Proceedings of Machine Learning Research vol 99:1–48, 2019 32nd Annual Conference on Learning Theory

The Gap Between Model-Based and Model-Free Methods on the Linear Quadratic Regulator: An Asymptotic Viewpoint

Stephen Tu STEPHENT@BERKELEY.EDU
Benjamin Recht BRECHT@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

The effectiveness of model-based versus model-free methods is a long-standing question in reinforcement learning (RL). Motivated by recent empirical success of RL on continuous control tasks, we study the sample complexity of popular model-based and model-free algorithms on the Linear Quadratic Regulator (LQR). We show that for policy evaluation, a simple model-based plugin method requires asymptotically less samples than the classical least-squares temporal difference (LSTD) estimator to reach the same quality of solution; the sample complexity gap between the two methods can be at least a factor of state dimension. For policy optimization, we study a simple family of problem instances and show that nominal (certainty equivalence principle) control also requires several factors of state and input dimension fewer samples than the policy gradient method to reach the same level of control performance on these instances. Furthermore, the gap persists even when employing baselines commonly used in practice. To the best of our knowledge, this is the first theoretical result which demonstrates a separation in the sample complexity between model-based and model-free methods on a continuous control task.

Keywords: Linear dynamical systems, linear quadratic regulator, least-squares temporal difference learning, policy gradients.

1. Introduction

The reinforcement learning (RL) community has been debating the relative merits of model-based and model-free methods for decades. This debate has become reinvigorated in the last few years due to the impressive success of RL techniques in various domains such as game playing, robotic manipulation, and locomotion tasks. A common rule of thumb amongst RL practitioners is that model-free methods have worse sample complexity compared to model-based methods, but are generally able to achieve better performance asymptotically since they do not suffer from biases in the model that lead to sub-optimal behavior Clavera et al. (2018); Nagabandi et al. (2018); Pong et al. (2018). However, there is currently no general theory which rigorously explains the gap between performance of model-based versus model-free methods. While there has been theoretical work studying both model-based and model-free methods in RL, prior work has primarily shown specific upper bounds Agrawal and Jia (2017); Azar et al. (2017); Jaksch et al. (2010); Jin et al. (2018); Strehl et al. (2006) which are not directly comparable, or information-theoretic lower bounds Jaksch et al. (2010); Jin et al. (2018) which are currently too coarse-grained to delineate between model-based and model-free methods. Furthermore, most of the prior work has focused primarily on the tabular Markov Decision Process (MDP) setting.

We take a first step towards a theoretical understanding of the differences between model-based and model-free methods for continuous control settings. While we are ultimately interested in comparing these methods for general MDPs with non-linear state transition dynamics, in this work we build upon recent progress in understanding the performance guarantees of data-driven methods for the Linear Quadratic Regulator (LQR). We study the asymptotic behavior of both *policy evaluation* and *policy optimization* on LQR, comparing the performance of simple model-based methods which use empirical state transition data to fit a dynamics model versus the performance of popular model-free methods from RL: *temporal-difference learning* for policy evaluation and *policy gradient methods* for policy optimization.

Our analysis shows that in the policy evaluation setting, a simple model-based plugin estimator is always asymptotically more sample efficient than the classical least-squares temporal difference (LSTD) estimator; the gap between the two methods can be at least a factor of state-dimension. For policy optimization, we consider a simple family of instances for which nominal control (also known as the certainty equivalence principle in control theory) is also at least several factors of state and input dimension more efficient than the widely used policy gradient method. Furthermore, the gap persists even when we employ commonly used baselines to reduce the variance of the policy gradient estimate. In both settings, we also show minimax lower bounds which highlight the near-optimality of model-based methods on the family of instances we consider. To the best of our knowledge, our work is the first to rigorously show a setting where a strict separation between a model-based and model-free method solving the same continuous control task occurs.

2. Main Results

In this paper, we study the performance of model-based and model-free algorithms for the *Linear Quadratic Regulator* (LQR) via two fundamental primitives in reinforcement learning: *policy evaluation* and *policy optimization*. In both tasks we fix an unknown dynamical system

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t \ ,$$

starting at $x_0 = 0$ (for simplicity) and driven by Gaussian white noise $w_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2 I_n)$. We let n denote the state dimension and d denote the input dimension, and assume the system is underactuated (i.e. $d \leq n$). We also fix two positive semi-definite cost matrices (Q, R).

