# Policy Gradient Methods for the Noisy Linear Quadratic Regulator over a Finite Horizon

Ben Hambly * Renyuan Xu * Huining Yang†*

November 20, 2020

**Abstract**

We explore reinforcement learning methods for finding the optimal policy in the linear quadratic regulator (LQR) problem. In particular we consider the convergence of policy gradient methods in the setting of known and unknown parameters. We are able to produce a global linear convergence guarantee for this approach in the setting of finite time horizon and stochastic state dynamics under weak assumptions. The convergence of a projected policy gradient method is also established in order to handle problems with constraints. We illustrate the performance of the algorithm with two examples. The first example is the optimal liquidation of a holding in an asset. We show results for the case where we assume a model for the underlying dynamics and where we apply the method to the data directly. The empirical evidence suggests that the policy gradient method can learn the global optimal solution for a larger class of stochastic systems containing the LQR framework and that it is more robust with respect to model mis-specification when compared to a model-based approach. The second example is an LQR system in a higher dimensional setting with synthetic data.

## 1 Introduction

The Linear Quadratic Regulator (LQR) problem is one of the most fundamental in optimal control theory. Its aim is to find a control for a linear dynamical system, that is the dynamics of the state of the system is described by a linear function of the current state and input, subject to a quadratic cost. It is an important problem for a number of reasons: (1) the LQR problem is one of the few optimal control problems for which there exists a closed-form analytical representation of the optimal feedback control; (2) when the dynamics are nonlinear and hard to analyze, a LQR approximation may be obtained as a local expansion and provide an approximation that is provably close to the original problem; (3) the LQR has been used in a wide variety of applications. In particular, in the set-up of fixed time horizon and stochastic dynamics, applications include portfolio optimization [3] and optimal liquidation [8] in finance, resource allocation in energy markets [39, 43], and biological movement systems [34].

Until recently much of the work on the LQR problem has focused on solving for the optimal controls under the assumption that the model parameters are *fully known*. See the book of Anderson and Moore [10] for an introduction to the LQR problem with known parameters. However, assuming that the controller has access to all the model parameters is not realistic for many applications, and this has lead to the exploration of learning approaches to the problem. We consider reinforcement learning (RL), one of the three basic machine learning paradigms (alongside supervised learning and unsupervised learning). Unlike the situation with full information on the model parameters, RL is learning to make

*Mathematical Institute, University of Oxford. **Email:** {hambly, xur, yang}@maths.ox.ac.uk

†Supported by the EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) in collaboration with BP plc.

decisions via trial and error, through interactions with the (partially) unknown environment. In RL, an agent takes an action and receives a reinforcement signal in terms of a numerical reward, which encodes the outcome of her action. In order to maximize the accumulated reward over time, the agent learns to select her actions based on her past experiences (exploitation) and/or by making new choices (exploration). There are two popular approaches in RL to handle the LQR with unknown parameters: the model-based approach and the model-free approach.

In the paradigm of the model-based approach, the controller estimates the unknown model parameters and then constructs a control policy based on the estimated parameters. The classical approach is the *certainty equivalence principle* [11]: the unknown parameters are estimated using observations (or samples), and a control policy is then designed by treating the estimated parameters as the truth. In the first step, the unknown model parameters can be estimated by standard statistical methods such as least-square minimization [21]. The second step is to show that when the estimated parameters are accurate enough, the policy using the "plug-in" estimates enjoys good theoretical guarantees of being close to optimal. See [21] and [25] for the optimal gap and sample complexities along this line and see [23] for the sample complexity with distributed robust learning. Another line of work in the model-based regime focuses on *uncertainty* quantification. The controller updates their posterior belief or the confidence bounds on the unknown model parameters and then makes decisions in an online manner, see [1, 2, 22, 31, 38].

Another recently developed approach is the *model-free approach*, where the controller learns the optimal policy *directly* via interacting with the system, without inferring the model parameters. As the optimal policy in the LQR problem is a linear function of the state, the aim is to determine this linear function. This is equivalent to learning a set of parameters in matrix form, called the policy matrix. One natural way to achieve this goal is to apply the gradient descent method in the parameter space of the policy matrix, also referred to as the *policy gradient method*. In particular, the policy gradient method computes the gradient of the cost function with respect to the policy matrix and then updates the policy in the steepest decent direction to find the optimal policy. The paper [24] was the first to show that policy gradients converge to the global optimal solution with polynomial (in the relevant quantities) sample complexity. However, [24] focuses on the case where the only noise in the system is in the initial state, and the rest of the state transitions are deterministic. There are other methods that fall into the category of the model-free approach, including the Actor-Critic method [44] and least-squares temporal difference learning [42].

Compared to the model-based approach, which strongly relies on the assumption that the stochastic system lies within the LQR framework and may, in practice, suffer from model mis-specification, the execution of the model-free algorithm does not rely on the assumptions of the model. It has been shown that the policy gradient method can learn the global optimal solution, not only for the LQR framework, but also for a more general class of deterministic systems in the setting of an infinite time horizon [14]. Thus the advantage of the model-free approach is that it is more robust against model mis-specification compared to the model-based approach.

**Our Contributions.** We now summarize our contributions. Motivated by many real-word decision-making problems with a fixed deadline and uncertainty in the underlying dynamics, such as the optimal liquidation problem that we discuss in Section 2, we extend the framework of [24] by incorporating a finite time horizon and sub-Gaussian noise (which includes Gaussian noise as a special case). In particular, we provide a global linear convergence guarantee and a polynomial sample complexity guarantee for the policy gradient method in this setting with both known parameters (Theorem 3.3) and unknown parameters (Theorem 4.4). The analysis with known parameters paves the way for learning LQR with unknown parameters. In addition, numerically solving the Riccati equation with known parameters in high dimensions may suffer from computational inaccuracy. The policy gradient method provides a direct way of searching for the optimal solution with known parameters in this case, which may be of separate interest. Note that the optimal policy is time-invariant for the LQR with infinite time horizon, whereas the optimal policy is time-dependent with finite time horizon and hence harder to learn in general. With noise in the dynamics, we need more careful choices of the hyper-parameters to retrieve compatible sample complexities with noisy observations. In addition, when optimal polices need to satisfy certain constraints, we provide a global convergence result for the projected policy gradient method in Theorem 4.5. This is required in the context of our application to the optimal liquidation problem.

We will formulate the optimal liquidation problem over a fixed horizon as a noisy LQR problem which is essentially the classical Almgren-Chriss formulation [8]. The performance of the algorithm on NASDAQ ITCH data is assessed. As well as using the method within this modelling approach, we also consider the performance of the policy gradient method when applied directly to the data with an appropriate cost function. This improves the performance of the LQR/Almgren-Chriss solution and shows promising results for the use of the policy gradient method for problems that are 'close' to the LQR framework.

## 1.1 Related Work

**Policy Gradient Methods for LQR Problems.** Since the policy gradient method is the main focus of our paper, here we provide a review of the previous theoretical work on this method in various LQR settings and extensions. The first global convergence result for the policy gradient method to learn the optimal policy for LQR problems was developed in [24] in the setting of infinite horizon and deterministic dynamics. The work of [24] was extended in [14] to give global optimality guarantees of policy gradient methods for a larger class of control problem that includes the linear-quadratic case. In particular, this class of control problem satisfies a closure condition under policy improvement and convexity of policy improvement steps. The paper [15] considers policy gradient methods for LQR problems in terms of optimizing a real valued matrix function over the set of feedback gains. The extension of the policy gradient method to continuous-time can be found in [16]. All of these methods are in the infinite horizon setting and without the addition of noise in the dynamics.

There has been some work on the case of noisy dynamics, but all in the setting of infinite horizon. In [27] the problem with a multiplicative noise was discussed, using a relatively straightforward extension of the deterministic dynamics considered in the original framework. In the case of additive noise [32] studies the global convergence of policy gradient and other learning algorithms for the LQR over an infinite time-horizon and with Gaussian noise. In particular, the policy considered in [32] is a randomized policy with Gaussian distribution. There is also [35] which studies derivative-free (zeroth-order) policy optimization methods for the LQR with bounded additive noise. Finally some other contributions can be found in [17, 45] for zero-sum LQR games and [18, 29] for mean-field LQR games.

There are several major technical differences compared to [24]. Due to the time-dependent nature of the admissible policies over a finite-horizon and randomness from the system noise, we need more careful handling of the system dynamics during the training to ensure they are well defined. We also need conditions to guarantee the gradient dominant condition, and for the proper sample size to get a good estimate on the gradient of the cost function with high probability. See the more detailed discussion in Remark 4.11.

**Optimal Liquidation.** An early mathematical framework for the optimal liquidation problem is due to Almgren and Chriss [8]. In this problem a trader is required to liquidate a portfolio of shares over a fixed horizon. The selling of a large number of shares at once has both temporary and permanent impacts on the share price causing it to decrease. The trader therefore wishes to find a trading strategy which maximizes their return from, or alternatively, minimizes the cost of, the liquidation of the portfolio subject to a given level of risk.

This problem has been considered in many papers and extended in many directions. See for instance [5], [7] and [26]. We will cast this as an LQR problem and show how the policy gradient method is a powerful tool for solving this problem even without assumptions on the model.

More recently techniques from reinforcement learning have been applied to the optimal liquidation problem. The first paper to do this was [36] where the authors showed promising results for this approach by designing a Q-learning based algorithm to optimally select price levels and passively place limit orders. This was further developed in [30] which designed a Q-learning based algorithm for liquidation within the standard Almgren-Chriss framework. For recent work incorporating deep learning see for example [12], [33], [37], and [46]. See [19] for a detailed review on reinforcement learning with applications in finance and economics, and the references therein. However, all these works focus on the model-free setting without taking advantage of even weak modelling assumptions on the market dynamics. In addition, the performances of these proposed algorithms are validated only through empirical studies and no theoretical guarantee of convergence is provided.

**Organization and Notation.** For any matrix $Z = (Z_1, \cdots, Z_d) \in \mathbb{R}^{m\times d}$ with $Z_j \in \mathbb{R}^m$ $(j = 1, 2, \cdots, d)$, $Z^\top \in \mathbb{R}^{m\times d}$ denote the transpose of $Z$, $\|Z\|$ denotes the spectral norm of a matrix $Z$; $\mathrm{Tr}(Z)$ denotes the trace of a square matrix $Z$; $\sigma_{\min}(Z)$ denotes the minimal singular value of a square matrix $Z$; and $\mathrm{vec}(Z) = (Z_1^\top, \cdots, Z_d^\top)^\top$ denote the vectorized version of a matrix $Z$. For a sequence of matrices $\boldsymbol{D} = (D_0, \cdots, D_T)$, we define a new norm $|||\boldsymbol{D}|||$ as:

$$|||\boldsymbol{D}||| = \sum_{t=0}^{T} \|D_t\|,$$

where $D_t \in \mathbb{R}^{m\times d}$. Further denote $\mathcal{N}(\mu, \Sigma)$ as the Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$.

The rest of the paper is organized as follows. We introduce the mathematical framework and problem set-up in Section 2. The first step in our convergence analysis of the policy gradient method is to consider the case of known model parameters in Section 3. When parameters are unknown, the convergence results for the sample-based policy gradient method and projected policy gradient method are obtained in Section 4. Finally, the algorithm is applied to liquidation problem. See Sections 2.1 and 5 for the corresponding set-up and algorithm performance, respectively.

# 2 Problem Set-up

We consider the following LQR problem over a finite time horizon $T$,

$$\min_{\{u_t\}_{t=0}^{T-1}} \mathbb{E}\left[\sum_{t=0}^{T-1} \left(x_t^\top Q_t x_t + u_t^\top R_t u_t\right) + x_T^\top Q_T x_T\right], \tag{2.1}$$

such that for $t = 0, 1, \cdots, T-1$,

$$x_{t+1} = Ax_t + Bu_t + w_t,\ x_0 \sim \mathcal{D}. \tag{2.2}$$

Here $x_t \in \mathbb{R}^d$ is the state of the system with the initial state $x_0$ drawn from a distribution $\mathcal{D}$, $u_t \in \mathbb{R}^k$ is the control at time $t$ and $\{w_t\}_{t=0}^{T-1}$ are zero-mean IID noises which are independent from $x_0$. At this moment, we only assume $x_0$ and $\{w_t\}_{t=0}^{T-1}$ have finite second moments. That is, $\mathbb{E}[x_0 x_0^\top]$ and $W := \mathbb{E}[w_t w_t^\top]$ $(\forall\, t = 0, 1, \cdots, T-1)$ exist. The system parameters $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times k}$ are referred to as system (transition) matrices; $Q_t \in \mathbb{R}^{d\times d}$ $(\forall\, t = 0, 1, \cdots, T)$ and $R_t \in \mathbb{R}^{k\times k}$ $(\forall t = 0, 1, \cdots, T-1)$ are matrices that parameterize the quadratic costs. Note that the expectation in (2.1) is taken with respect to both $x_0 \sim \mathcal{D}$ and $w_t$ ($t = 0, 1, \cdots, T-1$). We further denote by $\boldsymbol{u} := (u_0, \cdots, u_{T-1})$, $\boldsymbol{x} := (x_0, \cdots, x_T)$, $\boldsymbol{w} := (w_0, \cdots, w_{T-1})$, $\boldsymbol{Q} := (Q_0, \cdots, Q_T)$, and $\boldsymbol{R} := (R_0, \cdots, R_{T-1})$, the profile over the decision period $T$.

To solve the LQR problem (2.1)-(2.2), let us start with some conditions on the model parameters to assure the well-definedness of the problem.

**Assumption 2.1** (Cost Parameter)**.** *Assume $Q_t \in \mathbb{R}^{d\times d}$, for $t = 0, 1, \cdots, T$, and $R_t \in \mathbb{R}^{k\times k}$, for $t = 0, 1, \cdots, T-1$, are positive definite matrices.*

Under Assumption 2.1, we can properly define a sequence of matrices $\{P_t^*\}_{t=0}^T$ as the solution to the following dynamic Riccati equation [13]:

$$P_t^* = Q_t + A^\top P_{t+1}^* A - A^\top P_{t+1}^* B \left(B^\top P_{t+1}^* B + R_t\right)^{-1} B^\top P_{t+1}^* A, \tag{2.3}$$

with terminal condition

$$P_T^* = Q_T.$$

The matrices $\{P_t^*\}_{t=0}^T$ can be found by solving the Riccati equations iteratively backwards in time. In particular with a slight modification of the initial state distribution in [13, Chapter 4.1], we have the following result.

**Lemma 2.2** (Well-definedness and the Optimal Solution [13])**.** *Under Assumption 2.1,*

1. *The solution $P_t^*$ to the Riccati equation* (2.3) *is positive definite, $\forall\, t = 0, 1, \cdots, T$;*
2. *Then the optimal control sequence $\{u_t\}_{t=0}^{T-1}$ is given by*

$$u_t = -K_t^* x_t, \tag{2.4}$$

*where*

$$K_t^* = \left(B^\top P_{t+1}^* B + R_t\right)^{-1} B^\top P_{t+1}^* A. \tag{2.5}$$

To find the optimal solution in the linear feedback form (2.4), we only need to focus on the following class of linear *admissible policies* in feedback form

$$u_t = -K_t x_t, \qquad t = 0, 1, \cdots, T-1, \tag{2.6}$$

which can be fully characterized by $\boldsymbol{K} := (K_0, \cdots, K_{T-1})$.

## 2.1 Application: The Optimal Liquidation Problem

One application of the LQR framework (2.1)-(2.2) is the optimal liquidation problem. We give a slight variant of the setup of Almgren-Chriss [8]. Our aim is to liquidate an amount $q_0$ of an asset, with price $S_0$ at time 0, over the time period $[0, T]$ with trading decisions made at discrete time points $t = 0, 1, \ldots, T-1$. At each time $t$ our decision is to liquidate an amount $u_t$ of the asset. Any residual holding is then liquidated at time $T$. This will have two types of price impact. There will be a temporary price impact, caused when the order 'walks the book' and a permanent price impact as traders rearrange their positions in the light of the sell order. We will assume the impacts are linear in the number of traded shares.

We write $S_t$ for the asset price at time $t$. This evolves according to a Bachelier model with a linear permanent price impact in that

$$S_{t+1} = S_t + \sigma Z_{t+1} - \gamma u_t,$$

where, for each $t = 1, \ldots, T$, $Z_t$ is an independent standard normal random variable, $\sigma$ is the volatility and $\gamma$ is the permanent price impact parameter. The inventory process $q_t$ records the current holding in the asset at time $t$. Thus we have

$$q_{t+1} = q_t - u_t.$$

Therefore, the two-dimensional state process is

$$\begin{pmatrix} S_{t+1} \\ q_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S_t \\ q_t \end{pmatrix} + \begin{pmatrix} -\gamma \\ -1 \end{pmatrix} u_t + \begin{pmatrix} \sigma Z_{t+1} \\ 0 \end{pmatrix}. \tag{2.7}$$

When selling shares we incur a temporary price impact, parameter $\beta$, in that if, at time $t$, we trade $u_t$ of our asset then we obtain $\tilde{S}_t = S_t - \beta u_t$ per share. Therefore the total revenue is $\sum_{t=0}^{T-1} u_t \tilde{S}_t + q_T \tilde{S}_T$, and $C_T$, the total cost of execution over $[0, T]$, is the book value at time 0 minus the revenue:

$$C_T = q_0 S_0 - \sum_{t=0}^{T-1} u_t \tilde{S}_t - q_T \tilde{S}_T.$$

In a similar way to [8], after summation by parts, we have

$$C_T = -\sigma \sum_{t=1}^{T} q_t Z_t - \frac{\gamma}{2} \sum_{t=0}^{T-1} u_t^2 + \frac{\gamma}{2}\left(q_0^2 - q_T^2\right) + \beta \sum_{t=0}^{T-1} u_t^2 + \beta q_T^2.$$

The mean and variance of the total cost of execution are given by

$$\mathbb{E}(C) = \sum_{t=0}^{T-1} \delta u_t^2 + \delta q_T^2 + \frac{\gamma}{2} q_0^2, \quad \mathrm{var}(C) = \sum_{t=1}^{T} \sigma^2 q_t^2.$$

where $\delta = \beta - \gamma/2$ summarizes the impact and is assumed positive.

Following Almgren-Chriss [8], we minimize the following cost function

$$C_{\mathrm{AC}} = \min\left(\mathbb{E}(C) + \phi \, \mathrm{var}(C)\right), \tag{2.8}$$

where $\phi$ is a parameter balancing risk versus return. For our LQR framework we take the cost function to be

$$\begin{aligned} C_{\mathrm{LQR}}(\epsilon) &= \min\left(\mathbb{E}(C) + \phi \, \mathrm{var}(C) + \epsilon \sum_{t=0}^{T} S_t^2\right) \\ &= \min\left(\sum_{t=0}^{T-1} \delta u_t^2 + \delta q_T^2 + \frac{\gamma}{2} q_0^2 + \phi \sum_{t=1}^{T} \sigma^2 q_t^2 + \epsilon \sum_{t=0}^{T} S_t^2\right). \end{aligned} \tag{2.9}$$

Note that the term $\epsilon \sum_{t=0}^{T} S_t^2$, with some small $\epsilon > 0$, serves as a regularization term to guarantee Assumption 2.1 holds. In practice, we can show that the optimal solution with $\epsilon$ small is close to the Almgren-Chriss solution (when $\epsilon = 0$). In addition, the algorithm will still converge with $\epsilon = 0$. See more discussion in Section 5. Thus, in the LQR formulation we have $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $B = (-\gamma, -1)^\top$, and $w_t = (\sigma Z_{t+1}, 0)^\top$ and the objective function has $Q_T = \begin{pmatrix} \epsilon & 0 \\ 0 & \delta + \phi\sigma^2 \end{pmatrix}$, $Q_t = \begin{pmatrix} \epsilon & 0 \\ 0 & \phi\sigma^2 \end{pmatrix}$ and $R_t = \delta$. It is easy to see that $Q_t$, for $t = 0, 1, \cdots, T$ and $R_t$ for $t = 0, 1, \cdots, T-1$ are positive definite, hence Assumption 2.1 is satisfied.

We will show that the problem is well-defined and can be solved using the methods of this paper with rigorous convergence guarantees.

## 3 Exact Gradient Methods with Known Parameters

In this section we assume all the parameters in the model, $\{Q_t\}_{t=0}^{T}$, $\{R_t\}_{t=0}^{T-1}$, $A$, $B$, are known. The analysis of exact gradient methods with known parameters paves the way for learning LQR with unknown parameters in Section 4. In addition, numerically solving the Riccati equation (2.3) with known parameters in high dimensions may suffer from computational inaccuracy [6, 41]. The exact gradient provides a direct way of searching for the optimal solution in this case, which may be of separate interest. Since an admissible policy can be fully characterized by $\boldsymbol{K}$, the cost of a policy $\boldsymbol{K}$ can be correspondingly defined as

$$C(\boldsymbol{K}) = \mathbb{E}\left[\sum_{t=0}^{T-1}\left(x_t^\top Q_t x_t + u_t^\top R_t u_t\right) + x_T^\top Q_T x_T\right], \tag{3.1}$$

where $\{x_t\}_{t=1}^{T}$ and $\{u_t\}_{t=0}^{T-1}$ are the dynamics and controls induced by following $\boldsymbol{K}$, starting with $x_0 \sim \mathcal{D}$. Recall that $\boldsymbol{K}^*$ is the optimal policy for the problem, in that

$$\boldsymbol{K}^* = \arg\min_{\boldsymbol{K}} C(\boldsymbol{K}), \tag{3.2}$$

subject to the dynamics (2.2).

