v = 480$ for the second case. Interestingly, CTRACE with $C_v \in \{20, 40, 60\}$ in the first scenario appears to outperform slightly up to a first few thousand periods CTRACE with exact decay rates. Although the latter catches up and eventually attains lower regret than the former, the discrepancy would be irrelevant over a time scale of several months.

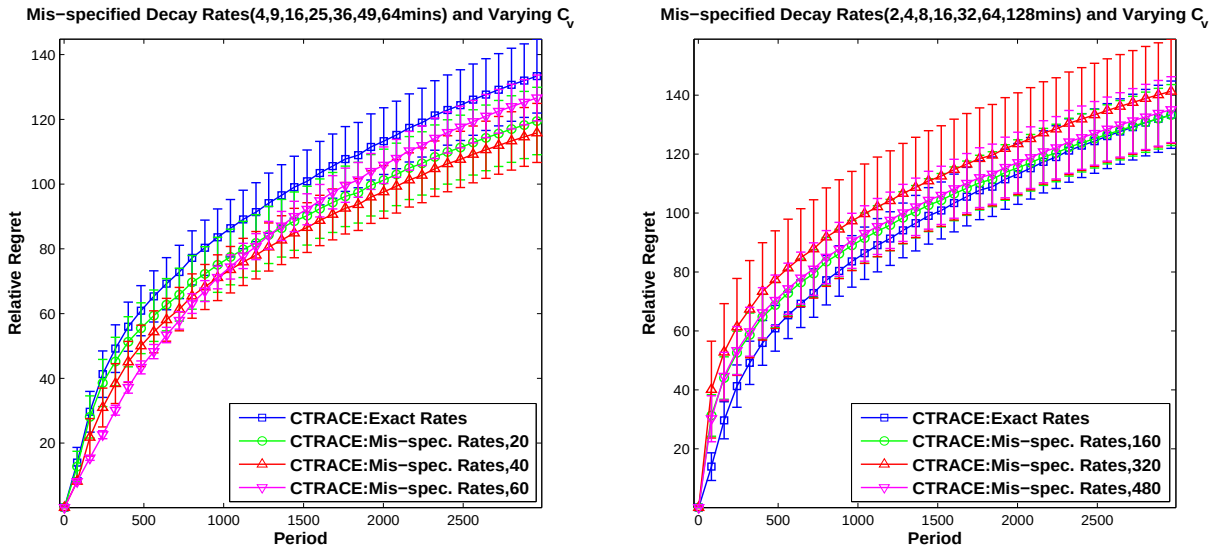

Figure 5: Relative regret with mis-specified half-lives and varying C_v: (a) Half-lives $[4, 9, 16, 25, 36, 49, 64]$ minutes (squares) and $C_v \in \{20, 40, 60\}$. (b) Half-lives $[2, 4, 8, 16, 32, 64, 128]$ minutes (powers of 2) and $C_v \in \{160, 320, 480\}$. $\kappa = 10^{11}$ for all cases.

5 Conclusion

We have considered a dynamic trading problem where a trader maximizes expected average risk-adjusted profit while trading a single risky security in the presence of unknown price impact. Our problem can be viewed as a special case of reinforcement learning: the trader can improve longer-term performance significantly by making decisions that explore efficiently to learn price impact at the expense of suboptimal short-term behavior such as execution of larger orders than appearing optimal with respect to current information. Like other reinforcement learning problems, it is crucial to strike a balance between exploration and exploitation. To this end, we have proposed the confidence-triggered regularized adaptive certainty equivalent policy (CTRACE) that improves purely exploitative certainty equivalent control (CE) in our problem. The enhancement is attributed to two properties of CTRACE: regularization and confidence-triggered update. Regularization encourages active exploration that accelerates learning as well as reduces the variance of an estimator. It helps keep CTRACE from being a passive learner due to overestimation of price impact that abates trading. Confidence-triggered update allows CTRACE to have monotonically nonincreasing upper bounds on estimation errors so that it reduces the frequency of overestimation. Using these two properties, we derived a finite-time expected regret bound for CTRACE of the form $O(\log^2 T)$. Finally, we have demonstrated through Monte Carlo simulation that CTRACE outperforms CE

and a reinforcement learning policy recently proposed in Abbasi-Yadkori and Szepesvàri [2010], and the performance of CTRACE is not so sensitive to using exact decay rates.

