SUPPLEMENTARY MATERIALS: POLICY GRADIENT METHODS FOR THE NOISY LINEAR QUADRATIC REGULATOR OVER A FINITE HORIZON

BEN HAMBLY *, RENYUAN XU *, AND HUINING YANG *†

SM1. 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 [1], namely that the price change ΔS is given by

$$\Delta S = \gamma \,\mathrm{MFI} + \sigma\,\epsilon, \tag{SM1.1}$$

with $\mathrm{MFI} = M^b - M^s$ 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 γ and σ from the data.

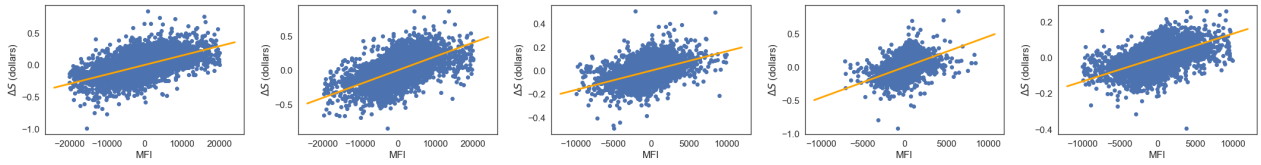

Fig. SM1: Relationship between MFI and ΔS. (Example (from left to right): AAP, FB, JPM, IBM and AAL, 10:00AM-11:00AM 01/01/2019-08/31/2019, $\Delta T = 1$min)

Temporary Price Impact. We assume the LOB has a flat shape with constant queue length l for the first few levels. Figure SM2 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

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

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

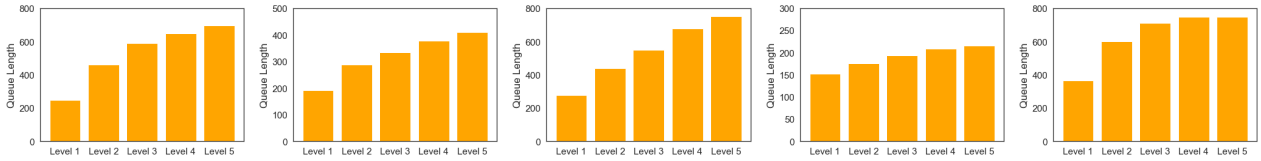

Fig. SM2: Average queue length (volume) of the first five levels on the limit buy side (Example (from left to right): AAP, FB, JPM, IBM and AAL, 10:00AM-11:00AM 01/01/2019-08/31/2019 with 5000 samples uniformly sampled with natural time clock in each trading day.)

Parameter Estimation. See the estimates for AAPL, FB, IBM, JPM, and AAL in Table SM1.

*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.

[1]https://lobsterdata.com/

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

Table SM1: Parameter estimation from NASDAQ ITCH Data (10:00AM-11:00PM 01/01/2019-08/31/2019).

SM2. Comparison between the Policy Gradient Method and Q-learning. The computational complexity of Q-learning is highly dependent on the size of the set of the (discrete) states and actions. Therefore Q-learning is typically less suited to problems with continuous and unbounded states and actions. In order to apply Q-learning for such problems, we need to discretize the continuous state and action space. Intuitively speaking, Q-learning suffers from low accuracy when the discretization scheme is less refined (see Figures SM3 and SM5). On the other hand, the computational complexity grows quadratically when increasing the level of granularity of discretization (see Figures SM4 and SM5).

To demonstrate this view point, we compare the performance of the Q-learning algorithm with the policy gradient method on a one-dimensional LQR problem with finite horizon as suggested by the reviewer. (We would expect the deep Q-learning algorithm and the deep policy gradient method to have similar comparison results.)

Q learning update. We initialize the Q table $\{q_t^{(0)}(x,u)\}_{x,u,t}$ with all zeros. In the i-th iteration, we update the Q table for $t = 0, 1, \cdots, T-1$,

$$q_t^{(i)}(x,u) = (1-\tilde{\eta})\, q_t^{(i-1)}(x,u) + \tilde{\eta} \left[c_t(x,u) + \min_{u'} \; q_{t+1}^{(i)}(x',u') \right], \tag{SM2.1}$$

with terminal condition $q_T^{(i)}(x,u) = x^2 Q_T$. Here $c_t(x,u) = x^2 Q_t + u^2 R_t$ is the instantaneous cost at time t; x' is the next state simulated from the system when the agent takes an action u in state x at time t; and $\tilde{\eta} \in (0,1)$ is the learning rate.

Model set-up. We set $d = k = 1$, $T = 5$, $A = 1.0$, $B = 0.2$, $Q_t = 0.2$ for $t = 0, 1, 2, 3, 4$, $Q_T = 0.4$, $R_t = 0.1(t+1)$ for $t = 0, 1, \cdots, 4$, $w_t \sim \mathcal{N}(0, 0.1)$, and $x_0 \sim \mathcal{N}(0, 0.1)$.

Parameter set-up. To perform Q-learning, we uniformly partition the states and actions in $[-1, 1]$. We set the learning rate for Q-learning as $\tilde{\eta} = 0.1$. For the policy gradient method, we set the learning rate $\eta = 0.2$ and the number of trajectories in the zero-th optimization as $m = 50$.

Conclusion. From Figure SM5, we observe that

- For LQR with finite horizon, the policy gradient method outperforms Q-learning algorithms (with the size of actions and states varying from 10 to 100) in terms of both sample efficiency and accuracy.
- When increasing the size of the states and actions from 10 to 100, the accuracy of the Q-learning algorithm improves, however, it requires many more samples to converge.

To conclude, Q-learning is less suited to handling decision-making problems with continuous and unbounded states and actions. More advanced approximation techniques may be needed in this case [2].