**Well-definedness of the State Process.** To prove the global convergence of policy gradient methods, the essential idea is to show the *gradient dominance condition*, which states that $C(\boldsymbol{K}) - C(\boldsymbol{K}^*)$ can be bounded by $\|\nabla C(\boldsymbol{K})\|_F$ for any admissible policy $\boldsymbol{K}$. One of the key steps to guarantee this gradient dominance condition is the well-definedness of the state covariance matrix. That is, $\mathbb{E}[x_t x_t^\top]$ is positive definite for $t = 0, 1, \cdots, T$. This condition holds almost for free for LQR problems with infinite time horizon and deterministic dynamics. The only condition needed there is the positive definiteness of $\mathbb{E}[x_0 x_0^\top]$ (See [24]). However, some effort needs to be made to ensure that the state covariance matrix is well-defined for LQR problems with finite horizon and stochastic dynamics. We show that this condition holds under moderate conditions.

**Assumption 3.1** (Initial State and Noise Process)**.** *We assume that*

1. *Initial state: $x_0 \sim \mathcal{D}$ such that $\mathbb{E}[x_0 x_0^\top]$ is positive definite;*
2. *Noise: $\{w_t\}_{t=0}^{T-1}$ are IID and independent from $x_0$ such that $\mathbb{E}[w_t] = 0$, and $W = \mathbb{E}[w_t w_t^\top]$ is positive definite, $\forall t = 0, 1, \cdots, T-1$.*

Define $\underline{\sigma}_{\boldsymbol{X}}$ as the lower bound over all the minimum singular values of $\mathbb{E}[x_t x_t^\top]$:

$$\underline{\sigma}_{\boldsymbol{X}} = \min_t \sigma_{\min}(\mathbb{E}[x_t x_t^\top]), \tag{3.3}$$

then we have the following result and the proof can be found in Appendix B.1.

**Lemma 3.2** (Well-definedness of the State Covariance Matrix)**.** *Under Assumption 3.1, we have $\mathbb{E}[x_t x_t^\top]$ is positive definite for $t = 0, 1, \cdots, T$ under any control policy $\boldsymbol{K}$. Therefore, $\underline{\sigma}_{\boldsymbol{X}} > 0$.*

Lemma 3.2 implies that if the initial state and the noise driving the dynamics are non-degenerate, the covariance matrices of the state dynamics are positive definite for any policy $\boldsymbol{K}$. However, the covariance matrix may be degenerate in many applications, especially when inventory processes are involved. (See, for example, the liquidation problem (2.7).) In this case, some problem-dependent conditions are needed to guarantee that $\underline{\sigma}_{\boldsymbol{X}} > 0$ holds. See more discussion on the condition $\underline{\sigma}_{\boldsymbol{X}} > 0$

for the liquidation problem in Section 5.1. In the light of this we will assume $\underline{\sigma}_{\boldsymbol{X}} > 0$ in the analysis of the convergence of the algorithm in Sections 3 and 4.

Similarly, we define $\underline{\sigma}_{\boldsymbol{R}}$ and $\underline{\sigma}_{\boldsymbol{Q}}$ to be the smallest values of all the minimum singular values of $\boldsymbol{R}$ and $\boldsymbol{Q}$:

$$\underline{\sigma}_{\boldsymbol{R}} = \min_t \sigma_{\min}(R_t), \tag{3.4}$$

and

$$\underline{\sigma}_{\boldsymbol{Q}} = \min_t \sigma_{\min}(Q_t). \tag{3.5}$$

Under Assumption 2.1, we have $\underline{\sigma}_{\boldsymbol{R}} > 0$ and $\underline{\sigma}_{\boldsymbol{Q}} > 0$.

We write $\mathcal{H} = \{h \,|\, h \text{ are polynomials in the model parameters}\}$ and $\mathcal{H}(.)$ when there are other dependencies. The model parameters are in terms of $d$, $k$, $\frac{1}{\|A\|}$, $\frac{1}{\|A\|+1}$, $\|A\|$, $\frac{1}{\|B\|}$, $\frac{1}{\|B\|+1}$, $\|B\|$, $\frac{1}{|||\boldsymbol{R}|||}$, $\frac{1}{|||\boldsymbol{R}|||+1}$, $|||\boldsymbol{R}|||$, $\frac{1}{\|W\|}$, $\frac{1}{\|W\|+1}$, $\|W\|$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $|||\boldsymbol{Q}|||$, $\mathbb{E}[x_0 x_0^\top]$, and $\frac{1}{\mathbb{E}[x_0 x_0^\top]}$.

**Exact Gradient Descent.** We consider the following *exact* gradient descent updating rule to find the optimal solution (3.2),

$$K_t^{n+1} = K_t^n - \eta \nabla_t C(\boldsymbol{K}^n), \ \forall\, 0 \le t \le T-1, \tag{3.6}$$

where $n$ is the number of iterations, $\nabla_t C(\boldsymbol{K}) = \frac{\partial C(\boldsymbol{K})}{\partial K_t}$ is the gradient of $C(\boldsymbol{K})$ with respect to $K_t$, and $\eta$ is the step size. We further denote $\nabla C(\boldsymbol{K}) = (\nabla_0 C(\boldsymbol{K}), \cdots, \nabla_{T-1} C(\boldsymbol{K}))$.

Let us define the state covariance matrix

$$\Sigma_t = \mathbb{E}\left[x_t x_t^\top\right], \ t = 0, 1, \cdots, T, \tag{3.7}$$

where $\{x_t\}_{t=1}^T$ is a state trajectory generated by $\boldsymbol{K}$. Further define a matrix $\Sigma_{\boldsymbol{K}}$ as the sum of $\Sigma_t$,

$$\Sigma_{\boldsymbol{K}} = \sum_{t=0}^T \Sigma_t = \mathbb{E}\Big[\sum_{t=0}^T x_t x_t^\top\Big]. \tag{3.8}$$

Then, the main result for this setting is the following.

**Theorem 3.3** (Global Convergence of Gradient Methods)**.** *Assume Assumption 2.1 holds. Further assume $\underline{\sigma}_{\boldsymbol{X}} > 0$ and $C(\boldsymbol{K}^0)$ is finite. Then, for an appropriate (constant) setting of the stepsize $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$, and for $\epsilon > 0$, if we have*

$$N \ge \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{2\eta\, \underline{\sigma}_{\boldsymbol{X}}^2\, \underline{\sigma}_{\boldsymbol{R}}} \log \frac{C(\boldsymbol{K}^0) - C(\boldsymbol{K}^*)}{\epsilon},$$

*the exact gradient descent method* (3.6) *enjoys the following performance bound:*

$$C(\boldsymbol{K}^N) - C(\boldsymbol{K}^*) < \epsilon.$$

The proof of Theorem 3.3 relies on the regularity of the LQR problem, some properties of the gradient descent dynamics, and the perturbation analysis of the covariance matrix of the controlled dynamics.

### 3.1 Regularity of the LQR Problem and Properties of the Gradient Descent Dynamics.

Let us start with the analysis of some properties of the LQR problem (2.1)-(2.2). To start, Proposition 3.4 focuses on the well-definedness of the Ricatti system $\{P_t^{\boldsymbol{K}}\}_{t=0}^T$ induced by a control $\boldsymbol{K}$; Lemma 3.5 gives a representation of the gradient term; Lemma 3.6 and Lemma 3.7 provide the gradient dominance condition and a smoothness condition on the cost function $C(\boldsymbol{K})$ with respect to policy $\boldsymbol{K}$, respectively; and finally, Lemma 3.8 gives two useful upper bounds on Ricatti system and state covariance matrices.

In the finite time horizon setting, define $P_t^{\boldsymbol{K}}$ as the solution to

$$P_t^{\boldsymbol{K}} = Q_t + K_t^\top R_t K_t + (A - BK_t)^\top P_{t+1}^{\boldsymbol{K}} (A - BK_t), \quad t = 0, 1, \cdots, T-1, \tag{3.9}$$

with terminal condition

$$P_T^{\boldsymbol{K}} = Q_T.$$

Note that (3.9) is equivalent to the Riccati equation (2.3) with optimal $K_t = K_t^*$ as given by (2.5). We have the following result on the well-definedness of $P_t^{\boldsymbol{K}}$ and the proof can be found in Appendix B.1.

**Proposition 3.4.** *Under Assumption 2.1, the matrices $P_t^{\boldsymbol{K}}$ for $t = 0, 1, \ldots, T$ derived from (3.9) are positive definite.*

To ease the exposition, we write $P_t^{\boldsymbol{K}}$ as $P_t$ when there is no confusion. Then the cost of $\boldsymbol{K}$ can be rewritten as

$$C(\boldsymbol{K}) = \mathbb{E}_{x_0 \sim \mathcal{D}}\Big[x_0^\top P_0 x_0 + L_0\Big],$$

where, for $t = 0, 1, \cdots, T-1$,

$$L_t = L_{t+1} + \mathbb{E}[w_t^\top P_{t+1} w_t] = L_{t+1} + \mathrm{Tr}(W P_{t+1}), \tag{3.10}$$

with $L_T = 0$. To see this,

$$\begin{aligned}\mathbb{E}[x_0^\top P_0 x_0] + L_0 &= \mathbb{E}\left[x_0^\top Q_0 x_0 + x_0^\top K_0^\top R_0 K_0 x_0 + x_0^\top (A - BK_0)^\top P_1 (A - BK_0) x_0 + \sum_{t=0}^{T-1} w_t^\top P_{t+1} w_t\right] \\ &= \mathbb{E}\left[x_0^\top Q_0 x_0 + u_0^\top R_0 u_0 + x_1^\top P_1 x_1 + \sum_{t=1}^{T-1} w_t^\top P_{t+1} w_t\right] = \mathbb{E}\Big[\sum_{t=0}^{T-1}\Big(x_t^\top Q_t x_t + u_t^\top R_t u_t\Big) + x_T^\top Q_T x_T\Big].\end{aligned}$$

In addition, define

$$E_t = (R_t + B^\top P_{t+1} B) K_t - B^\top P_{t+1} A, \; t = 0, 1, \cdots, T-1. \tag{3.11}$$

Then we have the following representation of the gradient term.

**Lemma 3.5.** *The policy gradient has the following representation, for $t = 0, 1, \cdots, T-1$,*

$$\begin{aligned}\nabla_t C(\boldsymbol{K}) &= 2\Big(\Big(R_t + B^\top P_{t+1} B\Big) K_t - B^\top P_{t+1} A\Big)\mathbb{E}\left[x_t x_t^\top\right] \\ &= 2 E_t \Sigma_t.\end{aligned}$$

*Proof.* Since

$$C(\boldsymbol{K}) = \mathbb{E}\Big[x_0^\top P_0 x_0 + L_0\Big] = \mathbb{E}\Big[x_0^\top (Q_0 + K_0^\top R_0 K_0) x_0 + x_0^\top (A - BK_0)^\top P_1 (A - BK_0) x_0 + \sum_{t=0}^{T-1} w_t^\top P_{t+1} w_t\Big],$$

we have

$$\nabla_0 C(\boldsymbol{K}) = \frac{\partial C(\boldsymbol{K})}{\partial K_0} = \mathbb{E}\Big[2R_0K_0x_0x_0^\top - 2B^\top P_1(A - BK_0)x_0x_0^\top\Big] = 2E_0\mathbb{E}\Big[x_0x_0^\top\Big] = 2E_0\Sigma_0.$$

Similarly, $\forall\, t = 0, 1, \cdots, T-1$,

$$\nabla_t C(\boldsymbol{K}) = 2\left(\left(R_t + B^\top P_{t+1}B\right)K_t - B^\top P_{t+1}A\right)\mathbb{E}[x_tx_t^\top] = 2E_t\mathbb{E}\Big[x_tx_t^\top\Big] = 2E_t\Sigma_t,$$

where the expectation $\mathbb{E}$ is taken with respect to both initial distribution $x_0 \sim \mathcal{D}$ and noises $\boldsymbol{w}$. □

In classical optimization theory [24], gradient domination and smoothness of the objective function are two key conditions to guarantee the global convergence of the gradient descent methods. To prove that $C(\boldsymbol{K})$ is gradient dominated, we first prove Lemma 3.6, which indicates that for a policy $\boldsymbol{K}$, the distance between $C(\boldsymbol{K})$ and the optimal cost $C(\boldsymbol{K}^*)$ is bounded by the sum of the magnitude of the gradient $\nabla_t C(\boldsymbol{K})$ for $t = 0, 1, \cdots, T-1$.

**Lemma 3.6.** *Assume Assumption 2.1 holds and $\underline{\sigma}_{\boldsymbol{X}} > 0$. Let $\boldsymbol{K}^*$ be an optimal policy and $C(\boldsymbol{K})$ be finite, then*

$$\underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t + B^\top P_{t+1}B\|} \operatorname{Tr}(E_t^\top E_t) \le C(\boldsymbol{K}) - C(\boldsymbol{K}^*) \le \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\,\underline{\sigma}_{\boldsymbol{X}}^2\,\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \operatorname{Tr}(\nabla_t C(\boldsymbol{K})^\top \nabla_t C(\boldsymbol{K})),$$

*where $\underline{\sigma}_{\boldsymbol{X}}$ and $\underline{\sigma}_{\boldsymbol{Q}}$ are defined in* (3.3) *and* (3.4).

We defer the proof of Lemma 3.6 to Appendix B.1. Lemma 3.6 implies that when the gradient becomes small, the value of the objective function is close to $C(\boldsymbol{K}^*)$. Now we consider the smoothness condition of the objective function. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be smooth if

$$|f(x) - f(y) - \nabla f(y)^\top (x - y)| \le \frac{M}{2}\|x - y\|^2, \ \forall\, x, y \in \mathbb{R}^n,$$

for some finite constant $M$. In general, it is difficult to characterize the smoothness of $C(\boldsymbol{K})$, since it may blow up when $A - BK_t$ is large. Here we will prove that $C(\boldsymbol{K})$ is "almost" smooth, in the sense that when $\boldsymbol{K}'$ is sufficiently close to $\boldsymbol{K}$, $C(\boldsymbol{K}') - C(\boldsymbol{K})$ is bounded by the sum of the first and second order terms in $\boldsymbol{K} - \boldsymbol{K}'$.

**Lemma 3.7** ("Almost Smoothness")**.** *Let $\{x_t'\}$ be the sequence of states for a single trajectory generated by $\boldsymbol{K}'$ starting from $x_0' = x_0$. Then, $C(\boldsymbol{K})$ satisfies*

$$C(\boldsymbol{K}') - C(\boldsymbol{K}) = \sum_{t=0}^{T-1} \Big[2\operatorname{Tr}\left(\Sigma_t'(K_t' - K_t)^\top E_t\right) + \operatorname{Tr}\left(\Sigma_t'(K_t' - K_t)^\top (R_t + B^\top P_{t+1}B)(K_t' - K_t)\right)\Big], \quad (3.12)$$

*where $\Sigma_t' = \mathbb{E}\left[x_t'(x_t')^\top\right]$.*

We defer the proof of Lemma 3.7 to Appendix B.1. To see why Lemma 3.7 is related to the smoothness, observe that when $\boldsymbol{K}'$ is sufficiently close to $\boldsymbol{K}$, in the sense that

$$\Sigma_t' \approx \Sigma_t + O(\|K_t - K_t'\|), \ \forall t = 0, 1, \cdots, T-1,$$

the first term in (3.12) will behave as $\operatorname{Tr}((K_t - K_t')\nabla_t C(\boldsymbol{K}))$ by Lemma 3.5, and the second term in (3.12) will be of second order in $K_t - K_t'$.

To utilize Lemmas 3.6 and 3.7 in the proof of Theorem 3.3, we need to further bound $P_t$ and $\Sigma_{\boldsymbol{K}}$, which is provided below in Lemma 3.8. The proof can be found in Appendix B.1.

**Lemma 3.8.** *Assume Assumption 2.1 holds, and $\underline{\sigma}_{\boldsymbol{X}} > 0$. Then we have*

$$\|P_t\| \le \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{X}}}, \ \|\Sigma_{\boldsymbol{K}}\| \le \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}},$$

*where $\underline{\sigma}_{\boldsymbol{X}}$ and $\underline{\sigma}_{\boldsymbol{Q}}$ are defined in* (3.3) *and* (3.5).

### 3.2 Perturbation Analysis of $\Sigma_K$.

First, let us define two linear operators on symmetric matrices. For $X \in \mathbb{R}^{d\times d}$ we set

$$\mathcal{F}_{K_t}(X) = (A - BK_t)X(A - BK_t)^\top, \quad \text{and} \quad \mathcal{T}_{\boldsymbol{K}}(X) := X + \sum_{t=0}^{T-1} \Pi_{i=0}^{t}(A - BK_i)\, X\, \Pi_{i=0}^{t}(A - BK_{t-i})^\top.$$

If we write $\mathcal{G}_t = \mathcal{F}_{K_t} \circ \mathcal{F}_{K_{t-1}} \circ \cdots \circ \mathcal{F}_{K_0}$, then

$$\mathcal{G}_t(X) = \mathcal{F}_{K_t} \circ \mathcal{G}_{t-1}(X) = \Pi_{i=0}^{t}(A - BK_i)\, X\, \Pi_{i=0}^{t}(A - BK_{t-i})^\top, \tag{3.13}$$

and

$$\mathcal{T}_{\boldsymbol{K}}(X) = X + \sum_{t=0}^{T-1} \mathcal{G}_t(X). \tag{3.14}$$

We first show the relationship between the operator $\mathcal{T}_{\boldsymbol{K}}$ and the quantity $\Sigma_{\boldsymbol{K}}$. The proof can be found in Appendix B.1.

**Proposition 3.9.** *For $T \geq 2$, we have that*

$$\Sigma_{\boldsymbol{K}} = \mathcal{T}_{\boldsymbol{K}}(\Sigma_0) + \Delta(\boldsymbol{K}, W), \tag{3.15}$$

*where*

$$\Delta(\boldsymbol{K}, W) = \sum_{t=1}^{T-1} \sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top + T\, W,$$

*with $D_{t,s} = \Pi_{u=s}^{t}(A - BK_u)$ (for $s = 1, 2, \cdots, t$), and $\Sigma_0 = \mathbb{E}\left[x_0 x_0^\top\right]$.*

Let

$$\rho := \max\Big\{ \max_{0\leq t\leq T-1} \|A - BK_t\|, \max_{0\leq t\leq T-1} \|A - BK_t'\|, 1 + \xi \Big\}, \tag{3.16}$$

for some small constant $\xi > 0$. Then we have the following result on perturbations of $\Sigma_{\boldsymbol{K}}$.

**Lemma 3.10** (Perturbation Analysis of $\Sigma_K$)**.** *Assume Assumption 2.1 holds. Then*

$$\begin{aligned}\Big\|\Sigma_{\boldsymbol{K}} - \Sigma_{\boldsymbol{K}'}\Big\| &\leq \Big\|(\mathcal{T}_{\boldsymbol{K}} - \mathcal{T}_{\boldsymbol{K}'})(\Sigma_0)\Big\| + \|\Delta(\boldsymbol{K}, W) - \Delta(\boldsymbol{K}', W)\| \\ &\leq \frac{\rho^{2T} - 1}{\rho^2 - 1}\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right)\Big(2\rho\,\|B\|\,|||\boldsymbol{K} - \boldsymbol{K}'||| + \|B\|^2\,|||\boldsymbol{K} - \boldsymbol{K}'|||^2\Big).\end{aligned}$$

**Remark 3.11.** By the definition of $\rho$ in (3.16), we have $\rho \geq 1 + \xi > 1$. This regularization term $1 + \xi$ is defined for ease of exposition. Alternatively, if we define $\rho := \max\Big\{\max_{0\leq t\leq T-1}\|A - BK_t\|, \max_{0\leq t\leq T-1}\|A - BK_t'\|\Big\}$, a similar analysis can still be carried out by considering the different cases: $\rho < 1$, $\rho = 1$ and $\rho > 1$.

The proof of Lemma 3.10 is based on the following Lemmas 3.12 and 3.13, which establish the Lipschitz property for the operators $\mathcal{F}_{K_t}$ and $\mathcal{G}_t$, respectively.

**Lemma 3.12.** *It holds that, $\forall\, t = 0, 1, \cdots, T-1$,*

$$\|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\| \leq 2\|A - BK_t\|\|B\|\|K_t - K_t'\| + \|B\|^2\|K_t - K_t'\|^2. \tag{3.17}$$

We refer to [24, Lemma 19] for the proof of Lemma 3.12.

Recall the definition of $\mathcal{G}_t$ in (3.13) associated with $\boldsymbol{K}$, similarly let us define $\mathcal{G}_t' = \mathcal{F}_{K_t'} \circ \mathcal{F}_{K_{t-1}'} \circ \cdots \circ \mathcal{F}_{K_0'}$ for policy $\boldsymbol{K}'$. Then we have the following perturbation analysis for $\mathcal{G}_t$.

**Lemma 3.13** (Perturbation Analysis for $\mathcal{G}_t$)**.** *For any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$, we have that*

$$\sum_{t=0}^{T-1} \left\|(\mathcal{G}_t - \mathcal{G}_t')(\Sigma)\right\| \leq \frac{\rho^{2T}-1}{\rho^2-1}\Big(\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\|\Big)\|\Sigma\|. \tag{3.18}$$

We defer the proof of Lemma 3.13 to Appendix B.2. The following perturbation analysis on $\mathcal{T}$ follows immediately from Lemma 3.13.

**Corollary 3.14.** *For any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$, we have*

$$\left\|(\mathcal{T}_{\boldsymbol{K}} - \mathcal{T}_{\boldsymbol{K}'})(\Sigma)\right\| \leq \frac{\rho^{2T}-1}{\rho^2-1}\Big(\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\|\Big)\|\Sigma\|, \tag{3.19}$$

*where $\rho$ is defined in* (3.16)*.*

Now we are ready for the proof of Lemma 3.10.