As extension to our current model, it would be interesting to develop an efficient reinforcement learning algorithm for a portfolio of securities. Another interesting direction is to incorporate a prior knowledge of particular structures of price impact coefficients, e.g. sparsity, to an estimation problem. It is worth considering other regularization schemes such as LASSO.

6 Uniform Bound on $\|z_t\|$

Using Assumption 2, we can obtain an upper bound on $\|z_t\|$ uniformly over $\theta \in \Theta$ and $t \geq 0$ as follows:

Lemma 6. *For any $0 < \xi < 1$, there exists $N \in \mathbb{N}$ being independent of θ such that $\|G^N(\theta)\| \leq \xi$ for all $\theta \in \Theta$. Thus, $\max_{0 \leq i \leq N-1} \sup_{\theta \in \Theta} \|G^i(\theta)\| \triangleq C_g$ is finite. For any fixed $\theta \in \Theta$, $\|z_t\| \leq C_g \|z_0\| + C_g C_\omega / (\xi(1-\xi^{1/N})) \triangleq C_z$, $\forall t \geq 0$ a.s. where $z_t = G(\theta) z_{t-1} + W_t$. Moreover, $\sup_{\theta \in \Theta} \|U(\theta)\| \leq C_g + 1$.*

Proof of Lemma 6 See e-companion.

Note that Lemma 6 can be applied only when θ is fixed over time. Throughout this paper, we assumed $\|z_0\| \leq 2C_g C_\omega / (\xi(1 - \xi^{1/N}))$ without loss of generality otherwise we can always set C_g to be greater than $\|z_0\| \xi (1 - \xi^{1/N}) / (2C_\omega)$.

7 Numerical Justification for Arithmetic Approximation of Price Dynamics

We present concrete numerical examples that support the validity of our price model as an approximation of the geometric model for practical purposes. As we discussed in Section 2.1, our numerical experiments conducted in Section 4 show that our infinite-horizon control problem could be approximated accurately by a finite-time control problem with a time horizon on a few week time scale. To be more precise, we define *relative error for* $P_0^{(T)}$ as $\|P_0^{(T)} - P\| / \|P\|$ where $P_t^{(T)}$ denotes a coefficient matrix of a quadratic value function at period t for a finite-horizon control problem with a terminal period T, and P denotes a coefficient matrix of a quadratic value function

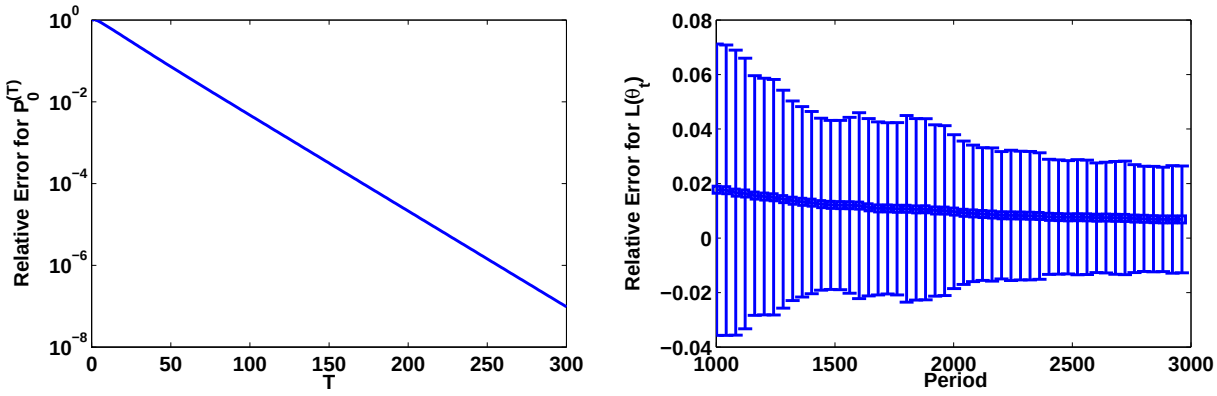

Figure 6: (a) Relative error for P_T: $T = 300$ corresponds to 3.8 trading days. (b) Relative error for $L(\theta_t)$ from CTRACE: Period 3000 corresponds to 38 trading days. The vertical bars represent two standard errors. In both figures, the simulation setting in Section 4 is used.

for our infinite-horizon control problem. As shown in Figure 6(a), the relative error for $P_0^{(T)}$ appears to decrease exponentially in T and the relative error for $P_0^{(300)}$ is almost 10^{-7} where $T = 300$ corresponds to 3.8 trading days.