2.1. Policy Evaluation

Given a controller $K \in \mathbb{R}^{d \times n}$ that stabilizes $(A_\star, B_\star)$, the policy evaluation task is to compute the (relative) value function $V^K(x)$:

$$V^K(x) := \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=0}^{T-1} (x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - \lambda_K) \,\middle|\, x_0 = x \right] \ , \ \ u_t = K x_t \ . \tag{2.1}$$

Above, λ_K is the infinite horizon average cost. It is well-known that $V^K(x)$ can be written as:

$$V^K(x) = \sigma_w^2 x^\mathsf{T} P_\star x \ , \tag{2.2}$$

where $P_\star = \mathsf{dlyap}(A_\star + B_\star K, Q + K^\mathsf{T} R K)$ solves the discrete-time Lyapunov equation:

$$(A_\star + B_\star K)^\mathsf{T} P_\star (A_\star + B_\star K) - P_\star + Q + K^\mathsf{T} R K = 0 \ . \tag{2.3}$$

From the Lyapunov equation, it is clear that given $(A_\star, B_\star)$, the solution to policy evaluation task is readily computable. In this paper, we study algorithms which only have input/output access to $(A_\star, B_\star)$. Specifically, we study *on-policy* algorithms that operate on a *single* trajectory, where the input u_t is determined by $u_t = Kx_t$. The variable that controls the amount of information available to the algorithm is T, the trajectory length. The trajectory will be denoted as $\{x_t\}_{t=0}^T$. We are interested in the asymptotic behavior of algorithms as $T \to \infty$.

Model-based algorithm. In light of Equation (2.3), the plugin estimator is a very natural model-based algorithm to use. Let $L_\star := A_\star + B_\star K$ denote the true closed-loop matrix. The plugin estimator uses the trajectory $\{x_t\}_{t=0}^T$ to estimate $L_\star$ via least-squares; call this $\widehat{L}(T)$. The estimator then returns $\widehat{P}_{\mathrm{plug}}(T)$ by using $\widehat{L}(T)$ in-place of $L_\star$ in (2.3). Algorithm 1 describes this estimator in more detail.

Algorithm 1 Model-based algorithm for policy evaluation.

Input: Policy $\pi(x) = Kx$, rollout length T, regularization $\lambda > 0$, thresholds $\zeta \in (0, 1)$ and $\psi > 0$.

1: Collect trajectory $\{x_t\}_{t=0}^T$ using the feedback $u_t = \pi(x_t) = Kx_t$.

2: Estimate the closed-loop matrix via least-squares:

$$\widehat{L}(T) = \left(\sum_{t=0}^{T-1} x_{t+1}x_t^\mathsf{T}\right)\left(\sum_{t=0}^{T-1} x_t x_t^\mathsf{T} + \lambda I_n\right)^{-1} .$$

3: **if** $\rho(\widehat{L}(T)) > \zeta$ or $\|\widehat{L}(T)\| > \psi$ **then**

4: Set $\widehat{P}_{\mathrm{plug}}(T) = 0$.

5: **else**

6: Set $\widehat{P}_{\mathrm{plug}}(T) = \mathsf{dlyap}(\widehat{L}(T), Q + K^\mathsf{T} RK)$.

7: **end if**

8: **return** $\widehat{P}_{\mathrm{plug}}(T)$.

Model-free algorithm. By observing that $V^K(x) = \sigma_w^2 x^\mathsf{T} P_\star x = \sigma_w^2\langle \mathrm{svec}(P_\star), \mathrm{svec}(xx^\mathsf{T})\rangle$, one can apply Least-Squares Temporal Difference Learning (LSTD) Boyan (1999); Bradtke and Barto (1996) with the feature map $\phi(x) := \mathrm{svec}(xx^\mathsf{T})$ to estimate $P_\star$. Here, $\mathrm{svec}(\cdot)$ vectorizes the upper triangular part of a symmetric matrix, weighting the off-diagonal terms by $\sqrt{2}$ to ensure consistency in the inner product. This is a classical algorithm in RL; the pseudocode is given in Algorithm 2.