*Proof of Lemma 3.10.* Using Lemma 3.12,

$$\begin{aligned}\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\| &= \sum_{t=0}^{T-1}\Big(2\|A-BK_t\|\|B\|\|K_t-K_t'\| + \|B\|^2\|K_t-K_t'\|^2\Big)\\ &\leq 2\rho\|B\|\sum_{t=0}^{T-1}\|K_t-K_t'\| + \|B\|^2\sum_{t=0}^{T-1}\|K_t-K_t'\|^2.\end{aligned}$$

In the same way as for the proof of Lemma 3.13, we have, $\forall\, t=1,\cdots,T-1$,

$$\sum_{s=1}^{t}\left\|D_{t,s}WD_{t,s}^\top - D_{t,s}'W(D_{t,s}')^\top\right\| \leq \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{s=0}^{t}\|\mathcal{F}_{K_s} - \mathcal{F}_{K_s'}\|\right)\|W\|. \tag{3.20}$$

By Proposition 3.9, Corollary 3.14, (3.14) and (3.20), we have

$$\begin{aligned}\left\|\Sigma_{\boldsymbol{K}} - \Sigma_{\boldsymbol{K}'}\right\| &\leq \left\|(\mathcal{T}_{\boldsymbol{K}} - \mathcal{T}_{\boldsymbol{K}'})(\Sigma_0)\right\| + \sum_{t=1}^{T-1}\sum_{s=1}^{t}\left\|D_{t,s}WD_{t,s}^\top - D_{t,s}'W(D_{t,s}')^\top\right\|\\ &\leq \frac{\rho^{2T}-1}{\rho^2-1}\Big(\sum_{t=0}^{T-1}\|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\|\Big)\left(\|\Sigma_0\| + T\|W\|\right)\\ &\leq \frac{\rho^{2T}-1}{\rho^2-1}\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right)\Big(2\rho\,\|B\|\,\|\!|\boldsymbol{K}-\boldsymbol{K}'|\!\| + \|B\|^2\,\|\!|\boldsymbol{K}-\boldsymbol{K}'|\!\|^2\Big).\end{aligned} \tag{3.21}$$

The last inequality holds since $\|\Sigma_0\| \leq \|\Sigma_{\boldsymbol{K}}\| \leq \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$ by Lemma 3.8. ∎

### 3.3 Convergence and Complexity Analysis.

We now provide the proof of Theorem 3.3 after two preliminary Lemmas.

**Lemma 3.15.** *Assume Assumption 2.1 holds,* $\underline{\sigma}_{\boldsymbol{X}} > 0$*, and that*

$$K_t' = K_t - \eta \nabla_t C(\boldsymbol{K}), \tag{3.22}$$

*where*

$$\eta \leq \min \left\{ \frac{(\rho^2 - 1)\, \underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{2T(\rho^{2T} - 1)(2\rho + 1)(C(\boldsymbol{K}) + \underline{\sigma}_{\boldsymbol{Q}}\, T\|W\|)\|B\| \max_t\{\|\nabla_t C(\boldsymbol{K})\|\}}, \frac{1}{2C_1} \right\}, \tag{3.23}$$

*with*

$$C_1 = \left( \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\| \right) \left( \frac{(2\rho + 1)\|B\|(\rho^{2T} - 1)}{(\rho^2 - 1)\, \underline{\sigma}_{\boldsymbol{X}}} \sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\| \right) + \frac{2C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} \sum_{t=0}^{T-1} \|R_t + B^\top P_{t+1} B\|. \tag{3.24}$$

*Then we have*

$$C(\boldsymbol{K}') - C(\boldsymbol{K}^*) \leq \Big(1 - 2\eta\, \underline{\sigma}_{\boldsymbol{R}} \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big).$$

We defer the proof of Lemma 3.15 to Appendix B.3.

**Lemma 3.16.** *Assume Assumption 2.1 holds and* $\underline{\sigma}_{\boldsymbol{X}} > 0$*. Then we have that*

$$\sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\|^2 \leq 4\Big(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\Big)^2 \frac{\max_t \|R_t + B^\top P_{t+1} B\|}{\underline{\sigma}_{\boldsymbol{X}}} (C(\boldsymbol{K}) - C(\boldsymbol{K}^*)),$$

*and that:*

$$\sum_{t=0}^{T-1} \|K_t\| \leq \frac{1}{\underline{\sigma}_{\boldsymbol{R}}} \Big( \sqrt{T \cdot \frac{\max_t \|R_t + B^\top P_{t+1} B\|}{\underline{\sigma}_{\boldsymbol{X}}} (C(\boldsymbol{K}) - C(\boldsymbol{K}^*))} + \sum_{t=0}^{T-1} \|B^\top P_{t+1} A\| \Big).$$

*Proof.* Using Lemma 3.8 we have

$$\sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\|^2 \leq 4 \sum_{t=0}^{T-1} \mathrm{Tr}(\Sigma_t E_t^\top E_t \Sigma_t) \leq 4 \sum_{t=0}^{T-1} \|\Sigma_t\|^2 \, \mathrm{Tr}(E_t^\top E_t) \leq 4 \Big( \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} \Big)^2 \sum_{t=0}^{T-1} \mathrm{Tr}(E_t^\top E_t).$$

From Lemma 3.6 we have

$$C(\boldsymbol{K}) - C(\boldsymbol{K}^*) \geq \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t + B^\top P_{t+1} B\|} \, \mathrm{Tr}(E_t^\top E_t) \geq \frac{\underline{\sigma}_{\boldsymbol{X}}}{\max_t \|R_t + B^\top P_{t+1} B\|} \sum_{t=0}^{T-1} \mathrm{Tr}(E_t^\top E_t), \tag{3.25}$$

and hence

$$\sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\|^2 \leq 4\Big(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\Big)^2 \frac{\max_t \|R_t + B^\top P_{t+1} B\|}{\underline{\sigma}_{\boldsymbol{X}}} (C(\boldsymbol{K}) - C(\boldsymbol{K}^*)).$$

For the second claim, using Lemma 3.6 again,

$$
\begin{aligned}
\sum_{t=0}^{T-1}\|K_t\| &= \sum_{t=0}^{T-1}\|(R_t+B^\top P_{t+1}B)^{-1}K_t(R_t+B^\top P_{t+1}B)\| \\
&\le \sum_{t=0}^{T-1}\frac{1}{\sigma_{\min}(R_t)}\|K_t(R_t+B^\top P_{t+1}B)\| \le \sum_{t=0}^{T-1}\frac{1}{\sigma_{\min}(R_t)}\Big(\|E_t\|+\|B^\top P_{t+1}A\|\Big) \\
&\le \sum_{t=0}^{T-1}\left(\frac{\sqrt{\mathrm{Tr}(E_t^\top E_t)}}{\sigma_{\min}(R_t)}+\frac{\|B^\top P_{t+1}A\|}{\sigma_{\min}(R_t)}\right) \le \frac{1}{\underline{\sigma}_{\boldsymbol{R}}}\Big(\sqrt{T\cdot\sum_{t=0}^{T-1}\mathrm{Tr}(E_t^\top E_t)}+\sum_{t=0}^{T-1}\|B^\top P_{t+1}A\|\Big) \\
&\le \frac{1}{\underline{\sigma}_{\boldsymbol{R}}}\Big(\sqrt{T\cdot\frac{\max_t\|R_t+B^\top P_{t+1}B\|}{\underline{\sigma}_{\boldsymbol{X}}}(C(\boldsymbol{K})-C(\boldsymbol{K}^*))}+\sum_{t=0}^{T-1}\|B^\top P_{t+1}A\|\Big).
\end{aligned}
$$

The second inequality holds by the definition of $E_t$ in (3.11), the second last step uses the Cauchy-Schwarz inequality, and the last inequality holds by (3.25). □

*Proof of Theorem 3.3.* In order to show the existence of a positive $\eta$ such that (3.23) holds, it suffices to show there exists a positive lower bound on the RHS of (3.23). By Lemma 3.16 and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\| &\le \sqrt{T\cdot\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\|^2} \\
&\le \sqrt{4T\cdot\Big(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\Big)^2\frac{\max_t\|R_t+B^\top P_{t+1}B\|}{\underline{\sigma}_{\boldsymbol{X}}}(C(\boldsymbol{K})-C(\boldsymbol{K}^*))}.
\end{aligned} \tag{3.26}
$$

Note that if $d<ab+c$ for some $a>0$, $b>0$ $c>0$ and $d>0$, then $\frac{1}{d}>\frac{1}{(a+1)(b+1)(c+1)}$. Also $\frac{1}{a^n+1}>\frac{1}{(a+1)^n}$ for $a>0$ and $n\in\mathbb{N}^+$. Therefore, based on (3.24) and (3.26), $\frac{1}{C_1}$ is bounded below by polynomials in $\frac{1}{\rho}$, $\frac{1}{C(\boldsymbol{K})+1}$, $\frac{1}{\|B\|+1}$, $\frac{1}{\|\boldsymbol{R}\|+1}$, $\frac{1}{\|W\|+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, and $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$.

Now we aim to show that $\frac{1}{\rho}$ is bounded below by some polynomials in the parameters. To see this, let us first show that $\rho$ is bounded above by polynomials in $\|A\|$, $\|B\|$, $\|\!|\boldsymbol{R}|\!\|$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$ and $C(\boldsymbol{K})$. Since $\|B\|\|K_t'-K_t\|\le\frac{\underline{\sigma}_{\boldsymbol{Q}}\underline{\sigma}_{\boldsymbol{X}}}{4C(\boldsymbol{K})}\le\frac{1}{2}$ holds under the assumptions in Lemma 3.15, we have

$$
\max_{0\le t\le T-1}\|A-BK_t'\|\le\max_{0\le t\le T-1}\big(\|A-BK_t\|+\|B\|\,\|K_t'-K_t\|\big)\le\max_{0\le t\le T-1}\|A-BK_t\|+\frac{1}{2},
$$

thus

$$
\begin{aligned}
\rho=\max\Big\{\max_{0\le t\le T-1}\|A-BK_t\|,\max_{0\le t\le T-1}\|A-BK_t'\|,\ 1+\xi\Big\} &\le \max\Big\{\max_{0\le t\le T-1}\|A-BK_t\|+\frac{1}{2},\ 1+\xi\Big\} \\
&\le \max\Big\{\|A\|+\|B\|\sum_{t=0}^{T-1}\|K_t\|+\frac{1}{2},\ 1+\xi\Big\}.
\end{aligned} \tag{3.27}
$$

Given the bound on $\sum_{t=0}^{T-1}\|K_t\|$ by Lemma 3.16 and $\|P_t\|\le\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{X}}}$ by Lemma 3.8, $\rho$ is bounded above by polynomials in $\|A\|$, $\|B\|$, $\|\!|\boldsymbol{R}|\!\|$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$ and $C(\boldsymbol{K})$, or a constant $1+\xi$. Therefore $\frac{1}{\rho}$ is bounded below by polynomials in $\frac{1}{\|A\|+1}$, $\frac{1}{\|B\|+1}$, $\frac{1}{\|\!|\boldsymbol{R}|\!\|+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\underline{\sigma}_{\boldsymbol{R}}$ and $\frac{1}{C(\boldsymbol{K})+1}$, or a constant $\frac{1}{1+\xi}$. Hence, by choosing

$\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ to be an appropriate polynomial in $\frac{1}{C(\boldsymbol{K}^0)}$, $\frac{1}{C(\boldsymbol{K}^0)+1}$, $\frac{1}{\|A\|+1}$, $\frac{1}{\|B\|+1}$, $\frac{1}{\|\boldsymbol{R}\|+1}$, $\frac{1}{\|W\|+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, and $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, (3.23) is satisfied, since by performing gradient descent, $C(\boldsymbol{K}^1) < C(\boldsymbol{K}^0)$. Therefore, by Lemma 3.15, we have

$$C(\boldsymbol{K}^1) - C(\boldsymbol{K}^*) \leq \Big(1 - 2\eta\, \underline{\sigma}_{\boldsymbol{R}} \frac{\underline{\sigma}_{\boldsymbol{X}}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}^0) - C(\boldsymbol{K}^*)\Big),$$

which implies that the cost decreases at $t = 1$. Suppose that $C(\boldsymbol{K}^n) \leq C(\boldsymbol{K}^0)$, then the stepsize condition in (3.23) is still satisfied by Lemma 3.16. Thus, Lemma 3.15 can again be applied for the update at round $n+1$ to obtain:

$$C(\boldsymbol{K}^{n+1}) - C(\boldsymbol{K}^*) \leq \Big(1 - 2\eta\, \underline{\sigma}_{\boldsymbol{R}} \frac{\underline{\sigma}_{\boldsymbol{X}}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}^n) - C(\boldsymbol{K}^*)\Big).$$

For $\epsilon > 0$, provided $N \geq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{2\eta\, \underline{\sigma}_{\boldsymbol{X}}^2\, \underline{\sigma}_{\boldsymbol{R}}} \log \frac{C(\boldsymbol{K}^0)-C(\boldsymbol{K}^*)}{\epsilon}$, we have

$$C(\boldsymbol{K}^N) - C(\boldsymbol{K}^*) < \epsilon.$$

□

# 4 Sample-based Policy Gradient Method with Unknown Parameters

In the setting with unknown parameters, the controller has only simulation access to the model; the model parameters, $A$, $B$, $\{Q_t\}_{t=0}^{T}$, $\{R_t\}_{t=0}^{T-1}$, are unknown. By using a zeroth-order optimization method to approximate the gradient, this section proves the policy gradient method with unknown parameters also leads to a global optimal policy, with both polynomial computational and sample complexities.

Note that in this section, when bounding the Frobenius norm of a matrix, we usually treat the matrix as a stacked vector. Therefore we denote by $D = k \times d$ the dimension of the corresponding vector formed from the $\boldsymbol{K}$ matrix for convenience in the proofs. Therefore in each iteration $n = 1, 2, \cdots, N$, we can update the policy as, for $t = 0, 1, \cdots, T-1$,

$$K_t^{n+1} = K_t^n - \eta \widehat{\nabla_t C(\boldsymbol{K}^n)}, \tag{4.1}$$

where $\widehat{\nabla_t C(\boldsymbol{K}^n)}$ is the estimate of $\nabla_t C(\boldsymbol{K}^n)$. We analyze the following Algorithm 1.

**Algorithm 1 Policy Gradient Estimation with Unknown Parameters**

1: **Input**: $\boldsymbol{K}$, number of trajectories $m$, smoothing parameter $r$, dimension $D$
2: **for** $i \in \{1, \ldots, m\}$ **do**
3: **for** $t \in \{0, \ldots, T-1\}$ **do**
4: Sample the (sub)-policy at time $t$: $\widehat{K}_t^i = K_t + U_t^i$ where $U_t^i$ is drawn uniformly at random over matrices such that $\|U_t^i\|_F = r$.
5: Denote $\widehat{c}_t^{\,i}$ as the single trajectory cost with policy $(\boldsymbol{K}_{-t}, \widehat{K}_t^i) := (K_0, \cdots, K_{t-1}, \widehat{K}_t^i, K_t, \cdots, K_{T-1})$ starting from $x_0^i \sim \mathcal{D}$.
6: **end for**
7: **end for**
8: Return the estimates of $\nabla_t C(\boldsymbol{K})$ for each $t$:

$$\widehat{\nabla_t C(\boldsymbol{K})} = \frac{1}{m} \sum_{i=1}^{m} \frac{D}{r^2}\, \widehat{c}_t^{\,i}\, U_t^i. \tag{4.2}$$

**Remark 4.1.** [Zeroth-order Optimization Approach in the Sub-routine (4.2)] In the estimation of the gradient term (4.2), we adopt a zeroth-order optimization method, using only query access to a sample of the reward function $c(\cdot)$ at input points $\boldsymbol{K}$, without querying the gradients and higher order derivatives of $c(\cdot)$. In a similar way to the observation in [24], the objective $C(\boldsymbol{K})$ may not be finite for every policy $\boldsymbol{K}$ when Gaussian smoothing is applied, therefore $\mathbb{E}_{\boldsymbol{U}\sim\mathcal{N}(0,\sigma^2 I)}[C(\boldsymbol{K}+\boldsymbol{U})]$ may not be well-defined. This is avoidable by smoothing over the surface of a ball. The step (4.2) (in Algorithm 1) provides a procedure to find an (bounded bias) estimate $\widehat{\nabla C(\boldsymbol{K})}$ of $\nabla C(\boldsymbol{K})$.

To guarantee the global convergence of the sample-based algorithm (Algorithm 1), we propose some conditions on the distribution of $x_0$ and $\{w_t\}_{t=0}^{T-1}$, in addition to the finite second moment condition specified in Section 2.

**Definition 4.2.** *A zero-mean random variable $X$*

1. *is said to be sub-Gaussian with variance proxy $\sigma^2$ and we write $X \in SG(\sigma^2)$ if its moment generating function satisfies $\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$.*
2. *is said to be sub-exponential with parameters $(\nu^2,\alpha)$ and we write $X \in SE(\nu^2,\alpha)$, if $\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right)$ for any $\lambda$ such that $|\lambda| \leq \frac{1}{\alpha}$.*

We assume the initial distribution and the noise in the state process dynamics satisfy the following assumptions.

**Assumption 4.3** (Initial State and Noise Process (II))**.**

1. *Initial state: $x_0 = \widetilde{W}_0 z_0$ where $z_0 = (z_{0,1}\cdots,z_{0,d}) \in \mathbb{R}^d$ is a random vector with independent components $z_{0,i}$ which are sub-Gaussian, mean-zero, and have sub-Gaussian parameter $\sigma_0^2$; $\widetilde{W}_0 \in \mathbb{R}^{d\times d}$ is an unknown and deterministic matrix.*
2. *Noise process: $w_t = \widetilde{W} v_t$ where $v_t := (v_{t,1}\cdots,v_{t,d}) \in \mathbb{R}^d$ are IID and independent from $x_0$. $v_t$ has independent components $v_{t,i}$ which are sub-Gaussian, mean-zero, and have sub-Gaussian parameter $\sigma_w^2$, $\forall\, t = 0,1,\cdots,T-1$. $\widetilde{W} \in \mathbb{R}^{d\times d}$ is an unknown and deterministic matrix.*

Note that Assumptions 3.1 and 4.3 serve different purposes in this paper. Assumption 3.1 provides one sufficient condition to assure $\underline{\sigma}_{\boldsymbol{X}} > 0$. Assumption 4.3 is used to guarantee the convergence of the sample based algorithm (Algorithm 1).

In addition to the model parameters specified in Section 3, here we assume $\mathcal{H}(\cdot)$ includes polynomials that are also functions of $\sigma_0$, $\frac{1}{\sigma_0}$, $\frac{1}{\sigma_0+1}$, $\sigma_w$, $\frac{1}{\sigma_w}$, $\frac{1}{\sigma_w+1}$ $\|\widetilde{W}\|$,$\frac{1}{\|\widetilde{W}\|}$, $\frac{1}{\|\widetilde{W}\|+1}$, $\|\widetilde{W}_0\|$, $\frac{1}{\|\widetilde{W}_0\|}$, and $\frac{1}{\|\widetilde{W}_0\|+1}$.

**Theorem 4.4.** *Assume Assumptions 2.1 and 4.3 hold and further assume $\underline{\sigma}_{\boldsymbol{X}} > 0$ and $C(\boldsymbol{K}^0)$ is finite. At every step the policy is updated as in* (4.1)*, that is*

$$K_t^{n+1} = K_t^n - \eta \widehat{\nabla_t C(\boldsymbol{K}^n)},$$

*with $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ and $\widehat{\nabla_t C(\boldsymbol{K}^n)}$ is computed with hyper-parameters $(r,m)$ such that $r < 1/\overline{h}_{radius}$ and $m > \overline{h}_{sample}$ with some fixed polynomials $\overline{h}_{radius} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$ and $\overline{h}_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$. Then for $\epsilon > 0$, if we have*

$$N \geq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\eta\, \underline{\sigma}_{\boldsymbol{X}}^2\, \underline{\sigma}_{\boldsymbol{R}}} \log \frac{C(\boldsymbol{K}^0) - C(\boldsymbol{K}^*)}{\epsilon},$$

*it holds that $C(\boldsymbol{K}^N) - C(\boldsymbol{K}^*) \leq \epsilon$ with high probability (at least $1-\exp(-D)$).*

The proof of Theorem 4.4 is based on a perturbation analysis of $C(\boldsymbol{K})$ and $\nabla_t C(\boldsymbol{K})$, smoothing and the gradient descent analysis of the procedures in Algorithm 1. We provide the perturbation analysis and the smoothing analysis in Sections 4.1 and 4.2, respectively. We defer the proof of Theorem 4.4 to Section 4.3.

**Projected Policy Gradient Method.** In many situations constrained optimization problems arise and the *projected* gradient descent method is one popular approach to solve such problems. Recall the projection of a point $y \in \mathbb{R}^{k\times(T\times d)}$ onto a set $S_{\boldsymbol{K}} \subset \mathbb{R}^{k\times(T\times d)}$ is defined as

$$\Pi_{S_{\boldsymbol{K}}}(y) = \underset{x\in S_{\boldsymbol{K}}}{\arg\min}\, \frac{1}{2}\left\|x-y\right\|_2^2. \tag{4.3}$$

Then the projected policy gradient (PPG) updating rule can be defined as

$$\boldsymbol{K}^{n+1} = \Pi_{S_{\boldsymbol{K}}}\left(\boldsymbol{K}^n - \eta\widehat{\nabla C(\boldsymbol{K}^n)}\right), \tag{4.4}$$

where $\widehat{\nabla C(\boldsymbol{K}^n)} = \left(\widehat{\nabla_0 C(\boldsymbol{K}^n)}, \cdots, \widehat{\nabla_{T-1} C(\boldsymbol{K}^n)}\right)$ denotes the estimate of $\nabla C(\boldsymbol{K}^n)$.

If the projection set $S_{\boldsymbol{K}}$ is convex and closed, the projection onto $S_{\boldsymbol{K}}$ is non-expansive, that is, $\left\|\Pi_{S_{\boldsymbol{K}}}(x) - \Pi_{S_{\boldsymbol{K}}}(y)\right\|_2 \le \left\|x-y\right\|_2$. Therefore the results in Theorem 4.4 for the standard policy gradient method can be easily generalized to the following Theorem 4.5 for the PPG version.