Furthermore, we could learn unknown θ^* fast enough to take actions that are close to optimal actions on a required time scale. An action from a current estimate could be quite close to an optimal action even if estimation error for the current estimate is large, especially in cases where a few "principal components" of $L(\theta)$ with large directional derivatives with respect to θ are learned accurately. To be more precise, we define *relative error for* $L(\theta_t)$ as

$$\frac{\mathrm{E}[(L(\theta_t)z^*_{t-1} - L(\theta^*)z^*_{t-1})^2]}{\mathrm{E}[(L(\theta^*)z^*_{t-1})^2]} = \frac{(L(\theta_t) - L(\theta^*))\Pi_{zz}(\theta^*)(L(\theta_t) - L(\theta^*))^\top}{L(\theta^*)\Pi_{zz}(\theta^*)L(\theta^*)^\top}$$

where z^*_t is a stationary process generated by $u^*_t = L(\theta^*)z^*_{t-1}$ and $\Pi_{zz}(\theta^*) = E[z^*_t z^{*\top}_t]$. The relative error for $L(\theta_t)$ indicates how different an action from an estimate θ_t is than an optimal action from the true value θ^*. Figure 6(b) shows how the relative error for $L(\theta_t)$ evolves over time with two-standard-error bars when θ_t's are obtained from CTRACE. As you can see, all the 95%-confidence intervals lie within $\pm 3\%$ range after period 2500 that corresponds to 32 trading days. It implies that actions from estimates learned over a few weeks could be sufficiently close to optimal actions.

8 Proofs

Due to space constraints, proofs of selected key results are presented here. Other proofs are provided in an e-companion.

Proof of Theorem 2 Let $t_0 = 0$. Conditioned on the event $\{t_{i-1} < \infty\}$, for $t > t_{i-1}$,

$$\lambda_{\min}(V_t) \geq \lambda_{\min}(V_{t_{i-1}}) + \lambda_{\min}\left(\sum_{j=t_{i-1}+1}^{t} \psi_j \psi_j^\top\right) \geq \kappa + C_v t_{i-1} + \lambda_{\min}\left(\sum_{j=t_{i-1}+1}^{t} \psi_j \psi_j^\top\right).$$

By Lemma 1, $\frac{1}{t-t_{i-1}} \sum_{j=t_{i-1}+1}^{t} \psi_j \psi_j^\top \to \lambda_{\min}\left(U(\theta)\Pi_{zz}(\theta)U(\theta)^\top\right) \geq \underline{\lambda}^*_{\psi\psi} > C_v$ *a.s.* as $t \to \infty$ so long as θ is fixed after t_{i-1}. It implies that there exists $t_{i-1} + \tau \leq t_i < \infty$ a.s. such that $\lambda_{\min}\left(\sum_{j=t_{i-1}+1}^{t_i} \psi_j \psi_j^\top\right) \geq C_v(t_i - t_{i-1})$. Since $\lambda_{\min}(V_{t_i}) \geq \kappa + C_v t_{i-1} + C_v(t_i - t_{i-1}) = \kappa + C_v t_i$, t_i is indeed a qualified update time. That is, $\Pr(t_i < \infty | \, t_{i-1} < \infty) = 1$. If $\Pr(t_{i-1} < \infty) = 1$, then $\Pr(t_i < \infty, \, t_{i-1} < \infty) = \Pr(t_{i-1} < \infty)\Pr(t_i < \infty | t_{i-1} < \infty) = 1$ and thus $\Pr(t_i < \infty) = 1$. Since $\Pr(t_0 < \infty) = 1$, it follows that $\Pr(t_i < \infty) = 1$ for all $i \geq 0$ by induction. Hence, $\Pr(t_i < \infty, \ \forall i \geq 0) = \cap_{i=0}^{\infty} \Pr(t_i < \infty) = 1$. By Proposition 1, on the event $\{\theta^* \in \mathcal{S}_t(\delta), \ \forall t \geq 1\}$ for any $i \geq 1$

$$\begin{aligned}\|\theta_{t_i} - \theta^*\| \leq \|\theta_{t_i} - \hat{\theta}_{t_i}\| + \|\hat{\theta}_{t_i} - \theta^*\| &\leq \frac{2}{\sqrt{\lambda_{\min}(V_{t_i})}} \left(C_\epsilon \sqrt{2 \log\left(\frac{\det(V_{t_i})^{1/2} \det(\kappa I)^{-1/2}}{\delta}\right)} + \kappa^{1/2} \|\theta_{\max}\| \right) \\ &\leq \frac{2C_\epsilon \sqrt{(M+1)\log\left(C_\psi^2 \, t_i/\kappa + M + 1\right) + 2\log(1/\delta)} + 2\kappa^{1/2}\|\theta_{\max}\|}{\sqrt{C_v t_i}} = b_{t_i}.\end{aligned}$$