We now proceed to compare the risk of Algorithm 1 versus Algorithm 2. Our notion of risk will be the expected squared error of the estimator: $\mathbb{E}[\|\widehat{P} - P_\star\|_F^2]$. Our first result gives an upper bound on the asymptotic risk of the model-based plugin Algorithm 1.

Theorem 2.1 *Let K stabilize $(A_\star, B_\star)$. Define $L_\star$ to be the closed-loop matrix $A_\star + B_\star K$ and let $\rho(L_\star) \in (0, 1)$ denote its stability radius. Recall that $P_\star$ is the solution to the discrete-time Lyapunov equation (2.3) that parameterizes the value function $V^K(x)$. We have that Algorithm 1 with thresholds (ζ, ψ) satisfying $\zeta \in (\rho(L_\star), 1)$ and $\psi \in (\|L_\star\|, \infty)$ and any fixed regularization parameter $\lambda > 0$ has the asymptotic risk upper bound:*

$$\lim_{T\to\infty} T \cdot \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|_F^2] \le 4\,\mathrm{tr}((I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1}(L_\star^\mathsf{T} P_\star^2 L_\star \otimes_s \sigma_w^2 P_\infty^{-1})(I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}}) .$$

Algorithm 2 Model-free algorithm for policy evaluation (LSTD) Bradtke and Barto (1996).

Input: Policy $\pi(x) = Kx$, rollout length T.

1: Collect trajectory $\{x_t\}_{t=0}^T$ using the feedback $u_t = \pi(x_t) = Kx_t$.
2: Estimate $\lambda_t \approx \sigma_w^2 \operatorname{tr}(P_\star)$ from $\{x_t\}_{t=0}^T$.
3: Compute (recall that $\phi(x) = \operatorname{svec}(xx^\mathsf{T})$):

$$\widehat{w}_{\mathrm{lstd}}(T) = \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t) - \phi(x_{t+1}))^\mathsf{T} \right)^{-1} \left(\sum_{t=0}^{T-1} (c_t - \lambda_t)\phi(x_t) \right) ,$$

4: Set $\widehat{P}_{\mathrm{lstd}}(T) = \operatorname{smat}(\widehat{w}_{\mathrm{lstd}}(T))$.
5: **return** $\widehat{P}_{\mathrm{lstd}}(T)$.

Here, $P_\infty = \mathsf{dlyap}(L_\star^\mathsf{T}, \sigma_w^2 I_n)$ *is the stationary covariance matrix of the closed-loop system* $x_{t+1} = L_\star x_t + w_t$ *and* $\otimes_s$ *denotes the symmetric Kronecker product.*

We make a few quick remarks regarding Theorem 2.1. First, while the risk bound is presented as an upper bound, the exact asymptotic risk can be recovered from the proof. Second, the thresholds (ζ, ψ) and regularization parameter λ do not affect the final asymptotic bound, but do possibly affect both higher order terms and the rate of convergence to the limiting risk. We include these thresholds as they simplify the proof. In practice, we find that thresholding or regularization is generally not needed, with the caveat that if the estimate $\widehat{L}(T)$ is not stable then the solution to the discrete Lyapunov equation is not guaranteed to exist (and when it exists is not guaranteed to be positive semidefinite). Finally, we remark that a non-asymptotic high probability upper bound for the risk of Algorithm 1 can be easily derived by combining the single trajectory learning results of Simchowitz et al. (2018) with standard results on perturbation of Lyapunov equations.

We now turn our attention to the model-free LSTD algorithm. Our next result gives a lower bound on the asymptotic risk of Algorithm 2.

Theorem 2.2 *Let* K *stabilize* $(A_\star, B_\star)$*. Define* $L_\star$ *to be the closed-loop matrix* $A_\star + B_\star K$*. Recall that* $P_\star$ *is the solution to the discrete-time Lyapunov equation (2.3) that parameterizes the value function* $V^K(x)$*. We have that Algorithm 2 with the cost estimates* λ_t *set to the true cost* $\lambda_\star := \sigma_w^2 \operatorname{tr}(P_\star)$ *satisfies the asymptotic risk lower bound:*

$$\begin{aligned} \liminf_{T\to\infty} T \cdot \mathbb{E}[\|\widehat{P}_{\mathrm{lstd}}(T) - P_\star\|_F^2] &\geq 4\mathcal{R}_{\mathsf{plug}} \\ &\quad + 8\sigma_w^2 \langle P_\infty, L_\star^\mathsf{T} P_\star^2 L_\star \rangle \operatorname{tr}((I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1}(P_\infty^{-1} \otimes_s P_\infty^{-1})(I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}}) \end{aligned}$$