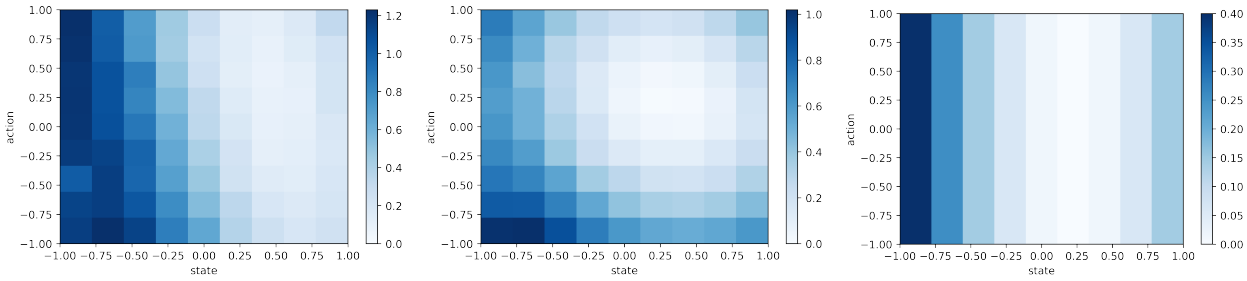

Fig. SM3: Q tables with 10 states and 10 actions: $q_0(s,a)$, $q_4(s,a)$ and $q_5(s,a)$ (from left to right).

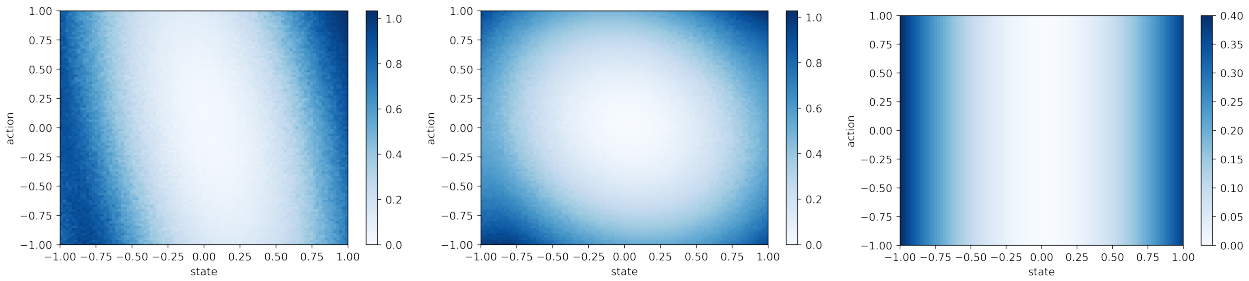

Fig. SM4: Q tables with 100 states and 100 actions: $q_0(s,a)$, $q_4(s,a)$ and $q_5(s,a)$ (from left to right).

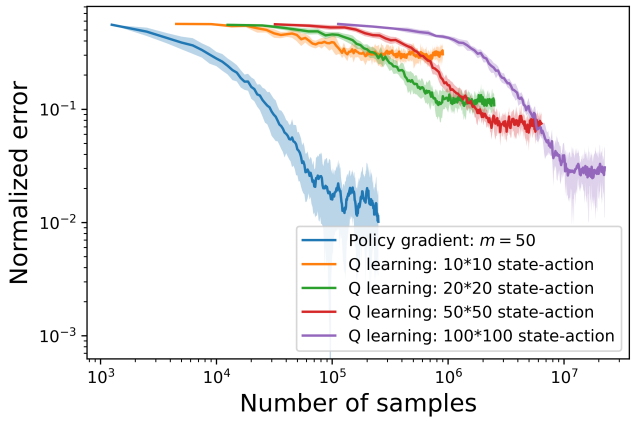

Fig. SM5: Comparison between Q-learning and the policy gradient method (log-log scale). (Average across 10 scenarios.)

SM3. Proofs of Technical Results. We now give the proofs that were omitted in the text.

SM3.1. 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 $\underline{\sigma}_{\boldsymbol{X}} = \min(\mathbb{E}[x_0 x_0^\top], \sigma_{\min}(W))$. ∎

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 $z \neq 0$,

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

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

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_τ 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 SM3.1. *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*

$$V_{\boldsymbol{K}'}(x,0) - V_{\boldsymbol{K}}(x,0) = \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x'_t,u'_t,t)\right], \tag{SM3.1}$$

and $A_{\boldsymbol{K}}(x,-K'_\tau x,\tau) = 2x^\top(K'_\tau-K_\tau)^\top E_\tau x + x^\top(K'_\tau-K_\tau)^\top(R_\tau+B^\top P_{\tau+1}B)(K'_\tau-K_\tau)x$, where E_τ 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 $u'_t = -K'_t x'_t$, $x'_{t+1} = Ax'_t + Bu'_t + w_t$, $x'_0 = x$.
Therefore,

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

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

(SM3.2)

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

Proof of Lemma 3.6. First for any K'_τ, from (SM3.2),

$$\begin{aligned} A_{\boldsymbol{K}}(x, -K'_\tau x, \tau) &= Q_{\boldsymbol{K}}(x, -K'_\tau x, \tau) - V_{\boldsymbol{K}}(x, \tau) \\ &= 2\,\mathrm{Tr}(xx^\top (K'_\tau - K_\tau)^\top E_\tau) + \mathrm{Tr}(xx^\top (K'_\tau - K_\tau)^\top (R_\tau + B^\top P_{\tau+1} B)(K'_\tau - K_\tau)) \\ &= \mathrm{Tr}\left(xx^\top (K'_\tau - K_\tau + (R_\tau + B^\top P_{\tau+1} B)^{-1} E_\tau)^\top (R_\tau + B^\top P_{\tau+1} B)\right. \\ &\quad \left.(K'_\tau - K_\tau + (R_\tau + B^\top P_{\tau+1} B)^{-1} E_\tau)\right) - \mathrm{Tr}(xx^\top E_\tau^\top (R_\tau + B^\top P_{\tau+1} B)^{-1} E_\tau) \\ &\geq -\mathrm{Tr}(xx^\top E_\tau^\top (R_\tau + B^\top P_{\tau+1} B)^{-1} E_\tau), \end{aligned} \tag{SM3.3}$$

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

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

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 (SM3.3). Using $C(\boldsymbol{K}^*) \leq C(\boldsymbol{K}')$

$$\begin{aligned} C(\boldsymbol{K}) - C(\boldsymbol{K}^*) \geq C(\boldsymbol{K}) - C(\boldsymbol{K}') &= -\mathbb{E} \sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x'_t, u'_t, t) = \mathbb{E} \sum_{t=0}^{T-1} \mathrm{Tr}(x'_t (x'_t)^\top E_t^\top (R_t + B^\top P_{t+1} B)^{-1} E_t) \\ &\geq \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t + B^\top P_{t+1} B\|} \mathrm{Tr}(E_t^\top E_t) \end{aligned} \tag{SM3.4}$$

□

Proof of Lemma 3.7. By lemma SM3.1 we have

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

□

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

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

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

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

$$\Sigma_t = \mathcal{G}_{t-1}(\Sigma_0) + \sum_{s=1}^{t-1} D_{t-1,s} W D_{t-1,s}^\top + W,\ \forall\, t = 2, 3, \cdots, T. \tag{SM3.5}$$

When $t = 2$,

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

which satisfies (SM3.5). Assume (SM3.5) holds for $t \le k$. Then for $t = k + 1$,

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

Therefore (SM3.5) holds, $\forall\, t = 1, 2, \cdots, T$. Finally,

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

∎

SM3.2. Proofs in Section 3.2.