The second inequality follows from the fact that

$$\theta_{t_i} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{t_i} \left(\Delta p_i - g^\top f_{i-1} - \psi_i^\top \theta\right)^2 + \kappa \|\theta\|^2 = \underset{\theta \in \Theta}{\operatorname{argmin}} \ (\theta - \hat{\theta}_{t_i})^\top V_{t_i} (\theta - \hat{\theta}_{t_i})$$

and thus on the event $\{\theta^* \in \mathcal{S}_t(\delta), \ \forall t \geq 1\}$

$$\begin{aligned}\lambda_{\min}(V_{t_i}) \|\theta_{t_i} - \hat{\theta}_{t_i}\|^2 \leq (\theta_{t_i} - \hat{\theta}_{t_i})^\top V_{t_i} (\theta_{t_i} - \hat{\theta}_{t_i}) &\leq (\theta^* - \hat{\theta}_{t_i})^\top V_{t_i} (\theta^* - \hat{\theta}_{t_i}) \\ &\leq \left(C_\epsilon \sqrt{2 \log\left(\frac{\det(V_{t_i})^{1/2} \det(\kappa I)^{-1/2}}{\delta}\right)} + \kappa^{1/2} \|\theta_{\max}\| \right)^2.\end{aligned}$$

In the third inequality, we use $\lambda_{\min}(V_{t_i}) \geq \kappa + C_v t_i \geq C_v t_i$ by definition of t_i and

$$\det(V_{t_i}) \leq \lambda_{\max}(V_{t_i})^{M+1} \leq \mathrm{tr}(V_{t_i})^{M+1} = \left(\kappa(M+1) + \sum_{j=1}^{t_i} \|\psi_j\|^2\right)^{M+1} \leq \left(\kappa(M+1) + C_\psi^2\, t_i\right)^{M+1}.$$

For any $t_i < t < t_{i+1}$, $\|\theta_t - \theta^*\| = \|\theta_{t_i} - \theta^*\| \leq b_{t_i} = b_t$. Now, we show the monotonicity of b_t. A key observation is that for any $C_\epsilon > 0$, $C_\psi > 0$, $\kappa > 0$, $0 < \delta < 1$ and $\|\theta_{\max}\|$,

$$h(t) = \frac{C_\epsilon \sqrt{(M+1)\log\left(C_\psi^2\, t/\kappa + M + 1\right) + 2\log(1/\delta)} + \kappa^{1/2}\|\theta_{\max}\|}{\sqrt{C_v t}}$$

is strictly decreasing in $t \geq 1$ if $M \geq 2$. It can be easily verified through elementary calculus. Since b_t is monotonically nonincreasing and bounded below from 0, it converges almost surely. It follows from $\lim_{i\to\infty} b_{t_i} = 0$ *a.s.* that $\lim_{t\to\infty} b_t = 0$ *a.s.* ■

Proof of Theorem 3 Using $\log(t + M + 1) \leq \sqrt{t} + \sqrt{M+1}$ for all $t \geq 0$,

$$\begin{aligned}
&\frac{2C_\epsilon \sqrt{(M+1)\log\left(C_\psi^2\, t/\kappa + M + 1\right) + 2\log(1/\delta)} + 2\kappa^{1/2}\|\theta_{\max}\|}{\sqrt{C_v t}} \\
&\leq \frac{2C_\epsilon \sqrt{(M+1)\left(\sqrt{C_\psi^2 t/\kappa} + \sqrt{M+1}\right) + 2\log(1/\delta)}}{\sqrt{C_v t}} + \frac{2\kappa^{1/2}\|\theta_{\max}\|}{\sqrt{C_v t}} \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon
\end{aligned}$$

$$\begin{aligned}
\text{if } \; t &\geq \left(\frac{8C_\epsilon^2 C_\psi (M+1) + 4\sqrt{4C_\epsilon^4 C_\psi^2 (M+1)^2 + \kappa C_\epsilon^2 C_v \epsilon^2\left((M+1)^{3/2} + 2\log(1/\delta)\right)}}{\sqrt{\kappa} C_v \epsilon^2}\right)^2 \vee \frac{(4\kappa\|\theta_{\max}\|)^2}{C_v \epsilon^2} \\
&= T_2(\epsilon, \delta, C_v)
\end{aligned}$$