Here, $\mathcal{R}_{\mathsf{plug}} := \lim_{T\to\infty} T \cdot \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|_F^2]$ *is the asymptotic risk of the plugin estimator,* $P_\infty = \mathsf{dlyap}(L_\star^\mathsf{T}, \sigma_w^2 I_n)$ *is the stationary covariance matrix of the closed loop system* $x_{t+1} = L_\star x_t + w_t$*, and* $\otimes_s$ *denotes the symmetric Kronecker product.*

Theorem 2.2 shows that the asymptotic risk of the model-free method always exceeds that of the model-based plugin method. We remark that we prove the theorem under an idealized setting where the infinite horizon cost estimate λ_t is set to the true cost $\lambda_\star$. In practice, the true cost is not known and must instead be estimated from the data at hand. However, for the purposes of our comparison this is not an issue because using the true cost $\lambda_\star$ over an estimator of $\lambda_\star$ only reduces the variance of the risk.

To get a sense of how much excess risk is incurred by the model-free method over the model-based method, consider the following family of instances, defined for $\rho \in (0,1)$ and $1 \le d \le n$:

$$\mathscr{F}(\rho, d, K) := \{(A_\star, B_\star) : A_\star + B_\star K = \tau P_E + \gamma I_n \ , \ (\tau, \gamma) \in (0,1) \ , \ \tau + \gamma \le \rho \ , \ \dim(E) \le d\} \ . \tag{2.4}$$

With this family, one can show with elementary computations that under the simplifying assumptions that $Q + K^\mathsf{T} R K = I_n$ and $d \asymp n$, Theorem 2.1 and Theorem 2.2 state that:

$$\lim_{T \to \infty} T \cdot \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|_F^2] \le O\left(\frac{\rho^2 n^2}{(1-\rho^2)^3} \right) \ ,$$

$$\liminf_{T \to \infty} T \cdot \mathbb{E}[\|\widehat{P}_{\mathrm{lstd}}(T) - P_\star\|_F^2] \ge \Omega\left(\frac{\rho^2 n^3}{(1-\rho^2)^3} \right) \ .$$

That is, for $\mathscr{F}(\rho, d, K)$, the plugin risk is a factor of state-dimension n less than the LSTD risk. Moreover, the non-asymptotic result for LSTD from Lemma 4.1 of Abbasi-Yadkori et al. (2019) (which extends the non-asymptotic *discounted* LSTD result from Tu and Recht (2018)) gives a bound of $\|\widehat{P}_{\mathrm{lstd}}(T) - P_\star\|_F^2 \le \widetilde{O}(n^3/T)$ w.h.p., which matches the asymptotic bound of Theorem 2.2 in terms of n up to logarithmic factors.

Our final result for policy evaluation is a minimax lower bound on the risk of any estimator over $\mathscr{F}(\rho, d, K)$.

Theorem 2.3 *Fix a $\rho \in (0,1)$ and suppose that K satisfies $Q + K^\mathsf{T} R K = I_n$. Suppose that n is greater than an absolute constant and $T \gtrsim n(1-\rho^2)/\rho^2$. We have that:*

$$\inf_{\widehat{P}} \sup_{(A_\star, B_\star) \in \mathscr{F}(\rho, \frac{n}{4}, K)} \mathbb{E}[\|\widehat{P} - P_\star\|_F^2] \gtrsim \frac{\rho^2 n^2}{(1-\rho^2)^3 T} \ ,$$

where the infimum is taken over all estimators $\widehat{P}$ taking input $\{x_t\}_{t=0}^T$.

Theorem 2.3 states that the rate achieved by the model-based Algorithm 3 over the family $\mathscr{F}(\rho, d, K)$ cannot be improved beyond constant factors, at least asymptotically; its dependence on both the state dimension n and stability radius ρ is optimal.