Proof of Lemma 3.13. By direct calculation,

$$\|\mathcal{G}_t\| \le \rho^{2(t+1)}, \quad \text{and} \quad \|\mathcal{G}_t'\| \le \rho^{2(t+1)}. \tag{SM3.6}$$

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 $t \ge 0$,

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

Therefore,

$$\|(\mathcal{G}_{t+1}' - \mathcal{G}_{t+1})(\Sigma)\| \le \rho^2\|(\mathcal{G}_t' - \mathcal{G}_t)(\Sigma)\| + \rho^{2(t+1)}\|\mathcal{F}_{t+1}' - \mathcal{F}_{t+1}\|\|\Sigma\|. \tag{SM3.7}$$

Summing (SM3.7) up for $t \in \{1, 2, \cdots, T-2\}$ with $\|\mathcal{G}_0' - \mathcal{G}_0\| = \|\mathcal{F}_0' - \mathcal{F}_0\|$, we have

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

∎

SM3.3. Proofs in Section 3.3.

Proof of Lemma 3.15. Given (3.22) and condition (3.23), we have $\|K_t' - K_t\| = \eta\|\nabla_t C(\boldsymbol{K})\| \le \frac{\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{2C(\boldsymbol{K})\|B\|}$. Therefore,

$$\|B\|\|K_t' - K_t\| \le \frac{\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{2C(\boldsymbol{K})} \le \frac{1}{2}. \tag{SM3.8}$$

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

$$\sum_{t=0}^{T-1}\|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\| \le (2\rho + 1)\|B\|\left(\sum_{t=0}^{T-1}\|K_t - K_t'\|\right). \tag{SM3.9}$$

By Lemmas 3.5 and 3.7,

(SM3.10)
$$\begin{aligned}
C(\boldsymbol{K}') - C(\boldsymbol{K}) &= \sum_{t=0}^{T-1} \Big[2\operatorname{Tr}\Big(\Sigma_t'(K_t'-K_t)^\top E_t\Big) + \operatorname{Tr}\Big(\Sigma_t'(K_t'-K_t)^\top (R_t + B^\top P_{t+1}B)(K_t'-K_t)\Big)\Big] \\
&= \sum_{t=0}^{T-1} \Big[-4\eta \operatorname{Tr}\Big(\Sigma_t'\Sigma_t E_t^\top E_t\Big) + 4\eta^2 \operatorname{Tr}\Big(\Sigma_t'\Sigma_t E_t^\top (R_t + B^\top P_{t+1}B) E_t \Sigma_t\Big)\Big] \\
&= \sum_{t=0}^{T-1} \Big[-4\eta \operatorname{Tr}\Big((\Sigma_t' - \Sigma_t + \Sigma_t)\Sigma_t E_t^\top E_t\Big) + 4\eta^2 \operatorname{Tr}\Big(\Sigma_t'\Sigma_t E_t^\top (R_t + B^\top P_{t+1}B) E_t \Sigma_t\Big)\Big] \\
&\le \sum_{t=0}^{T-1} \Big[-4\eta \operatorname{Tr}\Big(\Sigma_t E_t^\top E_t \Sigma_t\Big) + 4\eta \operatorname{Tr}((\Sigma_t' - \Sigma_t)\Sigma_t E_t^\top E_t \Sigma_t \Sigma_t^{-1}) \\
&\qquad + 4\eta^2 \operatorname{Tr}\Big(\Sigma_t'\Sigma_t E_t^\top (R_t + B^\top P_{t+1}B) E_t \Sigma_t\Big)\Big] \\
&\le \sum_{t=0}^{T-1} \Big[-4\eta \operatorname{Tr}\Big(\Sigma_t E_t^\top E_t \Sigma_t\Big) + 4\eta \frac{\|\Sigma_t' - \Sigma_t\|}{\sigma_{\min}(\Sigma_t)} \operatorname{Tr}\Big(\Sigma_t E_t^\top E_t \Sigma_t\Big) \\
&\qquad + 4\eta^2 \|\Sigma_t'(R_t + B^\top P_{t+1}B)\| \operatorname{Tr}\Big(\Sigma_t E_t^\top E_t \Sigma_t\Big)\Big] \\
&\le -\eta\Big(1 - \frac{\sum_{t=0}^{T-1}\|\Sigma_t' - \Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}} - \eta\|\Sigma_{\boldsymbol{K}'}\| \sum_{t=0}^{T-1} \|R_t + B^\top P_{t+1}B\|\Big) \sum_{t=0}^{T-1} \Big[\operatorname{Tr}(\nabla_t C(\boldsymbol{K})^\top \nabla_t C(\boldsymbol{K}))\Big].
\end{aligned}$$

By Lemma 3.6, we have

(SM3.11)
$$C(\boldsymbol{K}') - C(\boldsymbol{K}) \le -\eta\Big(1 - \frac{\sum_{t=0}^{T-1}\|\Sigma_t' - \Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}} - \eta\|\Sigma_{\boldsymbol{K}'}\| \sum_{t=0}^{T-1} \|R_t + B^\top P_{t+1}B\|\Big)\Big(\frac{4\,\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big)$$

provided that

(SM3.12)
$$1 - \frac{\sum_{t=0}^{T-1}\|\Sigma_t' - \Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}} - \eta\|\Sigma_{\boldsymbol{K}'}\| \sum_{t=0}^{T-1} \|R_t + B^\top P_{t+1}B\| > 0.$$