Let

$$T_1^*(\delta') = 4\left(\frac{32(C_z^* C_g)^2 (M+K+1)}{\xi^2 (1-\xi^{\frac{2}{N}})\underline{\lambda}_{zz}^*}\right)^2 \log\left(\frac{(M+K+2)^4}{432\delta'^2}\right) \vee 8\left(\frac{32(C_z^* C_g)^2 (M+K+1)}{\xi^2 (1-\xi^{\frac{2}{N}})\underline{\lambda}_{zz}^*}\right)^3 \vee 216.$$

Suppose for contradiction that $t_{N(\epsilon,\delta,C_v)} > T_1^*(\delta') \vee \tau + T_2(\epsilon, \delta, C_v) \triangleq \tilde{T}^*$. Let t_i be the last update time less than $T_2(\delta, \epsilon, C_v)$. t_i is zero if there is no update time before $T_2(\epsilon, \delta, C_v)$. Then, there is

no update time in the interval $[t_i+1, \tilde{T}^*]$ by definition of $t_{N(\epsilon,\delta,C_v)}$ and $T_2(\epsilon,\delta,C_v)$. Thus,

$$\begin{aligned}\lambda_{\min}\left(V_{\tilde{T}^*}\right) &\geq \lambda_{\min}(V_{t_i}) + \lambda_{\min}\left(\sum_{t=t_i+1}^{\tilde{T}^*}\psi_t\psi_t^\top\right) \geq \kappa + C_v t_i + \lambda_{\min}\left(\sum_{t=t_i+1}^{\tilde{T}^*}\psi_t\psi_t^\top\right) \\ &\geq \kappa + C_v t_i + \frac{7}{8}\underline{\lambda}^*_{\psi\psi}(\tilde{T}^* - t_i) \geq \kappa + C_v\tilde{T}^*\end{aligned}$$

where the second inequality holds by definition of t_i, the third inequality holds by Lemma 2, and the last inequality holds because $\frac{7}{8}\underline{\lambda}^*_{\psi\psi} > C_v$. Also, $\tilde{T}^* - t_i \geq \tau$. Consequently, $\tilde{T}^*$ is eligible for a next update time after t_i. It implies that $t_{N(\epsilon,\delta,C_v)} = \tilde{T}^*$ but this is a contradiction.

Proof of Theorem 4 Let

$$T_1(\delta/2) = 4\left(\frac{32(C_z^*C_g)^2(M+K+1)}{\xi^2(1-\xi^{\frac{2}{N}})\underline{\lambda}^*_{zz}}\right)^2\log\left(\frac{(M+K+2)^4}{432(\delta/2)^2}\right) \vee 8\left(\frac{32(C_z^*C_g)^2(M+K+1)}{\xi^2(1-\xi^{\frac{2}{N}})\underline{\lambda}^*_{zz}}\right)^3 \vee 216 \vee \tau,$$

$$T_2(\delta/2) = \left(\frac{8C_\epsilon^2C_\psi(M+1) + 4\sqrt{4C_\epsilon^4C_\psi^2(M+1)^2 + \kappa C_\epsilon^2C_v\epsilon^2\left((M+1)^{3/2} + 2\log(1/\delta)\right)}}{\sqrt{\kappa}C_v\epsilon^2}\right)^2 \vee \frac{(4\kappa\|\theta_{\max}\|)^2}{C_v\epsilon^2},$$

$$\begin{aligned}T_3(\delta/2) = 4&\left(\frac{32(C_z^*C_g)^2(M+K+1)}{\xi^2(1-\xi^{\frac{2}{N}})\lambda_{\min}(\Pi_{zz}(\theta^*))}\right)^2\log\left(\frac{(M+K+2)^4}{432(\delta/2)^2}\right) \vee 8\left(\frac{32(C_z^*C_g)^2(M+K+1)}{\xi^2(1-\xi^{\frac{2}{N}})\lambda_{\min}(\Pi_{zz}(\theta^*))}\right)^3 \vee 216 \\ &\vee \frac{3C_z^*(2C_\omega + C_z^*)}{C_\omega^2},\end{aligned}$$