2.2. Policy Optimization

Given a finite horizon length T, the policy optimization task is to solve the finite horizon optimal control problem:

$$J_\star := \min_{u_t(\cdot)} \mathbb{E}\left[\sum_{t=0}^{T-1} (x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t) + x_T^\mathsf{T} Q x_T \right] \ , \ x_{t+1} = A_\star x_t + B_\star u_t + w_t \ . \tag{2.5}$$

We will focus on a special case of this problem when there is no penalty on the input: $Q = I_n$, $R = 0$, and $\mathrm{range}(A_\star) \subseteq \mathrm{range}(B_\star)$. In this situation, the cost function reduces to $\mathbb{E}[\sum_{t=0}^T \|x_t\|_2^2]$ and the optimal solution simply chooses a u_t that cancels out the state x_t; that is $u_t = K_\star x_t$ with

$K_\star := -B_\star^\dagger A_\star$. We work with this simple class of instances so that we can ensure that policy gradient converges to the optimal solution; in general this is not guaranteed.

We consider a slightly different input/output oracle model in this setting than we did in Section 2.1. The horizon length T is now considered fixed, and N rounds are played. At each round $i = 1, ..., N$, the algorithm chooses a feedback matrix $K_i \in \mathbb{R}^{d\times n}$. The algorithm then observes the trajectory $\{x_t^{(i)}\}_{t=0}^T$ by playing the control input $u_t^{(i)} = K_i x_t^{(i)} + \eta_t^{(i)}$, where $\eta_t^{(i)} \sim \mathcal{N}(0, \sigma_u^2 I_d)$ is i.i.d. noise used for the policy. This process then repeats for N total rounds. After the N rounds, the algorithm is asked to output a $\widehat{K}(N)$ and is assigned the risk $\mathbb{E}[J(\widehat{K}(N)) - J_\star]$, where $J(\widehat{K}(N))$ denotes playing the feedback $u_t = \widehat{K}(N)x_t$ on the true system $(A_\star, B_\star)$. We will study the behavior of algorithms when $N \to \infty$ (and T is held fixed).

Model-based algorithm. Under this oracle model, a natural model-based algorithm is to first use random open-loop feedback (i.e. $K_i = 0$) to observe N independent trajectories (each of length T), and then use the trajectory data to fit the state transition matrices $(A_\star, B_\star)$; call this estimate $(\widehat{A}(N), \widehat{B}(N))$. After fitting the dynamics, the algorithm then returns the estimate of $K_\star$ by solving the finite horizon problem (2.5) with $(\widehat{A}(N), \widehat{B}(N))$ taking the place of $(A_\star, B_\star)$. In general, however, the assumption that $\mathrm{range}(\widehat{A}(N)) \subseteq \mathrm{range}(\widehat{B}(N))$ will not hold, and hence the optimal solution to (2.5) will not be time-invariant. Moreover, solving for the best time-invariant static feedback for the finite horizon problem in general is not tractable. In light of this, to provide the fairest comparison to the model-free policy gradient method, we use the time-invariant static feedback that arises from infinite horizon solution given by the discrete algebraic Riccati equation as a proxy. We note that under our range inclusion assumption, the infinite horizon solution is a consistent estimator of the optimal feedback. The pseudo-code for this model-based algorithm is described in Algorithm 3.

Algorithm 3 Model-based algorithm for policy optimization.

Input: Horizon length T, rollouts N, regularization λ, thresholds $\varrho \in (0,1), \zeta, \psi, \gamma$.

Collect trajectories $\{\{(x_t^{(i)}, u_t^{(i)})\}_{t=0}^T\}_{i=1}^N$ using the feedback $K_i = 0$ (open-loop).