By (3.21), (3.22), and (SM3.8),

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

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

(SM3.13)
$$\eta(2\rho+1)\|B\| \sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\| \le \eta(2\rho+1)\|B\|\Big(T \cdot \max_t\{\|\nabla_t C(\boldsymbol{K})\|\}\Big) \le \frac{(\rho^2-1)\,\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{2(\rho^{2T}-1)(C(\boldsymbol{K}) + \underline{\sigma}_{\boldsymbol{Q}}\,T\|W\|)}.$$

Then, by Corollary 3.14 and (SM3.9),

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

where the last step holds by (SM3.13). Therefore, the bound of $\|\Sigma_{\boldsymbol{K}'}\|$ in (SM3.12) is given by

(SM3.14)
$$\|\Sigma_{\boldsymbol{K}'}\| \le \|\Sigma_{\boldsymbol{K}'} - \Sigma_{\boldsymbol{K}}\| + \|\Sigma_{\boldsymbol{K}}\| \le \frac{1}{2}\,\underline{\sigma}_{\boldsymbol{X}} + \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} \le \frac{1}{2}\|\Sigma_{\boldsymbol{K}'}\| + \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}},$$

which indicates that $\|\Sigma_{\boldsymbol{K}'}\| \leq \frac{2C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$. Therefore, (SM3.12) 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,$C(\boldsymbol{K}') - C(\boldsymbol{K}) \leq -\frac{\eta}{2}\left(\frac{4\underline{\sigma}_{\boldsymbol{X}}{}^2 \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\right)\Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big)$, and

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

□

SM3.4. Proofs in Section 4. Before proceeding to the proof of Theorem 4.5, we show two important Lemmas which provide the intermediate steps. We first show the optimality condition for the projection operator in Lemma SM3.2. We then show the one-step convergence result in Lemma SM3.3.

LEMMA SM3.2 (Optimality Condition). *Fix a policy matrix $\boldsymbol{L}^1$ and write $\boldsymbol{L}^* = \Pi_{\mathcal{S}}(\boldsymbol{L}^1)$. Then for any $\boldsymbol{L}^0 \in \mathcal{S}$, we have*

$$
\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^0 - L_t^*)(L_t^* - L_t^1)^\top\right) \geq 0. \tag{SM3.15}
$$

Proof of Lemma SM3.2. We show condition (SM3.15) by contradiction. Assume condition (SM3.15) does not hold, then there exist some $\boldsymbol{L}^3 \in \mathcal{S}$ and some constant $b > 0$ such that

$$
\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^* - L_t^1)^\top\right) = -b < 0. \tag{SM3.16}
$$

Let

$$
M = 1 + \frac{\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right)}{-2\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^* - L_t^1)^\top\right)} = 1 + \frac{\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right)}{2b} > 1,
$$

and take

$$
\overline{\boldsymbol{L}} = \frac{1}{M}\boldsymbol{L}^3 + \left(1 - \frac{1}{M}\right)\boldsymbol{L}^*.
$$

By the convexity of $\mathcal{S}$, we have $\overline{\boldsymbol{L}} \in \mathcal{S}$. Hence from definition (4.3), we have

$$
\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^* - L_t^1)(L_t^* - L_t^1)^\top\right) \leq \sum_{t=0}^{T-1} \mathrm{Tr}\left((\overline{L}_t - L_t^1)(\overline{L}_t - L_t^1)^\top\right). \tag{SM3.17}
$$

On the other hand,

$$\begin{aligned}
&\sum_{t=0}^{T-1} \mathrm{Tr}\left((\overline{L}_t - L_t^1)(\overline{L}_t - L_t^1)^\top\right) \\
&= \sum_{t=0}^{T-1} \mathrm{Tr}\left(\left(\frac{1}{M}(L_t^3 - L_t^*) + (L_t^* - L_t^1)\right)\left(\frac{1}{M}(L_t^3 - L_t^*) + (L_t^* - L_t^1)\right)^\top\right) \\
&= \sum_{t=0}^{T-1} \frac{1}{M^2} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right) + \sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^* - L_t^1)(L_t^* - L_t^1)^\top\right) \\
&\qquad + \sum_{t=0}^{T-1} \frac{2}{M} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^* - L_t^1)^\top\right). \qquad \text{(SM3.18)}
\end{aligned}$$

By the definition of M we have,

$$\begin{aligned}
&\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right) + 2M \sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^* - L_t^1)^\top\right) \\
&= \sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right) - 2Mb \\
&= \sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right) - 2b - \sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right) = -2b < 0.
\end{aligned}$$

Thus substituting this in (SM3.18) contradicts (SM3.17) which completes the proof.

∎

LEMMA SM3.3. *Assume Assumption 2.1 holds,* $\underline{\sigma}_{\boldsymbol{X}} > 0$, $\boldsymbol{K} \in \mathcal{S}$ *and that*

$$K_t' = K_t - \eta \nabla_t C(\boldsymbol{K}), \qquad \text{where} \qquad \text{(SM3.19)}$$

$$\eta \leq \min\left\{C_1, C_2\right\}, \qquad \text{with} \qquad \text{(SM3.20)}$$

$$C_1 = \frac{(\rho^2 - 1)\,\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{4d\,T^2\sqrt{d+k}(\rho^{2T} - 1)(2\rho + 1)(C(\boldsymbol{K}) + \underline{\sigma}_{\boldsymbol{Q}}\,T\|W\|)\|B\|\max_t\{\|\nabla_t C(\boldsymbol{K})\|\}} \qquad \text{(SM3.21)}$$

$$C_2 = \frac{\underline{\sigma}_{\boldsymbol{Q}}}{8C(\boldsymbol{K})\sum_{t=0}^{T-1}\|R_t + B^\top P_{t+1} B\|}. \qquad \text{(SM3.22)}$$

Take

$$\widetilde{\boldsymbol{K}} = \Pi_{\mathcal{S}}\left(\boldsymbol{K}'\right), \qquad \text{(SM3.23)}$$

with $\boldsymbol{K}' = (K_0', \cdots, K_{T-1}')$ *and* K_t' *defined in* (SM3.19) $(t = 0, 1, \cdots, T-1)$. *Then we have*

$$2\eta \sum_{t=0}^{T} \|G_t(\boldsymbol{K})\|_F^2 = 2\eta \sum_{t=0}^{T} \mathrm{Tr}\left(G_t(\boldsymbol{K})^\top G_t(\boldsymbol{K})\right) \leq C(\boldsymbol{K}) - C(\widetilde{\boldsymbol{K}}).$$