$$\epsilon = \frac{1}{\sqrt{M+1}C_L}\left(\frac{\nu^3(1-\nu^{\frac{1}{N}})^3\lambda_{\min}(\Pi_{zz}(\theta^*))}{42NC_g^{N+1}C_\omega^2} \wedge \frac{\nu^3(1-\nu^{\frac{1}{N}})^3\lambda_{\min}(U(\theta^*)\Pi_{zz}(\theta^*)U(\theta^*)^\top)}{42NC_g^{N+1}C_\omega^2(1+\|U(\theta^*)\|)^2} \wedge \frac{\nu-\xi}{NC_g^{N-1}}\right).$$

Then,

$$\begin{aligned}&\sum_{t=1}^{T}(u_t - L^*z_{t-1})^\top(R+B^\top P^*B)(u_t - L^*z_{t-1}) \\ &= (R+B^\top P^*B)\sum_{t=1}^{T}(u_t - L^*z_{t-1})^2 = (R+B^\top P^*B)\sum_{t=1}^{T}((L(\theta_{t-1}) - L(\theta^*))z_{t-1})^2 \\ &\leq (R+B^\top P^*B)\sum_{t=1}^{T}\|L(\theta_{t-1}) - L(\theta^*)\|^2\|z_{t-1}\|^2 \leq (R+B^\top P^*B)C_z^{*2}C_L^2\sum_{t=1}^{T}\|\theta_{t-1} - \theta^*\|^2 \\ &= (R+B^\top P^*B)C_z^{*2}C_L^2\Bigg(\sum_{t=1}^{T_1(\delta/2)+T_2(\delta/2)}\|\theta_{t-1} - \theta^*\|^2\mathbf{1}\left\{1 \leq t \leq T_1(\delta/2)+T_2(\delta/2)\right\} \\ &\quad + \sum_{t=T_1(\delta/2)+T_2(\delta/2)+1}^{T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)}\|\theta_{t-1} - \theta^*\|^2\mathbf{1}\left\{T_1(\delta/2)+T_2(\delta/2) < t \leq T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)\right\}\end{aligned}$$

$$+\sum_{t=T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)+1}^{T}\|\theta_{t-1}-\theta^*\|^2\mathbf{1}\left\{T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)<t\le T\right\}\Bigg)$$

$$\le (R+B^\top P^*B)C_z^{*2}C_L^2\Bigg(\sum_{t=1}^{T_1(\delta/2)+T_2(\delta/2)}\|\theta_{t-1}-\theta^*\|^2+\sum_{t=T_1(\delta/2)+T_2(\delta/2)+1}^{T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)}\|\theta_{t-1}-\theta^*\|^2$$

$$+\sum_{t=T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)+1}^{T}\|\theta_{t-1}-\theta^*\|^2\mathbf{1}\left\{T>T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)\right\}\Bigg)$$

First, on the event $\{\theta^*\in\mathcal{S}_t(\delta/2),\ \forall t\ge 1\}\cap\mathcal{B}(\delta/2)$ with $\Pr(\{\theta^*\in\mathcal{S}_t(\delta/2),\ \forall t\ge 1\}\cap\mathcal{B}(\delta/2))\ge 1-\delta$,

$$\sum_{t=1}^{T_1(\delta/2)+T_2(\delta/2)}\|\theta_{t-1}-\theta^*\|^2\le(T_1(\delta/2)+T_2(\delta/2))\|\theta_{\max}\|^2,\ \sum_{t=T_1(\delta/2)+T_2(\delta/2)+1}^{T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)}\|\theta_{t-1}-\theta^*\|^2\le T_3(\delta/2)\epsilon^2.$$