Estimate the dynamics matrices $(A_\star, B_\star)$ via regularized least-squares:

$$\widehat{\Theta}(N) = \left(\sum_{i=1}^N \sum_{t=0}^{T-1} x_{t+1} (z_t^{(i)})^\mathsf{T}\right)\left(\sum_{i=1}^N \sum_{t=0}^{T-1} z_t^{(i)} (z_t^{(i)})^\mathsf{T} + \lambda I_{n+d}\right)^{-1}, \quad z_t^{(i)} := (x_t^{(i)}, u_t^{(i)}) .$$

Set $(\widehat{A}, \widehat{B}) = \widehat{\Theta}(N)$.

if $\rho(\widehat{A}) > \varrho$ or $\|\widehat{A}\| > \zeta$ or $\|\widehat{B}\| > \psi$ or $\sigma_d(\widehat{B}) < \gamma$ **then**

  Set $\widehat{K}_{\mathrm{plug}}(N) = 0$.

else

  Set $\widehat{P} = \mathsf{dare}(\widehat{A}, \widehat{B}, I_n, 0)$ as the positive definite solution to[1]:

$$P = \widehat{A}^\mathsf{T} P \widehat{A} - \widehat{A}^\mathsf{T} P \widehat{B}(\widehat{B}^\mathsf{T} P \widehat{B})^{-1} \widehat{B}^\mathsf{T} P \widehat{A} + I_n .$$

  Set $\widehat{K}_{\mathrm{plug}}(N) = -(\widehat{B}^\mathsf{T} \widehat{P} \widehat{B})^{-1} \widehat{B}^\mathsf{T} \widehat{P} \widehat{A}$.

end if

return $\widehat{K}_{\mathrm{plug}}(N)$.

Model-free algorithm. We study a model-free algorithm based on policy gradients (see e.g. Peters and Schaal (2008); Williams (1992)). Here, we choose to parameterize the policy as a time-invariant linear feedback. The algorithm is described in Algorithm 4.

Algorithm 4 Model-free algorithm for policy optimization (REINFORCE) Peters and Schaal (2008); Williams (1992).

Input: Horizon length T, rollouts N, baseline functions $\{\Psi_t(\cdot;\cdot)\}$, step-sizes $\{\alpha_i\}$, initial K_1, threshold ζ.

1: **for** $i = 1, ..., N$ **do**
2: Collect trajectory $\mathcal{T}^{(i)} := \{(x_t^{(i)}, u_t^{(i)})\}_{t=0}^{T}$ using feedback K_i.
3: Compute policy gradient g_i as: $g_i = \frac{1}{\sigma_u^2} \sum_{t=0}^{T-1} \eta_t^{(i)} (x_t^{(i)})^\mathsf{T} \Psi_t(\mathcal{T}^{(i)}; K_i)$.
4: Take policy gradient step: $K_{i+1} = \mathsf{Proj}_{\|\cdot\| \le \zeta}(K_i - \alpha_i g_i)$.
5: **end for**
6: Set $\widehat{K}_{\mathrm{pg}}(N) = K_N$.
7: **return** $\widehat{K}_{\mathrm{pg}}(N)$.

In general for problems with a continuous action space, when applying policy gradient one has many degrees of freedom in choosing how to represent the policy π. Some of these degrees of freedom include whether or not the policy should be time-invariant and how much of the history before time t should be used to compute the action at time t. More broadly, the question is what function class should be used to model the policy. Ideally, one chooses a function class which is both capable of expressing the optimal solution and is easy to optimize over.

Another issue that significantly impacts the performance of policy gradient in practice is choosing a baseline which effectively reduces the variance of the policy gradient estimate. What makes computing a baseline challenging is that good baselines (such as value or advantage functions) require knowledge of the unknown MDP transition dynamics in order to compute. Therefore, one has to estimate the baseline from the empirical trajectories, adding another layer of complexity to the policy gradient algorithm.

In general, these issues are still an active area of research in RL and present many hurdles to a general theory for policy optimization. However, by restriction our attention to LQR, we can sidestep these issues which enables our analysis. In particular, by studying problems with no penalty on the input and where the state can be cancelled at every step, we know that the optimal control is a static time-invariant linear feedback. Therefore, we can restrict our policy representation to static linear feedback controllers without introducing any approximation error. Furthermore, it turns out that the specific assumptions on $(A_\star, B_\star)$ that we impose imply that the optimization landscape satisfies a standard notion of restricted strong convexity. This allows us to study policy gradient by leveraging the existing theory on the asymptotic distribution of stochastic gradient descent for strongly convex objectives. Finally, we can compute many of the standard baselines used in closed form, which further enables our analysis.