Proof. By definition of $\widetilde{\boldsymbol{K}}$, and as $\boldsymbol{K} \in \mathcal{S}$, we have

$$\sum_{t=0}^{T-1} \mathrm{Tr}\left((\widetilde{K}_t - K_t')(\widetilde{K}_t - K_t')^\top\right) \leq \sum_{t=0}^{T-1} \mathrm{Tr}\left((K_t - K_t')(K_t - K_t')^\top\right). \qquad \text{(SM3.24)}$$

Take $\boldsymbol{L}^1 = \boldsymbol{K}'$ and $\boldsymbol{L}^0 = \boldsymbol{K}$ in Lemma SM3.2, we have

$$\sum_{t=0}^{T-1} \operatorname{Tr}\left((K_t - \widetilde{K}_t)(\widetilde{K}_t - K_t')^\top\right) \geq 0. \tag{SM3.25}$$

Combining (SM3.24) and (SM3.25) leads to

$$\sum_{t=0}^{T-1} \operatorname{Tr}\left((K_t - K_t')(K_t - \widetilde{K}_t)^\top\right) \geq \sum_{t=0}^{T-1} \operatorname{Tr}\left((K_t - \widetilde{K}_t)(K_t - \widetilde{K}_t)^\top\right). \tag{SM3.26}$$

Given the definition (SM3.23), we have $G(\boldsymbol{K}) = \frac{\widetilde{\boldsymbol{K}} - \boldsymbol{K}}{2\eta}$ and $G_t(\boldsymbol{K}) = \frac{\widetilde{K}_t - K_t}{2\eta}$. By Lemmas 3.5 and 3.7,

$$\begin{aligned} C(\widetilde{\boldsymbol{K}}) - C(\boldsymbol{K}) &= \sum_{t=0}^{T-1} \left[2\operatorname{Tr}\left(\widetilde{\Sigma}_t(\widetilde{K}_t - K_t)^\top E_t\right) + \operatorname{Tr}\left(\widetilde{\Sigma}_t(\widetilde{K}_t - K_t)^\top (R_t + B^\top P_{t+1} B)(\widetilde{K}_t - K_t)\right)\right] \\ &= \sum_{t=0}^{T-1} \Big[4\eta \operatorname{Tr}\left((\widetilde{\Sigma}_t - \Sigma_t)(G_t(\boldsymbol{K}))^\top E_t\right) + 4\eta \operatorname{Tr}\left(\Sigma_t (G_t(\boldsymbol{K}))^\top E_t\right) \\ &\quad + 4\eta^2 \operatorname{Tr}\left(\widetilde{\Sigma}_t (G_t(\boldsymbol{K}))^\top (R_t + B^\top P_{t+1} B) G_t(\boldsymbol{K})\right)\Big], \end{aligned} \tag{SM3.27}$$

with $\widetilde{\Sigma}_t := \mathbb{E}[\widetilde{x}_t \widetilde{x}_t^\top]$ and $\{\widetilde{x}_t\}_{t=0}^{T-1}$ is the trajectory under policy $\widetilde{\boldsymbol{K}}$.

First, we have

$$\begin{aligned} &\sum_{t=0}^{T-1} \operatorname{Tr}\left(\Sigma_t (G_t(\boldsymbol{K}))^\top E_t\right) = \sum_{t=0}^{T-1} \operatorname{Tr}\left((G_t(\boldsymbol{K}))^\top E_t \Sigma_t\right) = \frac{1}{4\eta^2} \sum_{t=0}^{T-1} \operatorname{Tr}((\widetilde{K}_t - K_t)^\top (K_t - K_t')) \\ &\leq -\frac{1}{4\eta^2} \sum_{t=0}^{T-1} \operatorname{Tr}((\widetilde{K}_t - K_t)(\widetilde{K}_t - K_t)^\top) = -\operatorname{Tr}((G_t(\boldsymbol{K}))^\top (G_t(\boldsymbol{K}))), \end{aligned}$$

in which the last inequality holds by (SM3.26).

Second given (SM3.19) and condition (SM3.21), we have

$$\|K_t' - K_t\| = \eta \|\nabla_t C(\boldsymbol{K})\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{2T\sqrt{d+k} C(\boldsymbol{K}) \|B\|}.$$

Therefore,

$$\begin{aligned} \sum_{t=0}^{T-1} \|B\| \|\tilde{K}_t - K_t\| &\leq \sum_{t=0}^{T-1} \|B\| \|\tilde{K}_t - K_t\|_F \leq \sum_{t=0}^{T-1} \|B\| \|K_t' - K_t\|_F \\ &\leq \sqrt{d+k} \|B\| \sum_{t=0}^{T-1} \|K_t' - K_t\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{2C(\boldsymbol{K})} \leq \frac{1}{2}. \end{aligned}$$

The second inequality holds by (SM3.26) and the last inequality holds since $\underline{\sigma}_{\boldsymbol{X}} \leq \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$ given by Lemma 3.8.

By (3.21),

$$\begin{aligned} &\sum_{t=0}^{T-1} \|\widetilde{\Sigma}_t - \Sigma_t\| \leq \frac{\rho^{2T} - 1}{\rho^2 - 1} \left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right) \left(2\rho \|B\| \, |||\boldsymbol{K} - \widetilde{\boldsymbol{K}}||| + \|B\|^2 \, |||\boldsymbol{K} - \widetilde{\boldsymbol{K}}|||^2\right) \\ &\leq \frac{\rho^{2T} - 1}{\rho^2 - 1} \left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right) \left(2(2\rho+1)\|B\| \sum_{t=0}^{T-1} \eta \|G_t(\boldsymbol{K})\|\right) \\ &\leq \frac{\underline{\sigma}_{\boldsymbol{X}}}{2dT^2 \sqrt{d+k} \cdot \max_t \|\nabla_t C(\boldsymbol{K})\|} \sum_{t=0}^{T-1} \|G_t(\boldsymbol{K})\|, \end{aligned} \tag{SM3.28}$$