Also, for $t\ge T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)+1$,

$$\lambda_{\min}\left(V_{t-1}\right)\ge\lambda_{\min}\left(V_{t_{N(\epsilon,\delta/2,C_v)}}\right)+\lambda_{\min}\left(\sum_{i=t_{N(\epsilon,\delta/2,C_v)}+1}^{t-1}\psi_i\psi_i^\top\right)\ge\kappa+C_v t_{N(\epsilon,\delta/2,C_v)}+\tilde{C}(t-1-t_{N(\epsilon,\delta/2,C_v)})$$

$$=\kappa+\tilde{C}(t-1)-(\tilde{C}-C_v)t_{N(\epsilon,\delta/2,C_v)}\ge\kappa+\tilde{C}(t-1)-(\tilde{C}-C_v)_+(T_1(\delta/2)+T_2(\delta/2))$$

where the second inequality holds by definition of $t_{N(\epsilon,\delta/2,C_v)}$ and $t-1-t_{N(\epsilon,\delta/2,C_v)}\ge T_3(\delta/2))$, and the last inequality holds by Lemma 3.

$$\therefore\sum_{t=T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)+1}^{T}\|\theta_{t-1}-\theta^*\|^2$$

$$\le\sum_{t=T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)+1}^{T}\frac{\tau}{\lambda_{\min}(V_{t-1})}\left(2C_\epsilon\sqrt{2\log\left(\frac{\det(V_{t-1})^{1/2}\det(\kappa I)^{-1/2}}{\delta/2}\right)}+2\kappa^{1/2}\|\theta_{\max}\|\right)^2$$

$$\le\sum_{t=T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)+1}^{T}\frac{\tau\left(2C_\epsilon\sqrt{(M+1)\log\left(C_\psi^2\,(t-1)/\kappa+M+1\right)+2\log(2/\delta)}+2\kappa^{1/2}\|\theta_{\max}\|\right)^2}{\kappa+\tilde{C}(t-1)-(\tilde{C}-C_v)_+(T_1(\delta/2)+T_2(\delta/2))}$$

$$\le\sum_{t=T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)+1}^{T}\frac{\tau\left(2C_\epsilon\sqrt{(M+1)\log\left(C_\psi^2\,(T-1)/\kappa+M+1\right)+2\log(2/\delta)}+2\kappa^{1/2}\|\theta_{\max}\|\right)^2}{\kappa+\tilde{C}(t-1)-(\tilde{C}-C_v)_+(T_1(\delta/2)+T_2(\delta/2))}$$

$$\le\frac{\tau\left(2C_\epsilon\sqrt{(M+1)\log\left(C_\psi^2\,T/\kappa+M+1\right)+2\log(2/\delta)}+2\kappa^{1/2}\|\theta_{\max}\|\right)^2}{\tilde{C}}$$

$$\times\log\left(\frac{\kappa+\tilde{C}(T-1)-(\tilde{C}-C_v)_+(T_1(\delta/2)+T_2(\delta/2))}{\kappa+\tilde{C}(T_1(\delta/2)+T_2(\delta/2)+T_3(\delta/2)-1)-(\tilde{C}-C_v)_+(T_1(\delta/2)+T_2(\delta/2))}\right).$$

Let $\mathcal{A} = \{\theta^* \in \mathcal{S}_t(\delta/2),\ \ \forall t \geq 1\} \cap \mathcal{B}(\delta/2)$ and $q = \Pr(\mathcal{A})$. Then,

$$
\begin{aligned}
\bar{R}_T^{\pi}(z_0) &= \mathsf{E}[z_T^{*\top} P^* z_T^* - z_T^\top P^* z_T] + \mathsf{E}\left[\sum_{t=1}^{T} (u_t - L^* z_{t-1})^\top (R + B^\top P^* B)(u_t - L^* z_{t-1})\right] \\
&\leq 2\|P^*\| C_z^{*2} + q\mathsf{E}\left[\sum_{t=1}^{T} (u_t - L^* z_{t-1})^\top (R + B^\top P^* B)(u_t - L^* z_{t-1}) \middle| \mathcal{A}\right] \\
&\quad + (1-q)\mathsf{E}\left[\sum_{t=1}^{T} (u_t - L^* z_{t-1})^\top (R + B^\top P^* B)(u_t - L^* z_{t-1}) \middle| \mathcal{A}^c\right] \\
&\leq 2\|P^*\| C_z^{*2} + \mathsf{E}\left[\sum_{t=1}^{T} (u_t - L^* z_{t-1})^\top (R + B^\top P^* B)(u_t - L^* z_{t-1}) \middle| \mathcal{A}\right] + \delta T (R + B^\top P^* B) C_z^{*2} C_L^2 \|\theta_{\max}\|^2 \\
&\leq 2\|P^*\| C_z^{*2} + (R + B^\top P^* B) C_z^{*2} C_L^2 \Bigg(\left(T_1(\delta/2) + T_2(\delta/2) + \delta T\right) \|\theta_{\max}\|^2 + T_3(\delta/2)\epsilon^2 \\
&\quad + \frac{\tau\left(2C_\epsilon \sqrt{(M+1)\log\left(C_\psi^2\, T/\kappa + M + 1\right) + 2\log(2/\delta)} + 2\kappa^{1/2}\|\theta_{\max}\|\right)^2}{\tilde{C}} \\
&\quad \times \log\left(\frac{\kappa + \tilde{C}(T-1) - (\tilde{C} - C_v)_+ (T_1(\delta/2) + T_2(\delta/2))}{\kappa + \tilde{C}(T_1(\delta/2) + T_2(\delta/2) + T_3(\delta/2) - 1) - (\tilde{C} - C_v)_+ (T_1(\delta/2) + T_2(\delta/2))}\right) \\
&\quad \times \mathbf{1}\left\{T > T_1(\delta/2) + T_2(\delta/2) + T_3(\delta/2)\right\}\Bigg) \quad \text{where} \quad \tilde{C} \triangleq \frac{3}{8}\lambda_{\min}(U(\theta^*)\Pi_{zz}(\theta^*)U(\theta^*)^\top).
\end{aligned}
$$