**Theorem 4.5.** *Assume Assumptions 2.1 and 4.3 hold, and the projection set of policy $\boldsymbol{K}$, denoted by $S_{\boldsymbol{K}}$, is convex and closed. Further assume $\underline{\sigma_{\boldsymbol{X}}} > 0$ and $C(\boldsymbol{K}^0)$ is finite. At every step the policy is updated as in (4.4), that is*

$$\boldsymbol{K}^{n+1} = \Pi_{S_{\boldsymbol{K}}}\left(\boldsymbol{K}^n - \eta\widehat{\nabla C(\boldsymbol{K}^n)}\right)$$

*with $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ and $\widehat{\nabla_t C(\boldsymbol{K}^n)}$ ($t = 0, 1, \cdots, T-1$) is computed with hyper-parameters $(r, m)$ such that $r < 1/\widehat{h}_{radius}$ and $m > \widehat{h}_{sample}$ with some fixed polynomials $\widehat{h}_{radius} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$ and $\widehat{h}_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$. Then for $\epsilon > 0$, if we have*

$$N \ge \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\eta\,\underline{\sigma_{\boldsymbol{X}}}^2\,\underline{\sigma_{\boldsymbol{R}}}}\log\frac{C(\boldsymbol{K}^0) - C(\boldsymbol{K}^*)}{\epsilon},$$

*it holds that $C(\boldsymbol{K}^N) - C(\boldsymbol{K}^*) \le \epsilon$ with high probability (at least $1 - \exp(-D)$).*

## 4.1 Perturbation analysis of $C(\boldsymbol{K})$ and $\nabla_t C(\boldsymbol{K})$

This section shows that the objective function $C(\boldsymbol{K})$ and its gradient are stable with respect to small perturbations. The proofs of the following Lemmas can be found in Appendix B.4.

**Lemma 4.6** ($C(\boldsymbol{K})$ Perturbation)**.** *Assume Assumptions 2.1 and 4.3 hold, $\underline{\sigma_{\boldsymbol{X}}} > 0$, and $\boldsymbol{K}'$ such that, $\forall\, t = 0, 1, \cdots, T-1$,*

$$\|K_t' - K_t\| \le \min\left\{\frac{(\rho^2-1)\,\underline{\sigma_{\boldsymbol{Q}}}\,\underline{\sigma_{\boldsymbol{X}}}}{2T(\rho^{2T}-1)(2\rho+1)(C(\boldsymbol{K}) + \underline{\sigma_{\boldsymbol{Q}}}\,T\|W\|)\|B\|}, \|K_t\|, \frac{1}{T}\right\}, \tag{4.5}$$

*where $\rho$ is defined in (3.16). Then there exists a polynomial $h_{cost} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ such that*

$$|C(\boldsymbol{K}') - C(\boldsymbol{K})| \le h_{cost}\left|\left|\left|\boldsymbol{K}' - \boldsymbol{K}\right|\right|\right|.$$

**Lemma 4.7** ($\nabla_t C(\boldsymbol{K})$ Perturbation)**.** *Under the same assumptions as in Lemma 4.6, there exists a polynomial $h_{grad} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ such that*

$$\|\nabla_t C(\boldsymbol{K}') - \nabla_t C(\boldsymbol{K})\| \le h_{grad}\left|\left|\left|\boldsymbol{K}' - \boldsymbol{K}\right|\right|\right|,$$

*and*

$$\|\nabla_t C(\boldsymbol{K}') - \nabla_t C(\boldsymbol{K})\|_F \le h_{grad}\left|\left|\left|\boldsymbol{K}' - \boldsymbol{K}\right|\right|\right|_F.$$

### 4.2 Smoothing and the Gradient Descent Analysis

In this section, Lemma 4.8 provides the formula for the perturbed gradient term, Lemma 4.9 provides the concentration inequality for finite samples, and Lemma 4.10 provides the guarantees for the gradient approximation.

Recall that $D = k \times d$. Let $\mathbb{S}_r$ represent the uniform distribution over the points with norm $r$ in dimension $D$, and $\mathbb{B}_r$ represent the uniform distribution over all points with norm at most $r$ in dimension $D$. For each $K_t$ ($t = 0, 1, \cdots, T-1$), the algorithm performs gradient descent on the following function:

$$C_t^r(\boldsymbol{K}) = \mathbb{E}_{V_t \sim \mathbb{B}_r}\left[C(\boldsymbol{K} + \boldsymbol{V}_t)\right], \tag{4.6}$$

where $\boldsymbol{V}_t := (0, \cdots, V_t, \cdots, 0)$ and $V_t \in \mathbb{R}^{k\times d}$.

**Lemma 4.8.** *Assume $C(\boldsymbol{K})$ is finite,*

$$\nabla_t C_t^r(\boldsymbol{K}) = \frac{D}{r^2}\mathbb{E}_{U_t \sim \mathbb{S}_r}[C(\boldsymbol{K} + \boldsymbol{U}_t)U_t]. \tag{4.7}$$

The proof of Lemma 4.8 is similar to the proof of [24, Lemma 29] and hence omitted.

We first state two facts on sub-Gaussian and sub-exponential random variables. Firstly, if $X$ and $Y$ are zero-mean independent random variables such that $X \in SG(\sigma_x^2)$ and $Y \in SG(\sigma_y^2)$, then $XY \in SE(\sigma_x\sigma_y, 4\sigma_x\sigma_y)$. Secondly, if $X_1, \cdots, X_n$ are zero-mean independent random variables such that $X_i \in SE(\nu_i^2, \alpha_i)$, then

$$\sum_{i=1}^n X_i \in SE\left(\sum_{i=1}^n \nu_i^2, \max_i \alpha_i\right).$$

Utilizing above two facts, we have the following.

**Lemma 4.9.** *Assume Assumptions 2.1 and 4.3 hold and $\underline{\sigma}_{\boldsymbol{X}} > 0$, then there exist polynomials $\nu \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ and $\alpha \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ such that*

$$\left[\sum_{t=0}^{T-1}\left(x_t^\top Q_t x_t + u_t^\top R_t u_t\right) + x_T^\top Q_T x_T\right]$$

*is sub-exponential with parameter $\left(\nu^2, \alpha\right)$. Here $\{x_t\}_{t=0}^T$ is the dynamics under policy $\boldsymbol{K}$.*

*Proof.* We first observe that, by direct calculation,

$$\left[\sum_{t=0}^{T-1}\left(x_t^\top Q_t x_t + u_t^\top R_t u_t\right) + x_T^\top Q_T x_T\right] = x_0^\top P_0 x_0 + \sum_{t=0}^{T-1} w_t^\top P_{t+1} w_t. \tag{4.8}$$

Note that by (3.9) and Proposition 3.4, $P_t$ is symmetric and positive definite. The Frobenius norm $\|\cdot\|_F$ and the spectral norm $\|\cdot\|$ of the matrix $P_t \in \mathbb{R}^{d\times d}$ have the following property:

$$\|P_t\| \le \|P_t\|_F \le \sqrt{d}\|P_t\|, \ \forall\, t = 0, 1, \cdots, T. \tag{4.9}$$

Let $\widehat{\sigma} = \max\{\sigma_0, \sigma_w\}$. Given the Hanson-Wright inequality (Theorem 2.5 in [4]),

$$\begin{aligned}\mathbb{P}\left(\left|w_t^\top P_{t+1} w_t - \mathbb{E}\left[w_t^\top P_{t+1} w_t\right]\right| \ge t\right) &= \mathbb{P}\left(\left|v_t^\top(\widetilde{W}^\top P_{t+1}\widetilde{W})v_t - \mathbb{E}\left[v_t^\top(\widetilde{W}^\top P_{t+1}\widetilde{W})v_t\right]\right| \ge t\right)\\ &\le 2\exp\left(-c\min\left\{\frac{t^2}{2\widehat{\sigma}^4\|\widetilde{W}^\top P_{t+1}\widetilde{W}\|_F^2}, \frac{t}{\widehat{\sigma}^2\|\widetilde{W}^\top P_{t+1}\widetilde{W}\|}\right\}\right),\end{aligned} \tag{4.10}$$

for some universal constant $c > 0$ which is independent of $P_{t+1}$ and $w_t$.

Combining (4.9), (4.10) and Lemma 3.8,

$$\begin{aligned}\mathbb{P}\left(\left|w_t^\top P_{t+1}w_t - \mathbb{E}\left[w_t^\top P_{t+1}w_t\right]\right| \geq t\right) &\leq 2\exp\left(-c\min\left\{\frac{t^2}{2\widehat{\sigma}^4\, d\,\|P_{t+1}\|^2\|\widetilde{W}\|^4}, \frac{t}{\widehat{\sigma}^2\|P_{t+1}\|\|\widetilde{W}\|^2}\right\}\right)\\ &\leq 2\exp\left(-c\min\left\{\frac{t^2}{2\widehat{\sigma}^4\|\widetilde{W}\|^4\, d\, C^2(\boldsymbol{K})/\underline{\sigma}_{\boldsymbol{X}}{}^2}, \frac{t}{\widehat{\sigma}^2\|\widetilde{W}\|^2 C(\boldsymbol{K})/\underline{\sigma}_{\boldsymbol{X}}}\right\}\right).\end{aligned}$$

Therefore the random variable $w_t^\top P_{t+1}w_t$ is sub-exponential with parameters $\left(\frac{\widehat{\sigma}^4\|\widetilde{W}\|^4 dC^2(\boldsymbol{K})}{c\,\underline{\sigma}_{\boldsymbol{X}}{}^2}, \frac{\widehat{\sigma}^2\|\widetilde{W}\|^2 C(\boldsymbol{K})}{2c\,\underline{\sigma}_{\boldsymbol{X}}}\right)$. In the same way $x_0^\top P_0 x_0$ is sub-exponential with parameters $\left(\frac{\widehat{\sigma}^4\|\widetilde{W}_0\|^4 dC^2(\boldsymbol{K})}{c\,\underline{\sigma}_{\boldsymbol{X}}{}^2}, \frac{\widehat{\sigma}^2\|\widetilde{W}_0\|^2 C(\boldsymbol{K})}{2c\,\underline{\sigma}_{\boldsymbol{X}}}\right)$. Let $\overline{\sigma} = \max\{\|\widetilde{W}_0\|, \|\widetilde{W}\|\}$. Since $\{w_t\}_{t=0}^{T-1}$ are IID and independent from $x_0$, we have (4.8) is sub-exponential with parameters

$$\left((T+1)\,\frac{\widehat{\sigma}^4\overline{\sigma}^4 dC^2(\boldsymbol{K})}{c\,\underline{\sigma}_{\boldsymbol{X}}{}^2}, \frac{\widehat{\sigma}^2\overline{\sigma}^2 C(\boldsymbol{K})}{2c\,\underline{\sigma}_{\boldsymbol{X}}}\right).$$

□

Define

$$\widetilde{\nabla}_t := \frac{1}{m}\sum_{i=1}^m\left(\frac{D}{r^2}C(\boldsymbol{K}+\boldsymbol{U}_t^i)U_t^i\right)$$

as the average of perturbed cost functions across $m$ scenarios which is an empirical approximation of (4.7). Similarly, define

$$\widehat{\nabla}_t := \frac{1}{m}\sum_{i=1}^m\left(\frac{D}{r^2}\left[\sum_{t=0}^{T-1}\left((x_t^i)^\top Q_t x_t^i + (u_t^i)^\top R_t u_t^i\right) + (x_T^i)^\top Q_T x_T^i\right]U_t^i\right) \tag{4.11}$$

as the average of perturbed and single-trajectory-based cost functions across $m$ scenarios, which is the same as (4.2) in Algorithm 1. Note that in order to calculate $\widetilde{\nabla}_t$, we require access to $C(\boldsymbol{K}+\boldsymbol{U}_t^i)$, which involves the calculation of expectations with respect to unknown initial states and state noises. This may be restrictive in some settings. On the other hand, the calculation of $\widehat{\nabla}_t$ only involves single-trajectory-based cost functions.

**Lemma 4.10.** *Assume Assumptions 2.1 and 4.3 hold, and $\underline{\sigma}_{\boldsymbol{X}} > 0$. Given any $\epsilon$, there are fixed polynomials $h_{radius} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ and $h_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ such that when $r \leq 1/h_{radius}$, with $m \geq h_{sample}$ samples of $U_t^1, \cdots, U_t^m \sim \mathbb{S}_r$ for each $t = 0, \cdots, T-1$,*

$$\left\|\widetilde{\nabla}_t - \nabla_t C(\boldsymbol{K})\right\|_F \leq \epsilon,$$

*holds with high probability (at least $1 - \left(\frac{D}{\epsilon}\right)^{-D}$). In addition, there is a polynomial $h_{sample,2} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ such that when $r \leq 1/h_{radius}$, with $m \geq h_{sample} + h_{sample,2}$ samples of $U_t^1, ..., U_t^m \sim \mathbb{S}_r$ for each $t = 0, \cdots, T-1$,*

$$\left\|\widehat{\nabla}_t - \nabla_t C(\boldsymbol{K})\right\|_F \leq \frac{3}{2}\epsilon,$$

*holds with high probability (at least $1 - 2\left(\frac{D}{\epsilon}\right)^{-D}$). Here, for each $i = 1, 2, \cdots, m$, $\{x_t^i\}_{t=0}^T$ and $\{u_t^i\}_{t=0}^{T-1}$ are the dynamics and controls for a single path sampled using policy $\boldsymbol{K} + \boldsymbol{U}_t^i$.*

*Proof.* Note that

$$\widetilde{\nabla}_t - \nabla_t C(\boldsymbol{K}) = (\nabla_t C_t^r(\boldsymbol{K}) - \nabla_t C(\boldsymbol{K})) + (\widetilde{\nabla}_t - \nabla_t C_t^r(\boldsymbol{K})),$$

where $C_t^r$ is defined in (4.6).

For the first term, choose $h_{radius} = \max\{1/r_0, 4h_{grad}/\epsilon\}$ ($r_0$ is chosen later), where $h_{grad} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ is defined in Lemma 4.7. By Lemma 4.7 when $r \leq 1/h_{radius} \leq \epsilon/4h_{grad}$, for $\boldsymbol{V}_t := (0, \cdots, V_t, \cdots, 0)$ where $V_t \sim \mathbb{B}_r$, we have

$$\|\nabla_t C(\boldsymbol{K} + \boldsymbol{V}_t) - \nabla_t C(\boldsymbol{K})\|_F \leq h_{grad} \|\boldsymbol{V}_t\|_F \leq h_{grad} \frac{\epsilon}{4h_{grad}} = \frac{\epsilon}{4}. \tag{4.12}$$

Since $\nabla_t C_t^r(\boldsymbol{K}) = \mathbb{E}_{V_t \sim \mathbb{B}_r}[\nabla_t C(\boldsymbol{K} + \boldsymbol{V}_t)]$, we have

$$\|\nabla_t C(\boldsymbol{K} + \boldsymbol{V}_t) - \nabla_t C_t^r(\boldsymbol{K})\|_F \leq \frac{\epsilon}{4},$$

by (4.12) and the continuity of $\nabla_t C$. Therefore

$$\|\nabla_t C_t^r(\boldsymbol{K}) - \nabla_t C(\boldsymbol{K})\|_F \leq \|\nabla_t C(\boldsymbol{K} + \boldsymbol{V}_t) - \nabla_t C(\boldsymbol{K})\|_F + \|\nabla_t C(\boldsymbol{K} + \boldsymbol{V}_t) - \nabla_t C_t^r(\boldsymbol{K})\|_F \leq \frac{\epsilon}{2} \tag{4.13}$$

holds by triangle inequality. We choose $r_0$ such that for any $\boldsymbol{U}_t \sim \mathbb{S}_r$, we have that $C(\boldsymbol{K} + \boldsymbol{U}_t) \leq 2C(\boldsymbol{K})$. By Lemma 4.6, we can pick $1/r_0 = h_{cost}/C(\boldsymbol{K})$, then $|C(\boldsymbol{K} + \boldsymbol{U}_t) - C(\boldsymbol{K})| \leq r_0 \cdot h_{cost} \leq C(\boldsymbol{K})$.

For the second term, by Lemma 4.8, $\mathbb{E}[\widetilde{\nabla}_t] = \nabla_t C_t^r(\boldsymbol{K})$, and each individual sample is bounded by $2DC(\boldsymbol{K})/r$, so by the Operator-Bernstein inequality [28, Theorem 12] with

$$m \geq h_{sample} = \Theta\left(D \left(\frac{D \cdot C(\boldsymbol{K})}{r\epsilon}\right)^2 \log(D/\epsilon)\right),$$

we have

$$\mathbb{P}\left[\left\|\widetilde{\nabla}_t - \nabla_t C_t^r(\boldsymbol{K})\right\|_F \leq \frac{\epsilon}{2}\right] \geq 1 - \left(\frac{D}{\epsilon}\right)^{-D}. \tag{4.14}$$

Note that $h_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ since $1/r > h_{radius} \in \mathcal{H}(1/\epsilon, C(\boldsymbol{K}))$. Adding these two terms together and applying the triangle inequality gives the result.

For the second part, note that

$$\mathbb{E}_{x_0, \boldsymbol{w}}[\widehat{\nabla}_t] = \widetilde{\nabla}_t. \tag{4.15}$$

By Lemma 4.9,

$$\left[\sum_{t=0}^{T-1} \left((x_t^i)^\top Q_t x_t^i + (u_t^i)^\top R_t u_t^i\right) + (x_T^i)^\top Q_T x_T^i\right]$$

is sub-exponential with parameters $(\nu^2, \alpha)$. Therefore,

$$Z_i := \left(\frac{D}{r^2}\left[\sum_{t=0}^{T-1} \left((x_t^i)^\top Q_t x_t^i + (u_t^i)^\top R_t u_t^i\right) + (x_T^i)^\top Q_T x_T^i\right] U_t^i\right)$$

is sub-exponential matrix with parameters $(\widetilde{\nu}^2, \widetilde{\alpha}) := \left(\frac{D}{r^2}\nu^2, \alpha\right)$. Then by Operator-Berinstein inequality [28, Theorem 12],

$$\mathbb{P}\left[\left\|\frac{1}{m}\sum_{i=1}^m Z_i - \mathbb{E}[Z_1]\right\|_F \leq t\right] \geq 1 - 2D\exp\left(-m\frac{t^2}{2\widetilde{\nu}^2}\right),$$

when $t \leq \frac{\widetilde{\nu}^2}{\widetilde{\alpha}}$. That is, there exists a polynomial $h_{sample,2} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ where

$$h_{sample,2} := h_{sample,2}\left(D, \frac{1}{\epsilon}, \frac{1}{r}, \sigma_0, \sigma_w, \|\widetilde{W_0}\|, \|\widetilde{W}\|, C(\boldsymbol{K}), \frac{1}{\underline{\sigma}_{\boldsymbol{X}}}\right) = \Theta\left(D\left(\frac{\widetilde{\nu}}{\epsilon}\right)^2 \log(D/\epsilon)\right),$$

such that when $m \geq h_{sample,2}$,

$$\mathbb{P}\left[\left\|\widehat{\nabla}_t - \widetilde{\nabla}_t\right\|_F \leq \frac{\epsilon}{2}\right] \geq 1 - \left(\frac{D}{\epsilon}\right)^{-D}. \tag{4.16}$$

Combining (4.16) with (4.14) and (4.13), we arrive at the desired result. □

### 4.3 Proof of Theorem 4.4

With the results in Section 4.1 and Section 4.2, now we are ready to prove the main theorem.

*Proof of Theorem 4.4.* By Lemma 3.15 and by choosing $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ such that the step size condition (3.23) is satisfied,

$$C(\boldsymbol{K}') - C(\boldsymbol{K}^*) \leq \left(1 - 2\eta\, \underline{\sigma}_{\boldsymbol{R}} \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\right)\left(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\right).$$

Recall the definition of $\widehat{\nabla}_t$ in (4.11) and let $K_t'' = K_t - \eta\widehat{\nabla}_t$ be the iterate that uses the approximate gradient. We will show later that given enough samples, the gradient can be estimated with enough accuracy that makes sure

$$|C(\boldsymbol{K}'') - C(\boldsymbol{K}')| \leq \eta\, \underline{\sigma}_{\boldsymbol{R}} \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\epsilon. \tag{4.17}$$

That means as long as $C(\boldsymbol{K}) - C(\boldsymbol{K}^*) \geq \epsilon$, we have

$$C(\boldsymbol{K}'') - C(\boldsymbol{K}^*) \leq \left(1 - \eta\, \underline{\sigma}_{\boldsymbol{R}} \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\right)\left(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\right).$$

Then the same proof as that of Theorem 3.3 gives the convergence guarantee.

Now let us prove (4.17). First note that $C(\boldsymbol{K}'') - C(\boldsymbol{K}')$ is bounded. By Lemma 4.6, if $\|K_t'' - K_t'\| \leq \eta\, \underline{\sigma}_{\boldsymbol{R}} \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{\|\Sigma_{\boldsymbol{K}^*}\|} \cdot \epsilon/(T \cdot h_{cost})$, where $h_{cost} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ is the polynomial in Lemma 4.6, then (4.17) holds. To get this bound, recall $K_t' = K_t - \eta\nabla_t C(\boldsymbol{K})$ in (3.22) and writing $\nabla_t = \nabla_t C(\boldsymbol{K})$ for ease of exposition, observe that $K_t'' - K_t' = \eta(\nabla_t - \widehat{\nabla}_t)$, therefore it suffices to make sure

$$\|\nabla_t - \widehat{\nabla}_t\| \leq \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}}{T\|\Sigma_{\boldsymbol{K}^*}\| h_{cost}}\epsilon.$$

By Lemma 4.10, it is enough to pick $\overline{h}_{radius} = h_{radius}(3T\|\Sigma_{\boldsymbol{K}^*}\| h_{cost}(C(\boldsymbol{K}))/(2\, \underline{\sigma}_{\boldsymbol{X}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}\, \epsilon), C(\boldsymbol{K})) \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$, and

$$\overline{h}_{sample} \;=\; h_{sample}\left(\frac{3h_{cost}(C(\boldsymbol{K}))\|\Sigma_{\boldsymbol{K}^*}\|}{2\, \underline{\sigma}_{\boldsymbol{X}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}\, \epsilon}, C(\boldsymbol{K})\right) + h_{sample,2}\left(\frac{3h_{cost}(C(\boldsymbol{K}))\|\Sigma_{\boldsymbol{K}^*}\|}{2\, \underline{\sigma}_{\boldsymbol{X}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}\, \epsilon}, C(\boldsymbol{K})\right).$$

This gives the desired upper bound on $\|\nabla_t - \widehat{\nabla}_t\|$ with high probability (at least $1 - 2(\epsilon/D)^D$).