where the last inequality holds by step size condition $\eta \le C_1$. Hence

$$
\begin{aligned}
(\text{SM3.27}) \le \sum_{t=0}^{T-1} \Bigg[& 2\eta \frac{d\|\widetilde{\Sigma}_t - \Sigma_t\|}{\sigma_{\min}(\Sigma_t)} \|G_t(\boldsymbol{K})\| \, \|\nabla_t C(\boldsymbol{K})\| - 4\eta \operatorname{Tr}((G_t(\boldsymbol{K}))^\top (G_t(\boldsymbol{K}))) \\
& + 4\eta^2 \|\Sigma_{\widetilde{\boldsymbol{K}}}\| \, \|R_t + B^\top P_{t+1} B\| \operatorname{Tr}\Big((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\Big) \Bigg] \\
\le & \, 2\eta \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \Big(\sum_{t=0}^{T-1} \|\widetilde{\Sigma}_t - \Sigma_t\| \Big) \Big(\sum_{t=0}^{T-1} \|G_t(\boldsymbol{K})\| \Big) \Big(\sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\| \Big) - 4\eta \sum_{t=0}^{T-1} \operatorname{Tr}((G_t(\boldsymbol{K}))^\top (G_t(\boldsymbol{K}))) \\
& + 4\eta^2 \sum_{t=0}^{T-1} \|\Sigma_{\widetilde{\boldsymbol{K}}}\| \, \|R_t + B^\top P_{t+1} B\| \operatorname{Tr}\Big((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\Big) \\
\le & \, \frac{\eta}{T} \Big(\sum_{t=0}^{T-1} \|G_t(\boldsymbol{K})\| \Big)^2 - 4\eta \sum_{t=0}^{T-1} \operatorname{Tr}((G_t(\boldsymbol{K}))^\top (G_t(\boldsymbol{K}))) \\
& + 4\eta^2 \sum_{t=0}^{T-1} \|\Sigma_{\widetilde{\boldsymbol{K}}}\| \, \|R_t + B^\top P_{t+1} B\| \operatorname{Tr}\Big((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\Big) \\
\le & \sum_{t=0}^{T-1} \Big[\eta \|G_t(\boldsymbol{K})\|^2 - 4\eta \operatorname{Tr}((G_t(\boldsymbol{K}))^\top (G_t(\boldsymbol{K}))) \\
& + 4\eta^2 \|\Sigma_{\widetilde{\boldsymbol{K}}}\| \, \|R_t + B^\top P_{t+1} B\| \operatorname{Tr}\Big((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\Big) \Big] \\
\le & \sum_{t=0}^{T-1} \eta \left[-3 + 4\eta \|\Sigma_{\widetilde{\boldsymbol{K}}}\| \, \|R_t + B^\top P_{t+1} B\| \right] \operatorname{Tr}\Big((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\Big),
\end{aligned}
$$

where the third inequality holds by (SM3.28) and the fourth inequality holds by Cauchy-Schwarz inequality. By (SM3.26) we have $\sqrt{d+k} \sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\| \ge \sum_{t=0}^{T-1} \|G_t(\boldsymbol{K})\|$ and thus (SM3.28) $\le \frac{\underline{\sigma}_{\boldsymbol{X}}}{2}$ and

$$
\|\Sigma_{\widetilde{\boldsymbol{K}}}\| \le \left\|\Sigma_{\widetilde{\boldsymbol{K}}} - \Sigma_{\boldsymbol{K}} + \Sigma_{\boldsymbol{K}}\right\| \le \frac{\underline{\sigma}_{\boldsymbol{X}}}{2} + \|\Sigma_{\boldsymbol{K}}\| \le \frac{\|\Sigma_{\widetilde{\boldsymbol{K}}}\|}{2} + \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}.
$$

Thus $\|\Sigma_{\widetilde{\boldsymbol{K}}}\| \le \frac{2C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$. Therefore when $\eta \le C_2$, we have

$$
(\text{SM3.27}) \le -2\eta \sum_{t=0}^{T} \operatorname{Tr}((G_t(\boldsymbol{K}))^\top (G_t(\boldsymbol{K}))).
$$

□

Proof of Theorem 4.5. The key step in this proof is Lemma SM3.3 and it suffices to show that the projected policy gradient method enjoys sublinear convergence rate in the setting of known paramaters. This is because moving from the analysis for the case of known parameters to that for the case of unknown parameters follows the same procedure of policy gradient descent (without projection). In particular, the zeroth order estimation of the gradient term $\eta \nabla_t C(\boldsymbol{K})$ in (SM3.19) is the same for policy gradient method and projected policy gradient method.

We now show that the projected policy gradient method with known parameters enjoys a sublinear convergence rate. Since the step size conditions (SM3.20)-(SM3.22) are independent of the term $G_t(\boldsymbol{K})$, the existence of η follows the analysis in Theorem 3.3. Hence when $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ is an appropriate polynomial in $\frac{1}{C(\boldsymbol{K}^0)+1}$ and model parameters, by Lemma SM3.3, we have for any $N \in \mathbb{N}^+$,

$$
\text{(SM3.29)} \qquad \sum_{n=1}^{N} \left(\sum_{t=0}^{T-1} \operatorname{Tr}((G_t^{\boldsymbol{K}^n})^\top (G_t^{\boldsymbol{K}^n})) \right) \le \frac{\sum_{n=1}^{N} C(\boldsymbol{K}^{n-1}) - C(\boldsymbol{K}^n)}{2\eta} \le \frac{C(\boldsymbol{K}^0) - C(\boldsymbol{K}^*)}{2\eta},
$$

Therefore $\left\{ \frac{1}{N} \sum_{n=0}^{N-1} \left(\sum_{t=0}^{T-1} \|G_t(\boldsymbol{K}^n)\|_F^2 \right) \right\}_{N \ge 1}$ converges at rate $\mathcal{O}\left(\frac{1}{N}\right)$, which thus completes the proof for the case of known parameters. □

Proof of Lemma 4.7. Under Assumption 4.3, we have $\mathbb{E}\left[x_0 x_0^\top\right] = \widetilde{W}_0 \mathbb{E}\left[z_0 z_0^\top\right] \widetilde{W}_0^\top$, and $\left\|\mathbb{E}\left[x_0 x_0^\top\right]\right\| \leq \sigma_0^2 \|\widetilde{W}_0\|^2$. 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 $t = 0, 1, \cdots, T-2$,

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

The last equality holds by (SM3.5). Therefore,

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

For the first term, $(I) \leq \mathrm{Tr}(\mathbb{E}[x_0 x_0^\top]) \|(K_0')^\top R_0 K_0' - K_0^\top R_0 K_0\|$. For the second term (II), since