A key observation is that the last inequality holds for any $0 < \delta < 1$ and δ is not an input to the CTRACE algorithm, i.e. operation of CTRACE is independent of δ. If we choose $\delta = 1/T$ for $T \geq 2$, then we obtain the desired regret bound. Note that $T_i(\delta/2)$'s are all $O(\log T)$. Therefore, it is not difficult to see that the expected regret bound is $O(\log^2 T)$.

References

Y. Abbasi-Yadkori and C. Szepesvàri. Regret bounds for the adaptive control of linear quadratic systems. In *24th Annual Conference on Learning Theory*. JMLR: Workshop and Conference Proceedings, 2010.

Y. Abbasi-Yadkori, D. Pal, and C. Szepesvàri. Online least squares estimation with self-normalized processes: An application to bandit problems. Working paper, 2011.

A. Alfonsi, A. Schied, and A. Schulz. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10:143–157, 2010.

R. Almgren and N. Chriss. Optimal control of portfolio transactions. *Journal of Risk*, 3:5–39, 2000.

D. Bertsimas and A. W. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1: 1–50, 1998.

J. Bouchaud, Yuval Gefen, Marc Potters, and Matthieu Wyart. Fluctuations and response in financial markets: the subtle nature of 'random' price changes. *Quantitative Finance*, 4:176–190, 2004.

H. Chen and L. Guo. Convergence rate of least squares identification and adaptive control for stochastic systems. *International Journal of Control*, 44:1459–1476, 1986.

R. Cont, A. Kukanov, and S. Stoikov. The price impact of order book events. Working paper, 2012.

A. Dufour and R. F. Engle. Time and the price impact of a trade. *The Journal of Finance*, 55(6): 2467–2498, 2000.

N. Garleanu and L. H. Pedersen. Dynamic trading with predictable returns and transaction costs. forthcoming in The Journal of Finance, 2012.

J. Gatheral. No-dynamic-arbitrage and market impact. *Quantitative Finance*, 10:749–759, 2010.

J. Hasbrouck. Measuring the information content of stock trades. *The Journal of Finance*, 46(1): 179–207, 1991.

G. Huberman and W. Stanzl. Price manipulation and quasi-arbitrage. *Econometrica*, 74(4):1247–1276, 2004.

R. Kissell and M. Glantz. *Optimal Trading Strategies: Quantitative Approaches for Managing Market Impact and Trading Risk*. Amacom Books, 2003.

A. S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.

T. L. Lai and C. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10:154–166, 1982.

T. L. Lai and C. Wei. Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Transactions on Automatic Control*, 31:898–906, 1986.

C. C. Moallemi, B. Park, and Van Roy B. Strategic execution in the presence of an uninformed arbitrageur. *Journal of Financial Markets*, 15:361–391, 2012.

A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16:1–32, 2013.

A. A. Obizhaeva. Liquidity estimates and selection bias. Working paper, 2012.

I. Rosu. A dynamic model of the limit order book. *Review of Financial Studies*, 22:4601–4641, 2009.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, 1998.