Since the number of steps is a polynomial, we have $TN = o(\epsilon^D)$. By the union bound with probability at least

$$\left(1 - 2\left(\frac{\epsilon}{D}\right)^D\right)^{TN} \geq 1 - 2TN\left(\frac{\epsilon}{D}\right)^D \geq 1 - \exp(-D),$$

we have $\|\nabla_t - \widehat{\nabla}_t\| \leq \frac{\underline{\sigma_{\boldsymbol{X}}}^2 \underline{\sigma_{\boldsymbol{R}}}}{T\|\Sigma_{\boldsymbol{K}^*}\| h_{cost}}\epsilon$, $\forall\, t = 0, 1, \cdots, T-1$. Therefore,

$$C(\boldsymbol{K}'') - C(\boldsymbol{K}^*) \leq \Big(1 - \eta\, \underline{\sigma_{\boldsymbol{R}}}\, \frac{\underline{\sigma_{\boldsymbol{X}}}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big). \tag{4.18}$$

This implies $C(\boldsymbol{K}'') < C(\boldsymbol{K})$. To guarantee that (4.18) holds at each iteration $n = 1, 2, \cdots, N$, it suffices to pick $\overline{h}_{radius} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$ and $\overline{h}_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$. The rest of the proof is the same as that of Theorem 3.3. Note again that in the smoothing, because the function value is monotonically decreasing, and by the choice of radius, all the function values encountered are bounded by $2C(\boldsymbol{K}^0)$, so the polynomials are indeed bounded throughout the algorithm. □

### 4.4 Discussion

**Remark 4.11** (Comparison with [24])**.** The proofs of our main results, Theorems 3.3 and 4.4, are different to those from [24]. Firstly, to prove the gradient dominant condition, [24] only required conditions on the distribution of the initial position. However, we need conditions to guarantee the non-degeneracy of the state covariance matrix at any time. Secondly, the extra randomness from the sub-Gaussian noise needs to be taken care of in the perturbation analysis of $\Sigma_{\boldsymbol{K}}$. Finally, we need more advanced concentration inequalities to provide the number of samples and number of simulation trajectories that leads to the theoretical guarantee in the case with unknown parameters.

**Remark 4.12** (Non-stationary Dynamics)**.** Note that our framework can be generalized to non-stationary dynamics, that is, for $t = 0, 1, \cdots, T-1$,

$$x_{t+1} = A_t x_t + B_t u_t + w_t,\; x_0 \sim \mathcal{D}. \tag{4.19}$$

with $\{A_t\}_{t=0}^{T-1}$ and $\{B_t\}_{t=0}^{T-1}$ time-dependent state parameters.

**Remark 4.13** (Other Policy Gradient Methods)**.** Our convergence and sample complexity analysis could be applied to other policy gradient methods, including the Natural policy gradient method and the Gauss-Newton method, in the framework of the LQR with stochastic dynamics and finite horizons.

## 5 Numerical Experiments

The performance of the PPG algorithm (4.4) is demonstrated for the optimal liquidation problem with single asset and the empirical analysis of the policy gradient method (4.1) in higher dimensions is also provided with synthetic data. We will specifically focus on the following questions.

- In practice, how fast do the policy gradient algorithm and the PPG algorithm with known and unknown parameters converge to the true solution?
- How does the deadline (finite horizon) influence the optimal policy?
- When the real-word system does not exactly follow the LQR framework, does policy-gradient outperforms mis-specified LQR models?

This section is organized as follows. We demonstrate the performance of the PPG algorithms for optimal liquidation problem with single asset in LQR framework in Section 5.1. We then show that without the LQR model specification, the learned policy from the policy gradient algorithm improves the Almgren-Chriss solution in Section 5.2. Finally, we test the performance of the algorithm with unknown parameters in high dimensions in Section 5.3.

### 5.1 Optimal Liquidation within the LQR Framework

Recall the set up of the optimal liquidation problem in (2.1). By convention, we write the control in the feedback form as $u_t = -K_t x_t$. Writing $K_t = (k_t^1, k_t^2)$, we have $u_t = -k_t^1 S_t - k_t^2 q_t$, the state equation becomes

$$x_{t+1} = \begin{pmatrix} 1+\gamma k_t^1 & \gamma k_t^2 \\ k_t^1 & 1+k_t^2 \end{pmatrix} x_t + w_t.$$

In the liquidation problem, we assume $u_t \geq 0$ $(0 \leq t \leq T-1)$. That is, $k_t^1 \leq 0$ and $k_t^2 \leq 0$ $(0 \leq t \leq T-1)$.

**Assumption 5.1** (Assumptions for the Optimal Liquidation Problems)**.** We assume

(1) $\gamma k_t^1 + k_t^2 > -1$ $(0 \leq t \leq T-1)$;

(2) $\beta > \frac{\gamma}{2}$.

**Justification of the Assumption.** Assumption 5.1-(1) is essential to ensure that the liquidation problem is well defined. First, $\gamma k_t^1 > -1$ makes sure that the stock price process $\{S_t\}_{t=0}^T$ is well-behaved:

$$\mathbb{E}[S_{t+1}] = \mathbb{E}[S_t] - \gamma \mathbb{E}[u_t] = (1+\gamma k_t^1)\mathbb{E}[S_t] + \gamma k_t^2 q_t.$$

If $\gamma k_t^1 < -1$, then $\mathbb{E}[S_{t+1}] \leq 0$ since $k_t^2 \leq 0$. Second, $k_t^2 \geq -1$ guarantees that inventory will not be negative. Note that

$$q_{t+1} = q_t - (-k_t^1 S_t - k_t^2 q_t) = (1+k_t^2) q_t + k_t^1 S_t.$$

If $k_t^2 \leq -1$ and $q_t > 0$, then $q_{t+1} < 0$. Assumption 5.1-(2) implies that the temporary market impact is "bigger" than one half of the permanent market impact, which is consistent with the empirical evidence [9] and assumptions in [8].

**Learning to Liquidate.** In practice, traders may not know the market impact parameter $\gamma$. But one can always take some $\bar{\gamma} > \gamma$ based on some basic understandings of the market and perform a PPG algorithm to the closed convex set $S_{\boldsymbol{K}}$:

$$S_{\boldsymbol{K}} := \left\{ \boldsymbol{K} = (K_0, \cdots, K_{T-1}) : K_t = (k_t^1, k_t^2),\ \bar{\gamma} k_t^1 + k_t^2 \geq -1 + \zeta,\ k_t^1 \leq 0,\ k_t^2 \leq 0,\ \forall t = 0, \cdots, T-1 \right\}, \tag{5.1}$$

with some small parameter $\zeta > 0$.

In practice $\gamma$ is usually on the order of $10^{-5} \sim 10^{-6}$ (See Table 3 in Appendix A) and hence a universal upper bound $\bar{\gamma}$ in (5.1) is not a strong assumption for a given portfolio of stocks to liquidate.

**Proposition 5.2.** *Assume $\boldsymbol{K} \in S_{\boldsymbol{K}}$ and Assumptions 2.1, 4.3 and 5.1 hold, we have $\underline{\sigma}_{\boldsymbol{X}} > 0$ and $\{P_t^{\boldsymbol{K}}\}_{t=0}^T$ derived from* (3.9) *are positive definite for the optimal liquidation problem* (2.7) *and* (2.9)*.*

The proof of Proposition 5.2 is deferred to Appendix B.5. It is easy to check that the projection set $S_{\boldsymbol{K}}$ defined in (5.1) is convex and closed. Along with Proposition 5.2, the convergence result in Theorem 4.5 holds for the liquidation problem (2.7) and (2.9) as long as the conditions in Proposition 5.2 are satisfied.

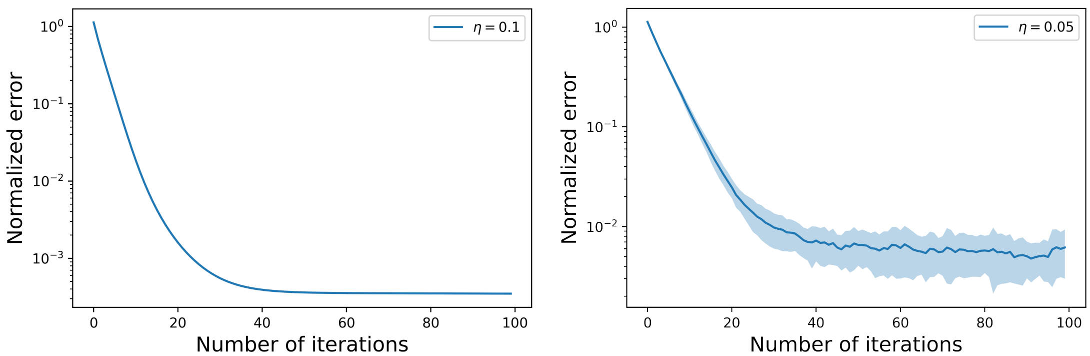
We test the performance of the PPG algorithm with projection set $S_{\boldsymbol{K}}$ on Apple (AAPL) and Facebook (FB) stocks. The market simulator of the associated LQR framework is constructed with NASDAQ ITCH data and the details can be found in Appendix A.

**Performance Measure.** We use the following *normalized error* to quantify the performance of a given policy $\boldsymbol{K}$,

$$\text{Normalized error} = \frac{C(\boldsymbol{K}) - C(\boldsymbol{K}^*)}{C(\boldsymbol{K}^*)},$$

where $\boldsymbol{K}^*$ is the optimal policy defined in (2.5).

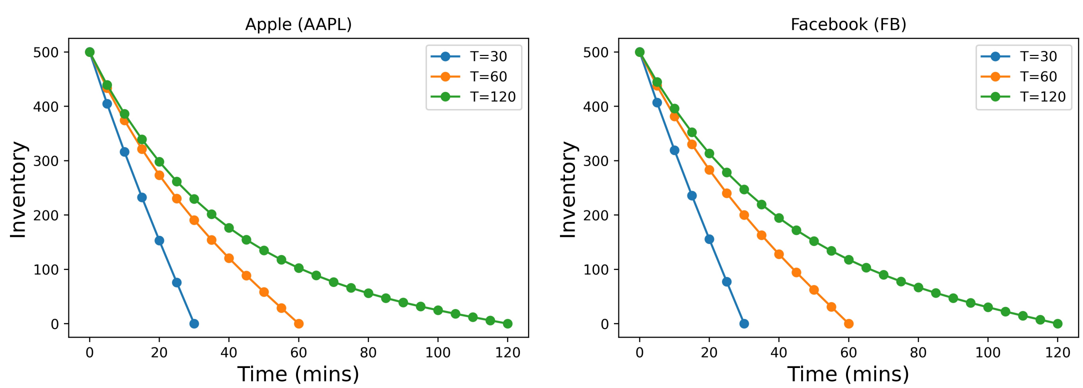
**Set-up.** (1) Parameters: $\phi = 5 \times 10^{-6}$ (for both AAPL and FB), $\epsilon = 10^{-8}$, $T = 10$; smoothing parameter $r = 0.6$, number of trajectories $m = 200$; initial policy $\boldsymbol{K}^0 \in \mathbb{R}^{1\times 2T}$ with $\{\boldsymbol{K}^0\}_{ij} = -0.2$ for all $i$, $j$, for both algorithms with known and unknown parameters; step sizes are indicated in the figures; $\bar{\gamma} = 5\times 10^{-5}$, $\zeta = 10^{-12}$ for the projection set. (2) Initialization: Assume the initial inventory $q_0$ follows $\mathcal{N}(500, 1)$. The small variance of the initial inventory distribution is used to guarantee the initial state covariance matrix is positive definite. In practice, the algorithm converges with deterministic initial inventories.




(a) PPG with known parameters ($\eta = 0.1$). (b) PPG with unknown parameters ($\eta = 0.05$).

Figure 1: Performance of the PPG algorithms (50 simulation scenarios).

**Convergence.** Both PPG algorithms with known parameters and unknown parameters show a reasonable level of accuracy within 50 iterations (that is the normalized error is less than $10^{-2}$). The PPG algorithm with known parameters has almost no fluctuations across the 50 scenarios. By choosing $m = 200$, the performance of the PPG algorithm with unknown parameters is stable with relatively small fluctuations (see the blue area in Figure 1b) across the 50 scenarios.




(a) AAPL. (b) FB.

Figure 2: Optimal inventory trajectory under different deadlines (200 simulation scenarios).

**Impact of the Deadline.** The optimal policy is sensitive to the deadline in that the shapes of the optimal inventory trajectories are different with different deadlines. See Figure 2 for both AAPL and FB with $T = 30, 60$ and 120 minutes. The liquidation speed is almost linear when $T$ is small; and it is faster in the initial trading phase and slower at the end when $T$ is relatively large.

**Impact of the Parameter $\phi$.** Recall that in (2.9) the parameter $\phi$ is used to balance the expected terminal wealth $\mathbb{E}[C]$ and the variance of the terminal wealth var$[C]$. To show the impact of $\phi$, we set $\phi$ to be $10^{-4}$, $10^{-5}$, $10^{-6}$, and $10^{-7}$ and show the corresponding inventory trajectories in Figure 3. The optimal liquidation speed is almost linear when $\phi$ is small, while it is faster in the initial trading phase and slower at the end when $\phi$ is relatively large.

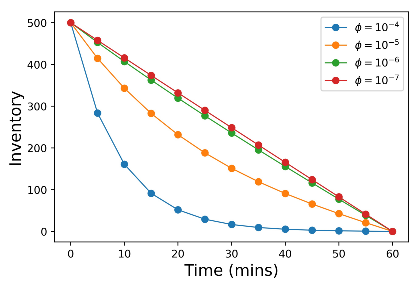
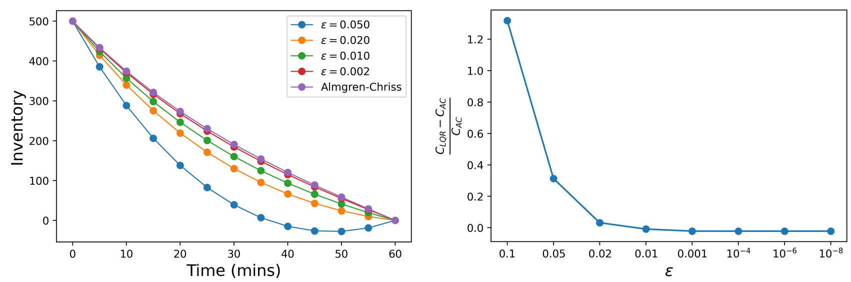



Figure 3: Inventory trajectories of AAPL under different $\phi$ (average across 200 simulation scenarios).




(a) Inventory trajectories.

(b) Relative cost difference.

Figure 4: Original Almgren-Chriss framework versus LQR formulation under different $\epsilon$ (AAPL).

**Impact of the Parameter $\epsilon$.** Recall that our liquidation formulation (2.9) differs from the Almgren-Chriss formulation (2.8) by an additional regularization term $\sum_{t=0}^{T} \epsilon S_t^2$. The role of this term is to enable the problem to be cast in the LQR framework and to guarantee the well-definedness of the Ricatti equation. From Figure 4a, the optimal policies and inventory trajectories are close to the Almgren-Chriss solution when $\epsilon \leq 0.01$. However, when $\epsilon = 0.05$, the optimal policy is far away from the Almgren-Chriss solution. We show the difference between $C_{\text{AC}}$, defined in (2.8), and $C_{\text{LQR}}(\epsilon)$, defined in (2.9), in Figure 4b. We see that $C_{\text{LQR}}(\epsilon)$ is close to $C_{\text{AC}}$ when $\epsilon < 0.02$ and is markedly different from $C_{\text{AC}}$ when $\epsilon \geq 0.02$. It is worth noticing that when $\epsilon = 0$, the algorithm does converge to the Almgren-Chriss solution in our setting although the convergence of the algorithm in this case is not guaranteed by our theoretical results.

## 5.2 Learning to Liquidate without Model Specification

In practice, the dynamics of the trading system may not be exactly those assumed in the LQR framework but we might expect that the policy gradient method could still perform well when the system is "nearly" linear quadratic as the execution of the policy gradient method does not rely on the model specification. In this section, we consider liquidation problems in the Limit Order Book (LOB) setting. A LOB is a list of orders that a trading venue, for example the NASDAQ exchange, uses to record the interest of buyers and sellers in a particular financial instrument. There are two types of orders the buyers (sellers) can submit: a limit buy (sell) order with a preferred price for a given volume or a market buy (sell) order with a given volume which will be immediately executed with the best available limit sell (buy) orders. Here we perform the policy gradient method to learn the optimal strategies to liquidate using market orders in the LOB.

We denote by $S_t$ the mid-price of the asset at time $t$, that is the average of the best-bid price and best-ask price. At each time $t$, the decision is to liquidate an amount $u_t$ of the asset. The action $u_t$ will have an impact on the market, with possibly both temporary and permanent impacts. Unlike the LQR framework or the classical Almgren-Chriss model, where dynamics are assumed to follow some stochastic model, here we run the policy gradient method directly on the LOB without any assumption on how the mid-price $S_t$ moves and what are the forms of the market impacts. Denote by $q_t = q_{t-1} - u_{t-1}$ the inventory at time $t$. We restrict the admissible controls to be of the linear feedback form $u_t = -K_t(S_t, q_t)^\top$ with some $K_t \in \mathbb{R}^{1\times 2}$.

The cost $c_t = \phi'(q_t - u_t)^2 - r_t(u_t)$ at time $t$ consists of two parts. The first part $\phi'(q_t - u_t)^2$ is the holding cost of the inventory weighted by a parameter $\phi'$. The quantity $r_t(u_t)$ is the amount we receive by liquidating $u_t$ shares at time $t$. Note that $r_t(\cdot)$ may depend on $S_t$ and other market observables. For example, if we liquidate $u_t = 1000$ shares of the asset with the market conditions given in Table 1, then the amount received would be

$$r_t(u_t) = 397 \times 200.1 + 412 \times 200.0 + (1000 - 397 - 412) \times 199.9 = 200020.6.$$

This transaction moves the best bid price two levels down. This is commonly referred to as the *temporary impact* of a market order.

| Bid level | One | Two | Three | Four | Five |
|---|---|---|---|---|---|
| Bid price (USD) | 200.1 | 200.0 | 199.9 | 199.8 | 199.7 |
| Volume available | 397 | 412 | 502 | 442 | 529 |

Table 1: One snapshot of the LOB.

**Performance Metric: Implementation Shortfall [40]**

$$\mathrm{IS}(\boldsymbol{u}) = \left(\sum_{t=0}^{T-1} c_t(u_t) + c_T\left(q_0 - \sum_{t=0}^{T-1} u_t\right)\right) - c_0(q_0). \tag{5.2}$$

The first term of (5.2) is the cost of implementing policy $\boldsymbol{u}$ over the horizon $[0, T]$. The second term is the cost when liquidating $q_0$ market orders at time 0. If we expect $\boldsymbol{u}$ is better than liquidating everything at time 0, then $\mathrm{IS}(\boldsymbol{u}) < 0$. A smaller implementation shortfall implies the strategy is more profitable.

We use the following *relative performance* (evaluated on a single trajectory) to compare the performance of two policies $\boldsymbol{u}^1$ and $\boldsymbol{u}^2$,

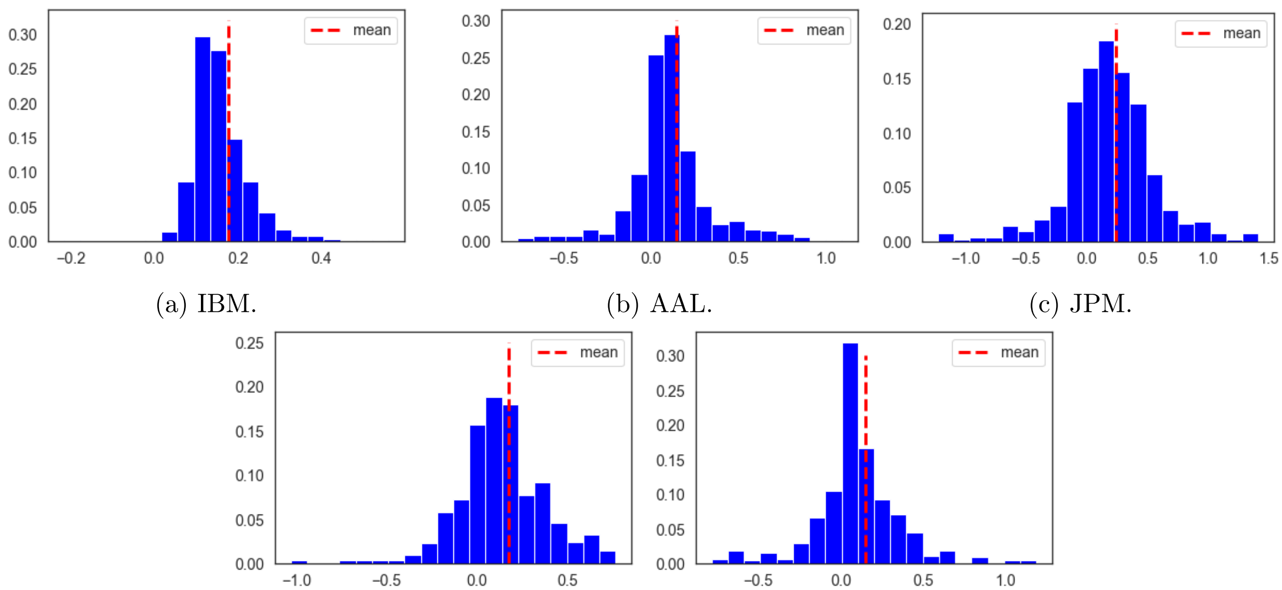
$$\text{Relative performance} = \frac{\mathrm{IS}(\boldsymbol{u}^2) - \mathrm{IS}(\boldsymbol{u}^1)}{|\mathrm{IS}(\boldsymbol{u}^2)|}.$$

**Experiment Set-up.** We consider the LOB data consisting of the best 5 levels and we assume the trading frequency $\Delta = 1$ minute and the trading horizon $T = 10$ minutes. We perform a numerical analysis for five different stocks, Apple (AAPL), Facebook (FB), International Business Machines Corporation (IBM), American Airlines (AAL) and JP Morgan (JPM), during the period from 01/01/2019 to 12/31/2019. The data is divided into two sets, a training set with data between 10:00AM-12:00AM 01/01/2019-08/31/2019 and a test set with data between 10:00AM-12:00AM 09/01/2019-12/31/2019.