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

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

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

$$\begin{aligned}\|\mathcal{G}_d\| \leq &\sum_{t=0}^{T-2} \Big\| \mathcal{G}_t'\left(Q_{t+1} + (K_{t+1}')^\top R_{t+1} K_{t+1}'\right) - \mathcal{G}_t\left(Q_{t+1} + (K_{t+1}')^\top R_{t+1} K_{t+1}'\right) - \\ &\mathcal{G}_t \circ \left(K_{t+1}^\top R_{t+1} K_{t+1} - (K_{t+1}')^\top R_{t+1} K_{t+1}'\right) \Big\| \\ \leq &\frac{\rho^{2T} - 1}{\rho^2 - 1} \left((2\rho + 1)\|B\| \sum_{t=0}^{T-2} \|K_t - K_t'\|\right) \left(\sum_{t=1}^{T-1} \|Q_t + (K_t')^\top R_t K_t'\|\right) \\ &+ \sum_{t=0}^{T-2} \|\mathcal{G}_t\| \left\|(K_{t+1}')^\top R_{t+1} K_{t+1}' - K_{t+1}^\top R_{t+1} K_{t+1}\right\| \\ \leq &\frac{\rho^{2T} - 1}{\rho^2 - 1} \left((2\rho + 1)\|B\| \sum_{t=0}^{T-2} \|K_t - K_t'\|\right) \left(\sum_{t=1}^{T-1} \|Q_t + (K_t')^\top R_t K_t' - K_t^\top R_t K_t + K_t^\top R_t K_t\|\right) \\ &+ \frac{\rho^2(\rho^{2(T-1)} - 1)}{\rho^2 - 1} \sum_{t=1}^{T-1} \left\|(K_t')^\top R_t K_t' - K_t^\top R_t K_t\right\| \\ \leq &\frac{\rho^{2T} - 1}{\rho^2 - 1} (2\rho + 1)\|B\| \,|||\boldsymbol{K}' - \boldsymbol{K}||| \left(|||\boldsymbol{Q}||| + |||\boldsymbol{K}|||^2 \,|||\boldsymbol{R}|||\right) \\ &+ \left(\frac{\rho^{2T} - 1}{\rho^2 - 1} (2\rho + 1)\|B\| \,|||\boldsymbol{K}' - \boldsymbol{K}||| + \frac{\rho^2(\rho^{2(T-1)} - 1)}{\rho^2 - 1}\right) \sum_{t=1}^{T-1} \left\|(K_t')^\top R_t K_t' - K_t^\top R_t K_t\right\|.\end{aligned} \tag{SM3.30}$$

where the second inequality holds by Lemma 3.13 and (SM3.9), and the third inequality holds by (SM3.6). For the first term in (III), we have

$$
\begin{aligned}
&\sum_{t=0}^{T-2} \operatorname{Tr}\left(\sum_{s=1}^{t} D'_{t,s} W (D'_{t,s})^\top S'_{t+1} - D_{t,s} W D_{t,s}^\top S_{t+1}\right) \\
&= \sum_{t=0}^{T-2} \operatorname{Tr}\left(\sum_{s=1}^{t} D'_{t,s} W (D'_{t,s})^\top (S'_{t+1} - S_{t+1}) + (D'_{t,s} W (D'_{t,s})^\top - D_{t,s} W D_{t,s}^\top) S_{t+1}\right) \\
&\le \Big(\sum_{t=0}^{T-2}\sum_{s=1}^{t} \operatorname{Tr}(W)\|D'_{t,s}\|^2\Big) \left\|\sum_{t=1}^{T-1} (K'_t)^\top R_t K'_t - K_t^\top R_t K_t\right\| \\
&\quad + \sum_{t=0}^{T-2} \left\|\sum_{s=1}^{t} D'_{t,s} W (D'_{t,s})^\top - D_{t,s} W D_{t,s}^\top\right\| \Big(\sum_{t=1}^{T-1} \operatorname{Tr}(Q_t) + \|K_t\|^2 \operatorname{Tr}(R_t)\Big) \\
&\le \operatorname{Tr}(W) \frac{(T-1)(\rho^{2(T-1)}-1)}{\rho^2-1} \left\|\sum_{t=1}^{T-1} (K'_t)^\top R_t K'_t - K_t^\top R_t K_t\right\| \\
&\quad + T\frac{(\rho^{2T}-1)}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||\left(\operatorname{Tr}\left(\sum_{t=1}^{T-1} Q_t\right) + |||\boldsymbol{K}|||^2 \operatorname{Tr}\left(\sum_{t=1}^{T-1} R_t\right)\right),
\end{aligned}
$$

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

$$
\sum_{t=0}^{T-2} \operatorname{Tr}\Big(W(S'_{t+1} - S_{t+1})\Big) \le \operatorname{Tr}(W) \sum_{t=1}^{T-1} \left\|(K'_t)^\top R_t K'_t - K_t^\top R_t K_t\right\|.
$$

Similarly, by (3.20) and (SM3.9), (IV) is bounded by

$$
\begin{aligned}
(IV) &\le \operatorname{Tr}(\mathbb{E}[x_0 x_0^\top]) \sum_{t=0}^{T-1} \left\|(\mathcal{G}'_t - \mathcal{G}_t)(Q_T)\right\| + \operatorname{Tr}\left(\sum_{s=1}^{T-1} D'_{T-1,s} W (D'_{T-1,s})^\top Q_T - D_{T-1,s} W D_{T-1,s}^\top Q_T\right) \\
&\le \operatorname{Tr}(\mathbb{E}[x_0 x_0^\top]) \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\,|||\boldsymbol{K}' - \boldsymbol{K}||| + \operatorname{Tr}(Q_T)\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||.
\end{aligned}
$$

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

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

The last step holds since $\|K'_t - K_t\| \le \|K_t\|$ by assumption.