We take $\phi' = 5 \times 10^{-6}$; $T = 10$; smoothing parameter $r = 0.4$; number of trajectories $m = 200$; initial policy $\boldsymbol{K}^0 \in \mathbb{R}^{1\times 20}$ with $(\boldsymbol{K}^0)_{ij} = -0.2$ for all $i, j$; and step size $\eta = 10^{-6}$. We assume the initial inventory follows $q_0 = 2000$. We compare the performance of the policy gradient method with the Almgren-Chriss solution with fitted parameters given in Table 3. In the Almgren-Chriss model, we set $\phi = \sigma^2\phi'$ to ensure a reasonable comparison.

**Results.** From Table 2 and Figure 5, the policy gradient method improves on the Almgren-Chriss solution by around 20% on five different stocks from different financial sectors. Note that the goal of the policy gradient method is to learn the global minimizer of the expected cost function, hence it is expected that the Almgren-Chriss solution could perform better than the policy gradient method for some sample trajectories, as shown in Figure 5. This result is compatible with the performance of the Q-learning algorithms [30]. The drawback of Q-learning algorithms is that the computational complexity is highly dependent on the size of the set of (discrete) states and actions, where as the policy gradient method can handle continuous states and actions.

We conjecture that the policy gradient method may be capable of learning the global "optimal" solution for a larger class of models that are "similar" to the LQR framework with stochastic dynamics and finite time horizon. In addition, as the policy gradient method is a model-free algorithm, it is more robust with respect to model mis-specification as compared to the Almgren-Chriss framework.




(a) IBM. (b) AAL. (c) JPM.

(d) FB. (e) AAPL.

Figure 5: Empirical distribution of the relative performance on the test set.

| Asset | IBM | AAL | JPM | FB | AAPL |
|---|---|---|---|---|---|
| In sample | 0.173 | 0.152 | 0.251 | 0.181 | 0.165 |
| (std) | (0.09) | (0.27) | (0.31) | (0.32) | (0.31) |
| Out of sample | 0.178 | 0.146 | 0.245 | 0.175 | 0.163 |
| (std) | (0.08) | (0.29) | (0.36) | (0.24) | (0.37) |

Table 2: Average relative performance of the policy gradient ($\boldsymbol{u}^1$) compared to Almgren-Chriss solution ($\boldsymbol{u}^2$).

### 5.3 Learning LQR in Higher Dimensions

In practice we can perform the policy gradient method for the optimal liquidation problem with multiple assets. However it is difficult to capture the cross impact and permanent impact with historical LOB data. Therefore we test the performance of the policy gradient method in higher dimensions on synthetic data consisting of a four-dimensional state variable and a two-dimensional control variable. The parameters are randomly picked such that the conditions for our LQR framework are satisfied.

**Set-up.** (1) Parameters:

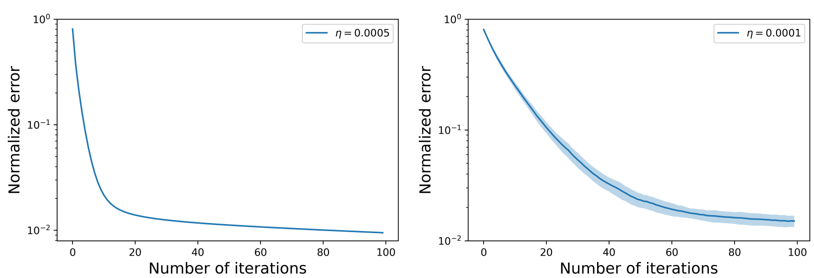
$$A = \begin{pmatrix} 0.5 & 0.05 & 0.1 & 0.2 \\ 0 & 0.2 & 0.3 & 0.1 \\ 0.06 & 0.1 & 0.2 & 0.4 \\ 0.05 & 0.2 & 0.15 & 0.1 \end{pmatrix}, \; B = \begin{pmatrix} -0.05 & -0.01 \\ -0.005 & -0.01 \\ -1 & -0.01 \\ -0.01 & -0.9 \end{pmatrix}, \; Q_t = \begin{pmatrix} 1 & 0.5 & -0.01 & 0 \\ -0.1 & 1.1 & 0.2 & 0 \\ 0 & 0.1 & 0.9 & -0.06 \\ 0.03 & 0 & -0.1 & 0.88 \end{pmatrix},$$

$$R_t = \begin{pmatrix} 0.4 & -0.2 \\ -0.3 & 0.7 \end{pmatrix}, \; W = \begin{pmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.3 \end{pmatrix},$$

$Q_T = Q_t$, $T = 10$; smoothing parameter $r = 1$, number of trajectories $m = 200$; initial policy $\boldsymbol{K}^0 \in \mathbb{R}^{2\times 40}$ with $\{\boldsymbol{K}^0\}_{ij} = 0.05$ for all $i$, $j$, for both known and unknown parameters;

(2) Initialization: We assume $x_0 = (x_0^1, x_0^2, x_0^3, x_0^4)^\top$ and $x_0^i$ are independent. $x_0^1, x_0^2, x_0^3$, and $x_0^4$ are sampled from $\mathcal{N}(5, 0.1)$, $\mathcal{N}(2, 0.3)$, $\mathcal{N}(8, 1)$, $\mathcal{N}(5, 0.5)$.

**Convergence.** For the high-dimensional case, the normalized error falls below the threshold $10^{-2}$ within 80 iterations for the policy gradient algorithm with known parameters. It takes substantially more iterations for the policy gradient algorithm with unknown parameters to have an error near such a threshold, which is as expected.




(a) Known parameters ($\eta = 0.0005$).

(b) Unknown parameters ($\eta = 0.0001$).

Figure 6: Performance of the policy gradient algorithms (50 simulation scenarios)

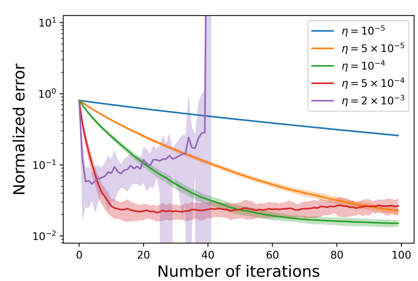



Figure 7: Performance of the policy gradient algorithm with unknown parameters under different step size $\eta$ (50 simulation scenarios).

**Outcomes from Varying the Parameter $\eta$.** The performance of the policy gradient algorithm also depends on the values of the step size $\eta$. We show how the values of the step size $\eta \in [10^{-5}, 2 \times 10^{-3}]$ affect the convergence of the policy gradient algorithm with unknown parameters in Figure 7. A tiny step size leads to slow convergence (see the blue line when $\eta = 10^{-5}$) and a larger step size may cause divergence (see the purple line when $\eta = 2 \times 10^{-3}$).

## References


[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

[2] Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. *AISTATS 2017 - 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[3] Marc Abeille, Alessandro Lazaric, Xavier Brokmann, et al. LQG for portfolio optimization. *Available at SSRN 2863925*, 2016.

[4] Radoslaw Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20, 2015.

[5] Aurélien Alfonsi, Antje Fruth, and Alexander Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10(2):143–157, 2010.

[6] Alessandro Alla and Valeria Simoncini. Order reduction approaches for the algebraic riccati equation and the lqr problem. In *Numerical Methods for Optimal Control Problems*, pages 89–109. Springer, 2018.

[7] Robert Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10(1):1–18, 2003.

[8] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

[9] Robert Almgren, Chee Thum, Emmanuel Hauptmann, and Hong Li. Direct estimation of equity market impact. *Risk*, 18(7):58–62, 2005.

[10] Brian D. O. Anderson and John B Moore. *Optimal Control: Linear Quadratic Methods*. Courier Corporation, 2007.

[11] Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Corporation, 2013.

[12] Wenhang Bao and Xiao-yang Liu. Multi-agent deep reinforcement learning for liquidation strategy analysis. *arXiv preprint arXiv:1906.11046*, 2019.

[13] Dimitri Bertsekas. *Dynamic Programming And Optimal Control*, volume 1. Athena Scientific, 3rd edition, 2005.

[14] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[15] Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. LQR through the lens of first order methods: discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

[16] Jingjing Bu, Afshin Mesbahi, and Mehran Mesbahi. Policy gradient-based algorithms for continuous-time linear quadratic control. *arXiv preprint arXiv:2006.09178*, 2020.

[17] Jingjing Bu, Lillian J Ratliff, and Mehran Mesbahi. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019.

[18] René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

[19] Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance. *arXiv preprint arXiv:2003.10014*, 2020.

[20] Rama Cont, Arseniy Kukanov, and Sasha Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88, 2014.

[21] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

[22] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, 2020.

[23] Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. Efficient learning of distributed linear-quadratic control policies. *SIAM Journal on Control and Optimization*, 58(5):2927–2951, 2020.

[24] Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *Proceedings of the 35th International Conference on Machine Learning*, pages 1467–1476, 2018.

[25] Claude-Nicolas Fiechter. PAC adaptive control of linear systems. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 72–80, 1997.

[26] Jim Gatheral and Alexander Schied. Optimal trade execution under geometric Brownian motion in the Almgren and Chriss framework. *International Journal of Theoretical and Applied Finance*, 14(03):353–368, 2011.

[27] Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning robust controllers for linear quadratic systems with multiplicative noise via policy gradient. *arXiv preprint arXiv:1905.13547*, 2019.

[28] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[29] Xin Guo, Renyuan Xu, and Thaleia Zariphopoulou. Entropy regularization for mean field games with learning. *arXiv preprint arXiv:2010.00145*, 2020.

[30] Dieter Hendricks and Diane Wilcox. A reinforcement learning extension to the Almgren-Chriss framework for optimal trade execution. In *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 457–464. IEEE, 2014.

[31] Morteza Ibrahimi, Adel Javanmard, and Benjamin V Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, pages 2636–2644, 2012.

[32] Zeyu Jin, Johann Michael Schmitt, and Zaiwen Wen. On the analysis of model-free methods for the linear quadratic regulator. *arXiv preprint arXiv:2007.03861*, 2020.

[33] Laura Leal, Mathieu Laurière, and Charles-Albert Lehalle. Learning a functional control for high-frequency finance. *arXiv preprint arXiv:2006.09611*, 2020.

[34] Weiwei Li and Emanuel Todorov. Iterative linear quadratic regulator design for nonlinear biological movement systems. In *ICINCO*, pages 222–229, 2004.

[35] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2916–2925. PMLR, 2019.

[36] Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 673–680, 2006.

[37] Brian Ning, Franco Ho Ting Ling, and Sebastian Jaimungal. Double deep Q-learning for optimal execution. *arXiv preprint arXiv:1812.06600*, 2018.

[38] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with Thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205. IEEE, 2017.

[39] Panagiotis Patrinos, Sergio Trimboli, and Alberto Bemporad. Stochastic MPC for real-time market-based optimal power dispatch. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 7111–7116. IEEE, 2011.

[40] Andre F Perold. The implementation shortfall: Paper versus reality. *Journal of Portfolio Management*, 14(3):4, 1988.

[41] Andreas Schmidt and Bernard Haasdonk. Reduced basis approximation of large scale parametric algebraic riccati equations. *ESAIM: Control, Optimisation and Calculus of Variations*, 24(1):129–151, 2018.

[42] Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 5005–5014, 2018.

[43] Yasuaki Wasa, Kengo Sakata, Kenji Hirata, and Kenko Uchida. Differential game-based load frequency control for power networks and its integration with electricity market mechanisms. In *2017 IEEE Conference on Control Technology and Applications (CCTA)*, pages 1044–1049. IEEE, 2017.

[44] Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. On the global convergence of actor-critic: a case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*, 2019.

[45] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11602–11614, 2019.

[46] Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deep reinforcement learning for trading. *The Journal of Financial Data Science*, 2(2):25–40, 2020.

# A Market Simulator for Linear Price Dynamics

We estimate the parameters for the LQR model using NASDAQ ITCH data taken from Lobster[1].

**Permanent Price Impact and Volatility** The model in (2.7) implies that prices changes are proportional to the *market-order flow imbalances* (MFI). We adopt the framework from [20], namely that the price change $\Delta S$ is given by

$$\Delta S = \gamma \, \text{MFI} + \sigma \, \epsilon, \tag{A.1}$$

with $\text{MFI} = M^b - M^s$ where $M^s$ and $M^b$ are the volumes of market sell orders and market buy orders respectively during a time interval $\Delta T = 5$mins and $\epsilon \sim \mathcal{N}(0, 1)$. We then estimate $\gamma$ and $\sigma$ from the data.

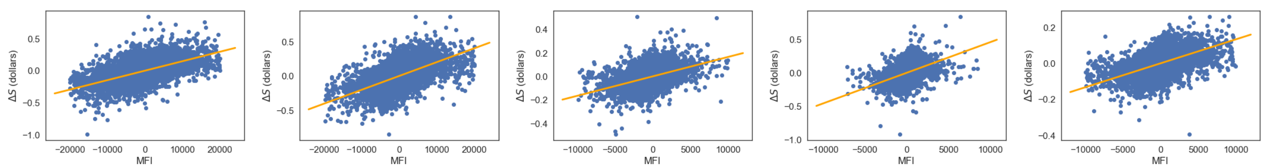



Figure 8: Relationship between MFI and $\Delta S$. (Example (from left to right): AAP, FB, JPM, IBM and AAL, 10:00AM-11:00AM 01/01/2019-08/31/2019, $\Delta T = 1$min)

**Temporary Price Impact** We assume the LOB has a flat shape with constant queue length $l$ for the first few levels. Figure 9 shows the average queue lengths for the first 5 levels so that our assumption is not too unreasonable. Therefore the following equation, on the amount received when we liquidate $u$ shares with best bid price $S$, holds

$$u(S - \beta u) = \int_{S - \frac{u\,\Delta}{l}}^{S} l v dv.$$

Therefore we have $\beta = \frac{\Delta}{2l}$, where $\Delta$ is the tick size and $l$ is the average queue length.

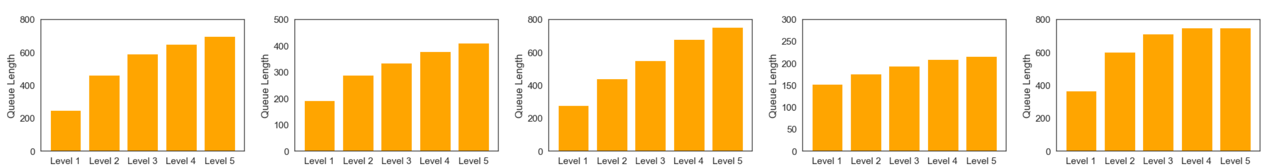



Figure 9: Average queue length (volume) of the first five levels on the limit buy side (Example (from left to right): AAP, FB, JPM, IBM and AAL, 10:00AM-11:00AM 01/01/2019-08/31/2019 with 5000 samples uniformly sampled with natural time clock in each trading day.)

**Parameter Estimation** See the estimates for AAPL, FB, IBM, JPM, and AAL in Table 3.

[1] https://lobsterdata.com/

| Paramters/Stock | AAPL | FB | IBM | JPM | AAL |
|---|---|---|---|---|---|
| $\beta$ | $1.03 \times 10^{-5}$ | $1.30 \times 10^{-5}$ | $2.65 \times 10**-5$ | $9.28 \times 10^{-6}$ | $3.27 \times 10^{-5}$ |
| $\gamma$ | $7.27 \times 10^{-6}$ | $1.40 \times 10^{-5}$ | $4.60 \times 10^{-5}$ | $1.65 \times 10^{-5}$ | $1.3310 \times 10^{-5}$ |
| $\sigma$ | 0.107 | 0.115 | 0.082 | 0.059 | 0.042 |

Table 3: Parameter estimation from NASDAQ ITCH Data (10:00AM-11:00PM 01/01/2019-08/31/2019).

# B Proofs of Technical Results

We now give the proofs that were omitted in the text.

## B.1 Proofs in Section 3.1

**Proof of Lemma 3.2.** Denote by $\{x_t\}_{t=0}^T$ the state trajectory induced by an arbitrary control $\boldsymbol{K}$. By Assumption 3.1 the matrix $\mathbb{E}[x_0 x_0^\top]$ is positive definite. For $t \geq 1$, we have

$$\mathbb{E}[x_t x_t^\top] = (A - BK_{t-1})\mathbb{E}[x_{t-1}x_{t-1}^\top](A - BK_{t-1})^\top + \mathbb{E}[w_{t-1}w_{t-1}^\top].$$

Now $(A-BK_{t-1})\mathbb{E}[x_{t-1}x_{t-1}^\top](A-BK_{t-1})^\top$ is positive semi-definite and $\mathbb{E}[w_{t-1}w_{t-1}^\top]$ is positive definite. Hence $\mathbb{E}[x_t x_t^\top]$ is positive definite and as a result $\underline{\sigma}_{\boldsymbol{X}} > 0$. In this case, we can simply take $\underline{\sigma}_{\boldsymbol{X}} = \min(\mathbb{E}[x_0 x_0^\top], \sigma_{\min}(W))$. $\square$

**Proof of Proposition 3.4.** This can be proved by backward induction. For $t = T$, $P_T^{\boldsymbol{K}} = Q_T$ is positive definite since $Q_T$ is positive definite. Assume $P_{t+1}^{\boldsymbol{K}}$ is positive definite for some $t+1$, then take any $z \in \mathbb{R}^d$ such that $z \neq 0$,

$$z^\top P_t^{\boldsymbol{K}} z = z^\top Q_t\, z + z^\top K_t^\top R_t K_t z + z^\top (A - BK_t)^\top P_{t+1}^{\boldsymbol{K}} (A - BK_t)\, z > 0.$$

The last inequality holds since $z^\top Q_t\, z > 0$, $z^\top K_t^\top R_t K_t z \geq 0$ and $z^\top (A - BK_t)^\top P_{t+1}^{\boldsymbol{K}} (A - BK_t)\, z \geq 0$. By backward induction, we have $P_t^{\boldsymbol{K}}$ positive definite, $\forall\, t = 0, 1, \cdots, T$. $\square$

To prove Lemma 3.6, let us start with a useful result for the value function. Define the value function $V_{\boldsymbol{K}}(x, \tau)$ for $\tau = 0, 1, \cdots, T-1$, as

$$V_{\boldsymbol{K}}(x, \tau) = \mathbb{E}_{\boldsymbol{w}} \left[ \sum_{t=\tau}^{T-1} (x_t^\top Q_t x_t + u_t^\top R_t u_t) + x_T^\top Q_T x_T \,\middle|\, x_\tau = x \right] = x^\top P_\tau x + L_\tau,$$

with terminal condition

$$V_{\boldsymbol{K}}(x, T) = x^\top Q_T x,$$

where $L_\tau$ is defined in (3.10). We then define the $Q$ function, $Q_{\boldsymbol{K}}(x, u, \tau)$ for $\tau = 0, 1, \cdots, T-1$ as

$$Q_{\boldsymbol{K}}(x, u, \tau) = x^\top Q_\tau x + u^\top R_\tau u + \mathbb{E}_{w_\tau} \left[ V_{\boldsymbol{K}}(Ax + Bu + w_\tau, \tau + 1) \right],$$

and the advantage function

$$A_{\boldsymbol{K}}(x, u, \tau) = Q_{\boldsymbol{K}}(x, u, \tau) - V_{\boldsymbol{K}}(x, \tau).$$

Note that $C(\boldsymbol{K}) = \mathbb{E}_{x_0 \sim \mathcal{D}}[V(x_0, 0)]$. Then we can write the difference of value functions between $\boldsymbol{K}$ and $\boldsymbol{K'}$ in terms of advantage functions.

**Lemma B.1.** *Assume $\boldsymbol{K}$ and $\boldsymbol{K'}$ have finite costs. Denote $\{x'_t\}_{t=0}^{T}$ and $\{u'_t\}_{t=0}^{T-1}$ as the state and control sequences of a single trajectory generated by $\boldsymbol{K'}$ starting from $x'_0 = x_0 = x$, then*

$$V_{\boldsymbol{K'}}(x,0) - V_{\boldsymbol{K}}(x,0) = \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x'_t, u'_t, t)\right], \tag{B.1}$$

*and*

$$\begin{aligned}&A_{\boldsymbol{K}}(x, -K'_\tau x, \tau)\\ &\quad= 2x^\top (K'_\tau - K_\tau)^\top E_\tau x + x^\top (K'_\tau - K_\tau)^\top (R_\tau + B^\top P_{\tau+1} B)(K'_\tau - K_\tau)x,\end{aligned}$$

*where $E_\tau$ is defined in* (3.11).

*Proof.* Denote by $c'_t(x)$ the cost generated by $\boldsymbol{K'}$ with a single trajectory starting from $x'_0 = x_0 = x$. That is,

$$c'_t(x) = (x'_t)^\top Q_t x'_t + (u'_t)^\top R_t u'_t,\ t = 0, 1, \cdots, T-1,$$

and

$$c'_T(x) = (x'_T)^\top Q_T x'_T,$$

with

$$u'_t = -K'_t x'_t, \quad x'_{t+1} = Ax'_t + Bu'_t + w_t, \quad x'_0 = x.$$

Therefore,

$$\begin{aligned}V_{\boldsymbol{K'}}(x,0) - V_{\boldsymbol{K}}(x,0) &= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T} c'_t(x)\right] - V_{\boldsymbol{K}}(x,0)\\ &= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T} \big(c'_t(x) + V_{\boldsymbol{K}}(x'_t,t) - V_{\boldsymbol{K}}(x'_t,t)\big)\right] - V_{\boldsymbol{K}}(x,0)\\ &= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} \big(c'_t(x) + V_{\boldsymbol{K}}(x'_{t+1},t+1) - V_{\boldsymbol{K}}(x'_t,t)\big)\right]\\ &= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} \big(Q_{\boldsymbol{K}}(x'_t,u'_t,t) - V_{\boldsymbol{K}}(x'_t,t)\big)\middle|\, x_0 = x\right]\\ &= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x'_t,u'_t,t)\middle|\, x_0 = x\right],\end{aligned}$$

where the third equality holds since $c'_T(x) = V_{\boldsymbol{K}}(x'_T, T)$ with the same single trajectory. For $u = -K'_\tau x$,

$$
\begin{aligned}
A_{\boldsymbol{K}}(x,-K'_\tau x,\tau) &= Q_{\boldsymbol{K}}(x,-K'_\tau x,\tau) - V_{\boldsymbol{K}}(x,\tau) \\
&= x^\top (Q_\tau + (K'_\tau)^\top R_\tau K'_\tau)x + \mathbb{E}_{w_\tau}\left[V_{\boldsymbol{K}}((A-BK'_\tau)x + w_\tau, \tau+1)\right] - V_{\boldsymbol{K}}(x,\tau) \\
&= x^\top (Q_\tau + (K'_\tau)^\top R_\tau K'_\tau)x + \Big(x^\top (A-BK'_\tau)^\top P_{\tau+1}(A-BK'_\tau)x + \mathrm{Tr}(WP_{\tau+1}) + L_{\tau+1}\Big) \\
&\quad - \Big(x^\top P_\tau x + L_\tau\Big) \\
&= x^\top (Q_\tau + (K'_\tau - K_\tau + K_\tau)^\top R_\tau (K'_\tau - K_\tau + K_\tau))x \\
&\quad + x^\top (A - BK_\tau - B(K'_\tau - K_\tau))^\top P_{\tau+1}(A - BK_\tau - B(K'_\tau - K_\tau))x \\
&\quad - x^\top (Q_\tau + K_\tau^\top R_\tau K_\tau + (A-BK_\tau)^\top P_{\tau+1}(A-BK_\tau))x \\
&= 2x^\top (K'_\tau - K_\tau)^\top ((R_\tau + B^\top P_{\tau+1}B)K_\tau - B^\top P_{\tau+1}A)x \\
&\quad + x^\top (K'_\tau - K_\tau)^\top (R_\tau + B^\top P_{\tau+1}B)(K'_\tau - K_\tau)x.
\end{aligned} \tag{B.2}
$$

□

**Proof of Lemma 3.6.** First for any $K'_\tau$, from (B.2),

$$
\begin{aligned}
A_{\boldsymbol{K}}(x,-K'_\tau x,\tau) &= Q_{\boldsymbol{K}}(x,-K'_\tau x,\tau) - V_{\boldsymbol{K}}(x,\tau) \\
&= 2\,\mathrm{Tr}(xx^\top (K'_\tau - K_\tau)^\top E_\tau) + \mathrm{Tr}(xx^\top (K'_\tau - K_\tau)^\top (R_\tau + B^\top P_{\tau+1}B)(K'_\tau - K_\tau)) \\
&= \mathrm{Tr}\Big(xx^\top (K'_\tau - K_\tau + (R_\tau + B^\top P_{\tau+1}B)^{-1}E_\tau)^\top (R_\tau + B^\top P_{\tau+1}B) \\
&\quad (K'_\tau - K_\tau + (R_\tau + B^\top P_{\tau+1}B)^{-1}E_\tau)\Big) - \mathrm{Tr}(xx^\top E_\tau^\top (R_\tau + B^\top P_{\tau+1}B)^{-1}E_\tau) \\
&\geq -\mathrm{Tr}(xx^\top E_\tau^\top (R_\tau + B^\top P_{\tau+1}B)^{-1}E_\tau),
\end{aligned} \tag{B.3}
$$

with equality holds when $K'_\tau = K_\tau - (R_\tau + B^\top P_{\tau+1}B)^{-1}E_\tau$. Then,

$$
\begin{aligned}
C(\boldsymbol{K}) - C(\boldsymbol{K}^*) &= -\mathbb{E}\sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x_t^*, u_t^*, t) \\
&\leq \mathbb{E}\sum_{t=0}^{T-1}\mathrm{Tr}\Big(x_t^*(x_t^*)^\top E_t^\top (R_t + B^\top P_{t+1}B)^{-1}E_t\Big) \\
&\leq \|\Sigma_{\boldsymbol{K}^*}\|\sum_{t=0}^{T-1}\mathrm{Tr}(E_t^\top (R_t + B^\top P_{t+1}B)^{-1}E_t) \\
&\leq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{R}}}\sum_{t=0}^{T-1}\mathrm{Tr}(E_t^\top E_t) \qquad \text{(B.4)} \\
&\leq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\,\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}}\sum_{t=0}^{T-1}\mathrm{Tr}(\nabla_t C(\boldsymbol{K})^\top \nabla_t C(\boldsymbol{K})), \qquad \text{(B.5)}
\end{aligned}
$$

where $\underline{\sigma}_{\boldsymbol{X}}$ is defined in (3.3) and $\underline{\sigma}_{\boldsymbol{R}}$ is defined in (3.4). For the lower bound, consider $K'_t = K_t - (R_t +$

$B^\top P_{t+1}B)^{-1}E_t$ where the equality holds in (B.3). Using $C(\boldsymbol{K}^*) \le C(\boldsymbol{K}')$

$$
\begin{aligned}
C(\boldsymbol{K}) - C(\boldsymbol{K}^*) \ge C(\boldsymbol{K}) - C(\boldsymbol{K}') & \\
&= -\mathbb{E}\sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x_t', u_t', t) \\
&= \mathbb{E}\sum_{t=0}^{T-1} \mathrm{Tr}(x_t'(x_t')^\top E_t^\top (R_t + B^\top P_{t+1}B)^{-1}E_t) \\
&\ge \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t + B^\top P_{t+1}B\|} \mathrm{Tr}(E_t^\top E_t)
\end{aligned} \tag{B.6}
$$

□

**Proof of Lemma 3.7.** By lemma B.1 we have

$$
\begin{aligned}
C(\boldsymbol{K}') - C(\boldsymbol{K}) &= \mathbb{E}\left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x_t', -K_t'x_t', t)\right] \\
&= \sum_{t=0}^{T-1}\Big(2\,\mathrm{Tr}(\Sigma_t'(K_t'-K_t)^\top E_t) + \mathrm{Tr}(\Sigma_t'(K_t'-K_t)^\top (R_t + B^\top P_{t+1}B)(K_t'-K_t))\Big).
\end{aligned}
$$

□

**Proof of Lemma 3.8.** For $t = 0, 1, \cdots, T$,

$$
C(\boldsymbol{K}) \ge \mathbb{E}[x_t^\top P_t x_t] \ge \|P_t\|\sigma_{\min}(\mathbb{E}[x_t x_t^\top]) \ge \underline{\sigma}_{\boldsymbol{X}}\,\|P_t\|,
$$

$$
C(\boldsymbol{K}) = \sum_{t=0}^{T-1} \mathrm{Tr}(\mathbb{E}[x_t x_t^\top](Q_t + K_t^\top R_t K_t)) + \mathrm{Tr}(\mathbb{E}[x_T x_T^\top]Q_T) \ge \underline{\sigma}_{\boldsymbol{Q}}\,\mathrm{Tr}(\Sigma_{\boldsymbol{K}}) \ge \underline{\sigma}_{\boldsymbol{Q}}\,\|\Sigma_{\boldsymbol{K}}\|.
$$

Therefore the statement in Lemma 3.8 follows provided that $\underline{\sigma}_{\boldsymbol{X}} > 0$ and Assumption 2.1 holds. □

**Proof of Proposition 3.9 .** Recall that $\Sigma_t = \mathbb{E}\left[x_t x_t^\top\right]$. Note that

$$
\begin{aligned}
\Sigma_1 &= \mathbb{E}\left[x_1 x_1^\top\right] = \mathbb{E}\left[((A - B\,K_0)x_0 + w_0)\,((A - B\,K_0)x_0 + w_0)^\top\right] \\
&= (A - B\,K_0)\Sigma_0\,(A - B\,K_0)^\top + W = \mathcal{G}_0(\Sigma_0) + W.
\end{aligned}
$$

Now we first prove that

$$
\Sigma_t = \mathcal{G}_{t-1}(\Sigma_0) + \sum_{s=1}^{t-1} D_{t-1,s} W D_{t-1,s}^\top + W,\ \forall\, t = 2, 3, \cdots, T. \tag{B.7}
$$

When $t = 2$,

$$
\begin{aligned}
\Sigma_2 &= \mathbb{E}\left[x_2 x_2^\top\right] = \mathbb{E}\left[((A - B\,K_1)x_1 + w_1)\,((A - B\,K_1)x_1 + w_1)^\top\right] \\
&= (A - B\,K_1)\Sigma_1\,(A - B\,K_1)^\top + W = \mathcal{G}_1(\Sigma_0) + (A - BK_1)W(A - BK_1)^\top + W,
\end{aligned}
$$

which satisfies (B.7). Assume (B.7) holds for $t \leq k$. Then for $t = k+1$,

$$\begin{aligned}
\mathbb{E}\left[x_{t+1}x_{t+1}^\top\right] &= \mathbb{E}\left[\left((A - B\,K_t)x_t + w_t\right)\left((A - B\,K_t)x_t + w_t\right)^\top\right] \\
&= (A - B\,K_t)\Sigma_t\,(A - B\,K_t)^\top + W = \mathcal{G}_t(\Sigma_0) + \sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top + W.
\end{aligned}$$

Therefore (B.7) holds, $\forall\, t = 1, 2, \cdots, T$. Finally,

$$\Sigma_{\boldsymbol{K}} = \sum_{t=0}^{T} \Sigma_t = \Sigma_0 + \sum_{t=1}^{T-1} \mathcal{G}_t(\Sigma_0) + \sum_{t=1}^{T-1}\sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top + TW = \mathcal{T}_{\boldsymbol{K}}(\Sigma_0) + \Delta(\boldsymbol{K}, W).$$

□

## B.2 Proofs in Section 3.2

**Proof of Lemma 3.13.** By direct calculation,

$$\|\mathcal{G}_t\| \leq \rho^{2(t+1)}, \quad \text{and} \quad \|\mathcal{G}'_t\| \leq \rho^{2(t+1)}. \tag{B.8}$$

Denote $\mathcal{F}_t = \mathcal{F}_{K_t}$ and $\mathcal{F}'_t = \mathcal{F}_{K'_t}$ to ease the exposition. Then for any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$ and $t \geq 0$,

$$\begin{aligned}
\|(\mathcal{G}'_{t+1} - \mathcal{G}_{t+1})(\Sigma)\| &= \|\mathcal{F}'_{t+1} \circ \mathcal{G}'_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\| \\
&= \|\mathcal{F}'_{t+1} \circ \mathcal{G}'_t(\Sigma) - \mathcal{F}'_{t+1} \circ \mathcal{G}_t(\Sigma) + \mathcal{F}'_{t+1} \circ \mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\| \\
&\leq \|\mathcal{F}'_{t+1} \circ \mathcal{G}'_t(\Sigma) - \mathcal{F}'_{t+1} \circ \mathcal{G}_t(\Sigma)\| + \|\mathcal{F}'_{t+1} \circ \mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\| \\
&= \|\mathcal{F}'_{t+1} \circ (\mathcal{G}'_t - \mathcal{G}_t)(\Sigma)\| + \|(\mathcal{F}'_{t+1} - \mathcal{F}_{t+1}) \circ \mathcal{G}_t(\Sigma)\| \\
&\leq \|\mathcal{F}'_{t+1}\|\,\|(\mathcal{G}'_t - \mathcal{G}_t)(\Sigma)\| + \|\mathcal{G}_t\|\,\|\mathcal{F}'_{t+1} - \mathcal{F}_{t+1}\|\,\|\Sigma\| \\
&\leq \rho^2 \|(\mathcal{G}'_t - \mathcal{G}_t)(\Sigma)\| + \rho^{2(t+1)} \|\mathcal{F}'_{t+1} - \mathcal{F}_{t+1}\|\|\Sigma\|.
\end{aligned}$$

Therefore,

$$\|(\mathcal{G}'_{t+1} - \mathcal{G}_{t+1})(\Sigma)\| \leq \rho^2 \|(\mathcal{G}'_t - \mathcal{G}_t)(\Sigma)\| + \rho^{2(t+1)} \|\mathcal{F}'_{t+1} - \mathcal{F}_{t+1}\|\|\Sigma\|. \tag{B.9}$$

Summing (B.9) up for $t \in \{1, 2, \cdots, T-2\}$ with $\|\mathcal{G}'_0 - \mathcal{G}_0\| = \|\mathcal{F}'_0 - \mathcal{F}_0\|$, we have

$$\sum_{t=0}^{T-1} \left\|(\mathcal{G}_t - \mathcal{G}'_t)(\Sigma)\right\| \leq \frac{\rho^{2T} - 1}{\rho^2 - 1} \Big(\sum_{t=0}^{T-1} \|\mathcal{F}_t - \mathcal{F}'_t\|\Big) \|\Sigma\|.$$

□

## B.3 Proofs in Section 3.3

**Proof of Lemma 3.15.** Given (3.22) and condition (3.23), we have

$$\|K'_t - K_t\| = \eta \|\nabla_t C(\boldsymbol{K})\| \leq \frac{\sigma_{\boldsymbol{Q}}\, \sigma_{\boldsymbol{X}}}{2C(\boldsymbol{K})\|B\|}.$$

Therefore,

$$\|B\|\|K'_t - K_t\| \leq \frac{\sigma_{\boldsymbol{Q}}\, \sigma_{\boldsymbol{X}}}{2C(\boldsymbol{K})} \leq \frac{1}{2}.$$

The last inequality holds since $\underline{\sigma_{\boldsymbol{X}}} \leq \frac{C(\boldsymbol{K})}{\underline{\sigma_{\boldsymbol{Q}}}}$ given by Lemma 3.8. Therefore, by Lemma 3.12,

$$\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\| \leq (2\rho+1)\|B\| \left( \sum_{t=0}^{T-1} \|K_t - K_t'\| \right). \tag{B.10}$$

By Lemmas 3.5 and 3.7,

$$\begin{aligned}
C(\boldsymbol{K}') - C(\boldsymbol{K}) &= \sum_{t=0}^{T-1} \Big[ 2 \operatorname{Tr}\Big(\Sigma_t'(K_t' - K_t)^\top E_t\Big) + \operatorname{Tr}\Big(\Sigma_t'(K_t' - K_t)^\top (R_t + B^\top P_{t+1} B)(K_t' - K_t)\Big)\Big] \\
&= \sum_{t=0}^{T-1} \Big[ -4\eta \operatorname{Tr}\Big(\Sigma_t' \Sigma_t E_t^\top E_t\Big) + 4\eta^2 \operatorname{Tr}\Big(\Sigma_t' \Sigma_t E_t^\top (R_t + B^\top P_{t+1} B) E_t \Sigma_t\Big)\Big] \\
&= \sum_{t=0}^{T-1} \Big[ -4\eta \operatorname{Tr}\Big((\Sigma_t' - \Sigma_t + \Sigma_t) \Sigma_t E_t^\top E_t\Big) + 4\eta^2 \operatorname{Tr}\Big(\Sigma_t' \Sigma_t E_t^\top (R_t + B^\top P_{t+1} B) E_t \Sigma_t\Big)\Big] \\
&\leq \sum_{t=0}^{T-1} \Big[ -4\eta \operatorname{Tr}\Big(\Sigma_t E_t^\top E_t \Sigma_t\Big) + 4\eta \operatorname{Tr}((\Sigma_t' - \Sigma_t)\Sigma_t E_t^\top E_t \Sigma_t \Sigma_t^{-1}) \\
&\qquad + 4\eta^2 \operatorname{Tr}\Big(\Sigma_t' \Sigma_t E_t^\top (R_t + B^\top P_{t+1} B) E_t \Sigma_t\Big)\Big] \\
&\leq \sum_{t=0}^{T-1} \Big[ -4\eta \operatorname{Tr}\Big(\Sigma_t E_t^\top E_t \Sigma_t\Big) + 4\eta \frac{\|\Sigma_t' - \Sigma_t\|}{\sigma_{\min}(\Sigma_t)} \operatorname{Tr}\Big(\Sigma_t E_t^\top E_t \Sigma_t\Big) \\
&\qquad + 4\eta^2 \|\Sigma_t'(R_t + B^\top P_{t+1} B)\| \operatorname{Tr}\Big(\Sigma_t E_t^\top E_t \Sigma_t\Big)\Big] \\
&\leq -\eta\Big(1 - \frac{\sum_{t=0}^{T-1} \|\Sigma_t' - \Sigma_t\|}{\underline{\sigma_{\boldsymbol{X}}}} - \eta \|\Sigma_{\boldsymbol{K}'}\| \sum_{t=0}^{T-1} \|R_t + B^\top P_{t+1} B\|\Big) \sum_{t=0}^{T-1} \Big[ \operatorname{Tr}(\nabla_t C(\boldsymbol{K})^\top \nabla_t C(\boldsymbol{K}))\Big].
\end{aligned} \tag{B.11}$$

By Lemma 3.6, we have

$$C(\boldsymbol{K}') - C(\boldsymbol{K}) \leq -\eta\Big(1 - \frac{\sum_{t=0}^{T-1} \|\Sigma_t' - \Sigma_t\|}{\underline{\sigma_{\boldsymbol{X}}}} - \eta \|\Sigma_{\boldsymbol{K}'}\| \sum_{t=0}^{T-1} \|R_t + B^\top P_{t+1} B\|\Big) \Big(\frac{4\, \underline{\sigma_{\boldsymbol{X}}}^2\, \underline{\sigma_{\boldsymbol{R}}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big) \Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big) \tag{B.12}$$

provided that

$$1 - \frac{\sum_{t=0}^{T-1} \|\Sigma_t' - \Sigma_t\|}{\underline{\sigma_{\boldsymbol{X}}}} - \eta \|\Sigma_{\boldsymbol{K}'}\| \sum_{t=0}^{T-1} \|R_t + B^\top P_{t+1} B\| > 0. \tag{B.13}$$

By (3.21), (3.22), and (B.10),

$$\sum_{t=0}^{T-1} \|\Sigma_t' - \Sigma_t\| \leq \frac{\rho^{2T} - 1}{\rho^2 - 1} \left( \frac{C(\boldsymbol{K})}{\underline{\sigma_{\boldsymbol{Q}}}} + T\|W\| \right) \left( \eta(2\rho+1)\|B\| \sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\| \right).$$

Given the step size condition in (3.23), we have

$$\eta(2\rho+1)\|B\| \sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\| \leq \eta(2\rho+1)\|B\| \Big(T \cdot \max_t \{\|\nabla_t C(\boldsymbol{K})\|\}\Big) \leq \frac{(\rho^2 - 1)\, \underline{\sigma_{\boldsymbol{Q}}}\, \underline{\sigma_{\boldsymbol{X}}}}{2(\rho^{2T} - 1)(C(\boldsymbol{K}) + \underline{\sigma_{\boldsymbol{Q}}}\, T\|W\|)}. \tag{B.14}$$

Then, by Corollary 3.14 and (B.10),

$$
\begin{aligned}
\frac{\|\Sigma_{\boldsymbol{K}'} - \Sigma_{\boldsymbol{K}}\|}{\underline{\sigma}_{\boldsymbol{X}}} &\leq \frac{\rho^{2T}-1}{\rho^2-1}\Big(\sum_{t=0}^{T-1}\|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\|\Big)\frac{\|\Sigma_0\| + T\|W\|}{\underline{\sigma}_{\boldsymbol{X}}} \\
&\leq \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\left(\sum_{t=0}^{T-1}\eta\|\nabla_t C(\boldsymbol{K})\|\right)\frac{C(\boldsymbol{K}) + \underline{\sigma}_{\boldsymbol{Q}}\, T\|W\|}{\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}} \\
&\leq \frac{1}{2},
\end{aligned}
$$

where the last step holds by (B.14). Therefore, the bound of $\|\Sigma_{\boldsymbol{K}'}\|$ in (B.13) is given by

$$
\|\Sigma_{\boldsymbol{K}'}\| \leq \|\Sigma_{\boldsymbol{K}'} - \Sigma_{\boldsymbol{K}}\| + \|\Sigma_{\boldsymbol{K}}\| \leq \frac{1}{2}\,\underline{\sigma}_{\boldsymbol{X}} + \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} \leq \frac{1}{2}\|\Sigma_{\boldsymbol{K}'}\| + \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}, \tag{B.15}
$$

which indicates that $\|\Sigma_{\boldsymbol{K}'}\| \leq \frac{2C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$. Therefore, (B.13) gives

$$
\begin{aligned}
&1 - \frac{\sum_{t=0}^{T-1}\|\Sigma_t' - \Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}} - \eta\|\Sigma_{\boldsymbol{K}'}\|\sum_{t=0}^{T-1}\|R_t + B^\top P_{t+1}B\| \\
&\geq 1 - \frac{(\rho^{2T}-1)}{(\rho^2-1)\,\underline{\sigma}_{\boldsymbol{X}}}\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right)\left(\eta(2\rho+1)\|B\|\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\|\right) \\
&- \eta\frac{2C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\sum_{t=0}^{T-1}\|R_t + B^\top P_{t+1}B\| \\
&= 1 - C_1\eta,
\end{aligned}
$$

where $C_1$ is defined in (3.24). So if $\eta \leq \frac{1}{2C_1}$, then,

$$
1 - \frac{\sum_{t=0}^{T-1}\|\Sigma_t' - \Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}} - \eta\|\Sigma_{\boldsymbol{K}'}\|\sum_{t=0}^{T-1}\|R_t + B^\top P_{t+1}B\| \geq 1 - C_1\eta \geq \frac{1}{2} > 0.
$$