Therefore,

$$
\begin{aligned}
|C(\boldsymbol{K}') - C(\boldsymbol{K})| \leq \mathrm{Tr}(\mathbb{E}[x_0 x_0^\top]) \Big\{ & 3|||\boldsymbol{K}|||\,|||R_0|||\,|||\boldsymbol{K}' - \boldsymbol{K}||| + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||Q_T\|\,|||\boldsymbol{K}' - \boldsymbol{K}||| \\
& + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||\left(|||\boldsymbol{Q}||| + |||\boldsymbol{K}|||^2\,|||\boldsymbol{R}|||\right) \\
& + \left(\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}' - \boldsymbol{K}||| + \frac{\rho^2(1-\rho^{2(T-1)})}{\rho^2-1}\right) 3|||\boldsymbol{K}|||\,|||\boldsymbol{R}|||\,|||\boldsymbol{K}' - \boldsymbol{K}||| \Big\} \\
& + 3\mathrm{Tr}(W)\Big(\frac{(T-1)(\rho^{2(T-1)}-1)}{\rho^2-1} + 1\Big)|||\boldsymbol{K}|||\,|||\boldsymbol{R}|||\,|||\boldsymbol{K}' - \boldsymbol{K}||| \\
& + \left(T\frac{(\rho^{2T}-1)}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||\right)\left(\mathrm{Tr}\left(\sum_{t=1}^{T-1} Q_t\right) + |||\boldsymbol{K}|||^2\,\mathrm{Tr}\left(\sum_{t=1}^{T-1} R_t\right)\right) \\
& + \mathrm{Tr}(Q_T)\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||.
\end{aligned}
$$

By (3.27), Lemma 3.8, and Lemma 3.16, ρ 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, we rewrite the above inequality by

$$
|C(\boldsymbol{K}') - C(\boldsymbol{K})| \leq h_{CK}|||\boldsymbol{K}' - \boldsymbol{K}||| + h'_{CK}|||\boldsymbol{K}' - \boldsymbol{K}|||^2, \tag{SM3.31}
$$

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 (4.7), we have $|||\boldsymbol{K}' - \boldsymbol{K}||| \leq 1$ and hence

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

Define $h_{cost} = h_{CK} + h'_{CK}$, then (SM3.31) gives

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

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

Proof of Lemma 4.8. 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,

$$
\|\nabla_t C(\boldsymbol{K}') - \nabla_t C(\boldsymbol{K})\| = \|2E'_t\Sigma'_t - 2E_t\Sigma_t\| \leq 2\|E'_t - E_t\|\|\Sigma'_t\| + 2\|E_t\|\|\Sigma'_t - \Sigma_t\|, \tag{SM3.32}
$$

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

$$
\|E_t\| \leq \sum_{t=0}^{T-1}\|E_t\| \leq \sum_{t=0}^{T-1}\sqrt{\mathrm{Tr}(E_t^\top E_t)} \leq \sqrt{T\cdot\frac{\max_t\|R_t + B^\top P_{t+1}B\|}{\underline{\sigma}_{\boldsymbol{X}}}\left(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\right)}. \tag{SM3.33}
$$

By (SM3.7) 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 (SM3.9) and (3.20), for $t = 1, 2, \cdots, T-1$,

$$
\begin{aligned}
\|\Sigma'_t - \Sigma_t\| &\leq \|(\mathcal{G}'_t - \mathcal{G}_t)(\Sigma_0)\| + \left\|\sum_{s=0}^{t-1} D_{t-1,s}WD_{t-1,s}^\top - D'_{t-1,s}W(D'_{t-1,s})^\top\right\| \\
&\leq \rho^{2t}(2\rho+1)\|B\|\,\|\Sigma_0\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||\ + \frac{(\rho^{2T}-1)}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||.
\end{aligned} \tag{SM3.34}
$$

Therefore the second term in (SM3.32) is bounded by the product of (SM3.33) and (SM3.34).

Next we bound the first term in (SM3.32). Similar to (SM3.14), $\|\Sigma'_t\| \le \|\sum_{t=0}^{T}\Sigma'_t\| = \|\Sigma_{\boldsymbol{K}'}\| \le \|\Sigma'_{\boldsymbol{K}} - \Sigma_{\boldsymbol{K}}\| + \|\Sigma_{\boldsymbol{K}}\| \le \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + \|\Sigma_{\boldsymbol{K}}\|$. 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 (SM3.30), we have

(SM3.35)

$$\begin{aligned}\|P'_t - P_t\| \le \|P'_0 - P_0\| &\le 3\|K_0\|\|R_0\|\|K'_0 - K_0\| + \|\mathcal{G}_d\| + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\left(\sum_{t=0}^{T-1}\|K_t - K'_t\|\right)\\ &\le \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||\left(|||\boldsymbol{Q}||| + |||\boldsymbol{K}|||^2|||\boldsymbol{R}|||\right)\\ &\quad + 3\left(1 + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}' - \boldsymbol{K}||| + \frac{\rho^2(1-\rho^{2(T-1)})}{\rho^2-1}\right)\cdot |||\boldsymbol{K}|||\,|||\boldsymbol{R}|||\,|||\boldsymbol{K}' - \boldsymbol{K}|||\\ &\quad + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\||||\boldsymbol{K}' - \boldsymbol{K}|||.\end{aligned}$$

Thus,

$$\begin{aligned}\|E'_t - E_t\| &= \left\|R_t(K'_t - K_t) - B^\top(P'_{t+1} - P_{t+1})A + B^\top(P'_{t+1} - P_{t+1})BK'_t + B^\top P_{t+1}B(K'_t - K_t)\right\|\\ &\le \left(\|R_t\| + \|B\|^2\|P_0\|\right)|||\boldsymbol{K}' - \boldsymbol{K}||| + \|B\|\,\|P'_0 - P_0\|\,\|A\| + 2\|B\|^2\|P'_0 - P_0\|||\boldsymbol{K}|||.\end{aligned}$$

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 (SM3.32) 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 4.7, we have $|||\boldsymbol{K}' - \boldsymbol{K}||| \le 1$ and

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

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

SM3.5. 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 $\gamma k_t^1 + k_t^2 > -1$ $(0 \le t \le T-1)$.

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

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

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 $z \ne 0$,

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

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 $t = 0, 1, 2, \cdots, T$. $\square$

References.

[1] R. Cont, A. Kukanov, and S. Stoikov, *The price impact of order book events*, Journal of Financial Econometrics, 12 (2014), pp. 47–88.

[2] S. Tu and B. Recht, *Least-squares temporal difference learning for the linear quadratic regulator*, in International Conference on Machine Learning, 2018, pp. 5005–5014.