Hence,

$$
C(\boldsymbol{K}') - C(\boldsymbol{K}) \leq -\frac{\eta}{2}\Big(\frac{4\,\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big),
$$

and

$$
C(\boldsymbol{K}') - C(\boldsymbol{K}^*) = \big(C(\boldsymbol{K}') - C(\boldsymbol{K})\big) + (C(\boldsymbol{K}) - C(\boldsymbol{K}^*)) \leq \Big(1 - 2\eta\frac{\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big).
$$

□

### B.4 Proofs in Section 4

**Proof of Lemma 4.6.** Under Assumption 4.3,

$$
\mathbb{E}\left[x_0 x_0^\top\right] = \widetilde{W}_0\mathbb{E}\left[z_0 z_0^\top\right]\widetilde{W}_0^\top,\ \left\|\mathbb{E}\left[x_0 x_0^\top\right]\right\| \leq \sigma_0^2\|\widetilde{W}_0\|^2.
$$

With the sub-Gaussian distributed noise,

$$
W = \mathbb{E}\left[w_t w_t^\top\right] = \widetilde{W}\mathbb{E}\left[v_t v_t^\top\right]\widetilde{W}^\top,
$$

then we have $\|W\| \leq \sigma_w^2 \left\| \widetilde{W}^2 \right\|$.
Denote $S_t = Q_t + K_t^T R_t K_t$, $\forall\, t = 1, \cdots, T-1$. Thus, for $t = 0, 1, \cdots, T-2$,

$$\begin{aligned}
\mathbb{E}[x_{t+1}^\top Q_{t+1} x_{t+1} + u_{t+1}^\top R_{t+1} u_{t+1}] = \mathbb{E}[x_{t+1}^\top S_{t+1} x_{t+1}] = \operatorname{Tr}(\mathbb{E}[x_{t+1}^\top S_{t+1} x_{t+1}]) = \operatorname{Tr}(\mathbb{E}[x_{t+1} x_{t+1}^\top] S_{t+1}) \\
= \operatorname{Tr}\left( \mathcal{G}_t(\Sigma_0) S_{t+1} + \sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top S_{t+1} + W S_{t+1} \right).
\end{aligned}$$

The last equality holds by (B.7). Therefore,

$$\begin{aligned}
C(\boldsymbol{K}') - C(\boldsymbol{K}) = &\underbrace{\mathbb{E}[x_0^\top (K_0')^\top R_0 K_0' x_0 - x_0^\top K_0^\top R_0 K_0 x_0]}_{(I)} + \underbrace{\sum_{t=0}^{T-2} \operatorname{Tr}\left( \mathcal{G}_t'(\Sigma_0) S_{t+1}' - \mathcal{G}_t(\Sigma_0) S_{t+1} \right)}_{(II)} \\
&+ \underbrace{\sum_{t=0}^{T-2} \operatorname{Tr}\left( \sum_{s=1}^{t} \left( D_{t,s}' W (D_{t,s}')^\top S_{t+1}' - D_{t,s} W D_{t,s}^\top S_{t+1} \right) + W(S_{t+1}' - S_{t+1}) \right)}_{(III)} \\
&+ \underbrace{\operatorname{Tr}\left( \mathcal{G}_{T-1}(\Sigma_0) Q_T - \mathcal{G}_{T-1}'(\Sigma_0) Q_T + \sum_{s=1}^{T-1} \left( D_{T-1,s}' W (D_{T-1,s}')^\top Q_T - D_{T-1,s} W D_{T-1,s}^\top Q_T \right) \right)}_{(IV)}.
\end{aligned}$$

For the first term $(I)$,

$$(I) \leq \operatorname{Tr}(\mathbb{E}[x_0 x_0^\top]) \| (K_0')^\top R_0 K_0' - K_0^\top R_0 K_0 \|$$

For the second term $(II)$, since

$$\begin{aligned}
\sum_{t=0}^{T-2} \left( \operatorname{Tr}\left( \mathcal{G}_t(\Sigma_0) S_{t+1} \right) \right) &= \mathbb{E}\left[ \sum_{t=0}^{T-2} \left( \operatorname{Tr}\left( \Pi_{i=0}^{t} (A - BK_i) x_0 x_0^\top \Pi_{i=0}^{t} (A - BK_{t-i})^\top S_{t+1} \right) \right) \right] \\
&\leq \operatorname{Tr}\left( \mathbb{E}\left[ x_0 x_0^\top \right] \right) \left\| \sum_{t=0}^{T-2} \mathcal{G}_t(S_{t+1}) \right\|,
\end{aligned}$$

we have,

$$(II) \leq \operatorname{Tr}\left( \mathbb{E}\left[ x_0 x_0^\top \right] \right) \left\| \sum_{t=0}^{T-2} \left( \mathcal{G}_t'\left( S_{t+1}' \right) - \mathcal{G}_t\left( S_{t+1} \right) \right) \right\|.$$

We denote $\mathcal{G}_d := \sum_{t=0}^{T-2}\left(\mathcal{G}_t'\left(S_{t+1}'\right) - \mathcal{G}_t\left(S_{t+1}\right)\right)$, then

$$
\begin{aligned}
\|\mathcal{G}_d\| \leq & \sum_{t=0}^{T-2}\Big\|\mathcal{G}_t'\left(Q_{t+1} + (K_{t+1}')^\top R_{t+1}K_{t+1}'\right) - \mathcal{G}_t\left(Q_{t+1} + (K_{t+1}')^\top R_{t+1}K_{t+1}'\right) - \\
& \mathcal{G}_t \circ \left(K_{t+1}^\top R_{t+1}K_{t+1} - (K_{t+1}')^\top R_{t+1}K_{t+1}'\right)\Big\| \\
\leq & \frac{\rho^{2T}-1}{\rho^2-1}\left((2\rho+1)\|B\|\sum_{t=0}^{T-2}\|K_t - K_t'\|\right)\left(\sum_{t=1}^{T-1}\|Q_t + (K_t')^\top R_t K_t'\|\right) \\
& + \sum_{t=0}^{T-2}\|\mathcal{G}_t\|\left\|(K_{t+1}')^\top R_{t+1}K_{t+1}' - K_{t+1}^\top R_{t+1}K_{t+1}\right\| \\
\leq & \frac{\rho^{2T}-1}{\rho^2-1}\left((2\rho+1)\|B\|\sum_{t=0}^{T-2}\|K_t - K_t'\|\right)\left(\sum_{t=1}^{T-1}\|Q_t + (K_t')^\top R_t K_t' - K_t^\top R_t K_t + K_t^\top R_t K_t\|\right) \\
& + \frac{\rho^2(\rho^{2(T-1)}-1)}{\rho^2-1}\sum_{t=1}^{T-1}\left\|(K_t')^\top R_t K_t' - K_t^\top R_t K_t\right\| \\
\leq & \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\left\|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right\|\left(\|\!|\boldsymbol{Q}|\!\| + \|\!|\boldsymbol{K}|\!\|^2\,\|\!|\boldsymbol{R}|\!\|\right) \\
& + \left(\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\left\|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right\| + \frac{\rho^2(\rho^{2(T-1)}-1)}{\rho^2-1}\right)\sum_{t=1}^{T-1}\left\|(K_t')^\top R_t K_t' - K_t^\top R_t K_t\right\|.
\end{aligned}
\tag{B.16}
$$

where the second inequality holds by Lemma 3.13 and (B.10), and the third inequality holds by (B.8). For the first term in $(III)$, we have

$$
\begin{aligned}
& \sum_{t=0}^{T-2}\mathrm{Tr}\left(\sum_{s=1}^{t} D_{t,s}'W(D_{t,s}')^\top S_{t+1}' - D_{t,s}WD_{t,s}^\top S_{t+1}\right) \\
= & \sum_{t=0}^{T-2}\mathrm{Tr}\left(\sum_{s=1}^{t} D_{t,s}'W(D_{t,s}')^\top (S_{t+1}' - S_{t+1}) + (D_{t,s}'W(D_{t,s}')^\top - D_{t,s}WD_{t,s}^\top)S_{t+1}\right) \\
\leq & \left(\sum_{t=0}^{T-2}\sum_{s=1}^{t}\mathrm{Tr}(W)\|D_{t,s}'\|^2\right)\left\|\sum_{t=1}^{T-1}(K_t')^\top R_t K_t' - K_t^\top R_t K_t\right\| \\
& + \sum_{t=0}^{T-2}\left\|\sum_{s=1}^{t} D_{t,s}'W(D_{t,s}')^\top - D_{t,s}WD_{t,s}^\top\right\|\left(\sum_{t=1}^{T-1}\mathrm{Tr}(Q_t) + \|K_t\|^2\,\mathrm{Tr}(R_t)\right) \\
\leq & \mathrm{Tr}(W)\frac{(T-1)(\rho^{2(T-1)}-1)}{\rho^2-1}\left\|\sum_{t=1}^{T-1}(K_t')^\top R_t K_t' - K_t^\top R_t K_t\right\| \\
& + T\frac{(\rho^{2T}-1)}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,\left\|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right\|\left(\mathrm{Tr}\left(\sum_{t=1}^{T-1}Q_t\right) + \|\!|\boldsymbol{K}|\!\|^2\,\mathrm{Tr}\left(\sum_{t=1}^{T-1}R_t\right)\right),
\end{aligned}
$$

where the last step holds by (3.20). The second term in $(III)$ is bounded by

$$
\sum_{t=0}^{T-2}\mathrm{Tr}\left(W(S_{t+1}' - S_{t+1})\right) \leq \mathrm{Tr}(W)\sum_{t=1}^{T-1}\left\|(K_t')^\top R_t K_t' - K_t^\top R_t K_t\right\|.
$$

Similarly, by (3.20) and (B.10), $(IV)$ is bounded by

$$
\begin{aligned}
(IV) \quad &\leq \quad \mathrm{Tr}(\mathbb{E}[x_0x_0^\top])\sum_{t=0}^{T-1}\left\|(\mathcal{G}_t'-\mathcal{G}_t)(Q_T)\right\| + \mathrm{Tr}\left(\sum_{s=1}^{T-1} D_{T-1,s}'W(D_{T-1,s}')^\top Q_T - D_{T-1,s}WD_{T-1,s}^\top Q_T\right) \\
&\leq \quad \mathrm{Tr}(\mathbb{E}[x_0x_0^\top])\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\,\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\| + \mathrm{Tr}(Q_T)\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\|.
\end{aligned}
$$

Now we bound the term $\sum_{t=1}^{T-1}\left\|(K_t')^\top R_tK_t' - K_t^\top R_tK_t\right\|$, which appears several times in previous inequalities:

$$
\begin{aligned}
\sum_{t=1}^{T-1}\left\|(K_t')^\top R_tK_t' - K_t^\top R_tK_t\right\| &= \sum_{t=1}^{T-1}\left\|(K_t'-K_t+K_t)^\top R_t(K_t'-K_t+K_t) - K_t^\top R_tK_t\right\| \\
&\leq \sum_{t=1}^{T-1}\|K_t'-K_t\|^2\|R_t\| + 2\|K_t\|\|R_t\|\|K_t'-K_t\| \\
&\leq 3|||\boldsymbol{K}|||\;|||\boldsymbol{R}|||\;\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\|.
\end{aligned}
$$

The last step holds since $\|K_t'-K_t\| \leq \|K_t\|$ by assumption.

Therefore,

$$
\begin{aligned}
|C(\boldsymbol{K}')-C(\boldsymbol{K})| \leq \mathrm{Tr}(\mathbb{E}[x_0x_0^\top])\Big\{&3|||\boldsymbol{K}|||\|R_0\|\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\| + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\,\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\| \\
&+ \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\|\left(|||\boldsymbol{Q}||| + |||\boldsymbol{K}|||^2\,|||\boldsymbol{R}|||\right) \\
&+ \left(\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\| + \frac{\rho^2(1-\rho^{2(T-1)})}{\rho^2-1}\right)3|||\boldsymbol{K}|||\,|||\boldsymbol{R}|||\,\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\|\Big\} \\
&+ 3\mathrm{Tr}(W)\Big(\frac{(T-1)(\rho^{2(T-1)}-1)}{\rho^2-1}+1\Big)|||\boldsymbol{K}|||\,|||\boldsymbol{R}|||\,\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\| \\
&+ \left(T\frac{(\rho^{2T}-1)}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\|\right)\left(\mathrm{Tr}\left(\sum_{t=1}^{T-1}Q_t\right) + |||\boldsymbol{K}|||^2\,\mathrm{Tr}\left(\sum_{t=1}^{T-1}R_t\right)\right) \\
&+ \mathrm{Tr}(Q_T)\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\|.
\end{aligned}
$$

By (3.27), Lemma 3.8, and Lemma 3.16, $\rho$ is bounded above by polynomials in $\|A\|$, $\|B\|$, $|||\boldsymbol{R}|||$, $\frac{1}{\sigma_{\boldsymbol{X}}}$, $\frac{1}{\sigma_{\boldsymbol{R}}}$ and $C(\boldsymbol{K})$, or a constant $1+\xi$. Therefore, we rewrite the above inequality by

$$
|C(\boldsymbol{K}')-C(\boldsymbol{K})| \leq h_{CK}\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\| + h_{CK}'\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\|^2, \tag{B.17}
$$

where $h_{CK} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ and $h_{CK}' \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ are polynomials in $C(\boldsymbol{K})$ and model parameters. Given assumption (4.5), we have $|||\boldsymbol{K}'-\boldsymbol{K}||| \leq 1$ and hence

$$
\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\| \geq \left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\|^2.
$$

Define $h_{cost} = h_{CK} + h_{CK}'$, then (B.17) gives

$$
|C(\boldsymbol{K}')-C(\boldsymbol{K})| \leq h_{cost}\left\|\left\|\boldsymbol{K}'-\boldsymbol{K}\right\|\right\|,
$$

with $h_{cost} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$. □

**Proof of Lemma 4.7.** Recall $\nabla_t C(\boldsymbol{K}) = 2E_t\Sigma_t$ and $W = \mathbb{E}\left[w_t w_t^\top\right] = \widetilde{W}\mathbb{E}\left[v_t v_t^\top\right]\widetilde{W}^\top$. We have,

$$\|\nabla_t C(\boldsymbol{K}') - \nabla_t C(\boldsymbol{K})\| = \|2E_t'\Sigma_t' - 2E_t\Sigma_t\| \leq 2\|E_t' - E_t\|\|\Sigma_t'\| + 2\|E_t\|\|\Sigma_t' - \Sigma_t\|, \tag{B.18}$$

For the second term, by Lemma 3.6 and Cauchy-Schwarz inequality,

$$\|E_t\| \leq \sum_{t=0}^{T-1}\|E_t\| \leq \sum_{t=0}^{T-1}\sqrt{\mathrm{Tr}(E_t^\top E_t)} \leq \sqrt{T \cdot \frac{\max_t \|R_t + B^\top P_{t+1}B\|}{\underline{\sigma}_{\boldsymbol{X}}}(C(\boldsymbol{K}) - C(\boldsymbol{K}^*))}. \tag{B.19}$$

By (B.9) and direct calculation, we have

$$\|(\mathcal{G}_{t+1}' - \mathcal{G}_{t+1})(\Sigma_0)\| \leq \rho^{2(t+1)}\left(\sum_{s=0}^{t+1}\|\mathcal{F}_{K_s'} - \mathcal{F}_{K_s}\|\,\|\Sigma_0\|\right).$$

By (B.10) and (3.20), for $t = 1, 2, \cdots, T-1$,

$$\begin{aligned}\|\Sigma_t' - \Sigma_t\| &\leq \|(\mathcal{G}_t' - \mathcal{G}_t)(\Sigma_0)\| + \left\|\sum_{s=0}^{t-1} D_{t-1,s}WD_{t-1,s}^\top - D_{t-1,s}'W(D_{t-1,s}')^\top\right\| \\ &\leq \rho^{2t}(2\rho+1)\|B\|\,\|\Sigma_0\|\,\left|\!\left|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right|\!\right| \; + \frac{(\rho^{2T}-1)}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,\left|\!\left|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right|\!\right|.\end{aligned} \tag{B.20}$$

Therefore the second term in (B.18) is bounded by the product of (B.19) and (B.20).

Next we bound the first term in (B.18). Similar to (B.15), $\|\Sigma_t'\| \leq \|\sum_{t=0}^{T}\Sigma_t'\| = \|\Sigma_{\boldsymbol{K}'}\| \leq \|\Sigma_{\boldsymbol{K}}' - \Sigma_{\boldsymbol{K}}\| + \|\Sigma_{\boldsymbol{K}}\| \leq \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + \|\Sigma_{\boldsymbol{K}}\|$. For $\|E_t' - E_t\|$, we first need a bound on $\|P_t' - P_t\|$. Since $P_0 = S_0 + \sum_{t=0}^{T-2}\mathcal{G}_t(S_{t+1}) + \mathcal{G}_{T-1}(Q_T)$, by (B.16), we have

$$\begin{aligned}\|P_t' - P_t\| \leq \|P_0' - P_0\| &\leq 3\|K_0\|\|R_0\|\|K_0' - K_0\| + \|\mathcal{G}_d\| + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\left(\sum_{t=0}^{T-1}\|K_t - K_t'\|\right) \\ &\leq \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\left|\!\left|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right|\!\right|\left(|||\boldsymbol{Q}||| + |||\boldsymbol{K}|||^2|||\boldsymbol{R}|||\right) \\ &\quad + 3\left(1 + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\left|\!\left|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right|\!\right| + \frac{\rho^2(1-\rho^{2(T-1)})}{\rho^2-1}\right)\cdot|||\boldsymbol{K}|||\;|||\boldsymbol{R}|||\;\left|\!\left|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right|\!\right| \\ &\quad + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\left|\!\left|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right|\!\right|.\end{aligned} \tag{B.21}$$

Thus,

$$\begin{aligned}\left\|E_t' - E_t\right\| &= \left\|R_t(K_t' - K_t) - B^\top(P_{t+1}' - P_{t+1})A + B^\top(P_{t+1}' - P_{t+1})BK_t' + B^\top P_{t+1}B(K_t' - K_t)\right\| \\ &\leq \left(\|R_t\| + \|B\|^2\|P_0\|\right)\left|\!\left|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right|\!\right| + \|B\|\,\|P_0' - P_0\|\,\|A\| + 2\|B\|^2\|P_0' - P_0\||||\boldsymbol{K}|||.\end{aligned}$$

Given the bound on $|||\boldsymbol{K}||| = \sum_{t=0}^{T-1}\|K_t\|$ in Lemma 3.16 and the bound on $\|P_t\|$ in Lemma 3.8, all the terms in (B.18) can be bounded by polynomials of related parameters multiplied by $|||\boldsymbol{K}' - \boldsymbol{K}|||$ and $|||\boldsymbol{K}' - \boldsymbol{K}|||^2$. Similarly to the proof of Lemma 4.6, we have $|||\boldsymbol{K}' - \boldsymbol{K}||| \leq 1$ and

$$\|\nabla_t C(\boldsymbol{K}') - \nabla_t C(\boldsymbol{K})\| \leq h_{grad}\left|\!\left|\!\left|\boldsymbol{K}' - \boldsymbol{K}\right|\!\right|\!\right|,$$

for some polynomial $h_{grad} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$.

$\square$

## B.5 Proofs in Section 5

*Proof of Proposition 5.2.* Denote $H_t := \begin{pmatrix} 1+\gamma k_t^1 & \gamma k_t^2 \\ k_t^1 & 1+k_t^2 \end{pmatrix}$. Since $H_t$ has two eigenvalues 1 and $\gamma k_t^1 + k_t^2 + 1$, $H_t$ is positive definite when $\gamma k_t^1 + k_t^2 > -1$ $(0 \leq t \leq T-1)$.

Then let us show the first claim by induction. Assume $\mathbb{E}[x_s x_s^\top]$ is positive definite for all $s \leq t$, then

$$\begin{aligned} \mathbb{E}[x_{t+1}x_{t+1}^\top] &= \mathbb{E}[((A-BK_t)x_t + w_t)\,((A-BK_t)x_t + w_t)^\top] \\ &= \mathbb{E}[(H_t x_t + w_t)\,(H_t x_t + w_t)^\top] \\ &= \mathbb{E}[H_t x_t x_t^\top H_t^\top + w_t w_t^\top + w_t w_t^\top + 2H_t x_t w_t^\top] \\ &= H_t \mathbb{E}[x_t x_t^\top] H_t^\top + \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

Hence $\mathbb{E}[x_{t+1}x_{t+1}^\top]$ is positive definite since $\mathbb{E}[x_t x_t^\top]$ is positive definite and $H_t$ is positive definite. Therefore $\underline{\sigma}_{\boldsymbol{X}} > 0$.

The second claim can be proved by backward induction. For $t = T$, $P_T^{\boldsymbol{K}} = Q_T$ is positive definite since $Q_T$ is positive definite. Assume $P_{t+1}^{\boldsymbol{K}}$ is positive definite for some $t+1$, then take any $z \in \mathbb{R}^d$ such that $z \neq 0$,

$$z^\top P_t^K z = z^\top Q_t\, z + z^\top K_t^\top R_t K_t z + z^\top H_t^\top P_{t+1}^{\boldsymbol{K}} H_t z > 0.$$

Note that $H_t$ is positive definite when $\gamma k_t^1 + k_t^2 > -1$ and $1 + \gamma k_t^1 > 0$. The last inequality holds since $Q_t$ and $H_t^\top P_{t+1}^{\boldsymbol{K}} H_t$ are positive definite, and $K_t^\top R_t K_t$ is positive semi-definite. Hence we have $P_t^{\boldsymbol{K}}$ positive definite for all $t = 0, 1, 2, \cdots, T$. $\square$