We note that in the literature, the model-based method is often called *nominal control* or the *certainty equivalence principle*. As noted in Dean et al. (2017), one issue with this approach is that

1. A sufficient condition for the existence of a unique positive definite solution to the discrete algebraic Riccati equation when $R = 0$ is that (A, B) is stabilizable and B has full column rank (Lemma C.1).

on an infinite horizon, there is no guarantee of robust stability with nominal control. However, as we are dealing with only finite horizon problems, the notion of stability is irrelevant.

Our first result for policy optimization gives the asymptotic risk of the model-based Algorithm 3.

Theorem 2.4 *Let $(A_\star, B_\star)$ be such that $A_\star$ is stable, $\mathrm{range}(A_\star) \subseteq \mathrm{range}(B_\star)$, and $B_\star$ has full column rank. We have that the model-based plugin Algorithm 3 with thresholds $(\varrho, \zeta, \psi, \gamma)$ such that $\varrho \in (\rho(A_\star), 1)$, $\zeta \in (\|A_\star\|, \infty)$, $\psi \in (\|B_\star\|, \infty)$, and $\gamma \in (0, \sigma_d(B_\star))$ satisfies the asymptotic risk bound:*

$$\lim_{N \to \infty} N \cdot \mathbb{E}[J(\widehat{K}_{\mathrm{plug}}(N)) - J_\star] = O(d(\mathrm{tr}(P_\infty^{-1}) + \|K_\star\|_F^2)) + o_T(1) .$$

Here, $P_\infty = \mathsf{dlyap}(A_\star, \sigma_u^2 B_\star B_\star^\mathsf{T} + \sigma_w^2 I_n)$ is the steady-state convariance of the system driven with control input $u_t \sim \mathcal{N}(0, \sigma_u^2 I_d)$, $K_\star$ is the optimal controller, and $O(\cdot)$ hides constants depending only on σ_w^2, σ_u^2.

We can interpret Theorem 2.4 by upper bounding $P_\infty^{-1} \preceq \sigma_w^{-2} I_n$. In this case if $\|K_\star\|_F^2 \leq O(n)$, then this result states that the asymptotic risk scales as $O(nd/N)$. Similar to Theorem 2.1, Theorem 2.4 requires the setting of thresholds $(\varrho, \zeta, \psi, \gamma)$. These thresholds serve two purposes. First, they ensure the existence of a unique positive definite solution to the discrete algebraic Riccati solution with the input penalty $R = 0$ (the details of this are worked out in Section C.2). Second, they simplify various technical aspects of the proof related to uniform integrability. In practice, such strong thresholds are not needed, and we leave either removing them or relaxing their requirements to future work.

Next, we look at the model-free case. As mentioned previously, baselines are very influential on the behavior of policy gradient. In our analysis, we consider three different baselines:

$$\Psi_t(\mathcal{T}; K) = \sum_{\ell=t+1}^{T} \|x_\ell\|_2^2 \,, \qquad (\textbf{Simple} \text{ baseline } b_t(x_t; K) = \|x_t\|_2^2.)$$

$$\Psi_t(\mathcal{T}; K) = \sum_{\ell=t}^{T} \|x_\ell\|_2^2 - V_t^K(x_t) \,, \qquad (\textbf{Value function} \text{ baseline } b_t(x_t; K) = V_t^K(x_t).)$$

$$\Psi_t(\mathcal{T}; K) = A_t^K(x_t, u_t) \,. \qquad (\textbf{Advantage} \text{ baseline } A_t^K(x_t, u_t) = Q_t^K(x_t, u_t) - V_t^K(x_t).)$$

Above, the simple baseline should be interpreted as having effectively no baseline; it turns out to simplify the variance calculations. On the other hand, the value function baseline V_t^K is a very popular heuristic used in practice Peters and Schaal (2008). Typically one has to actually estimate the value function for a given policy, since computing it requires knowledge of the model dynamics. In our analysis however, we simply assume the true value function is known. While this is an unrealistic assumption in practice, we note that this assumption substantially reduce the variance of policy gradient, and hence only serves to reduce the asymptotic risk. The last baseline we consider is to use the advantage function A_t^K. Using advantage functions has been shown to be quite effective in practice Schulman et al. (2016). It has the same issue as the value function baseline in that it needs to be estimated from the data; once again in our analysis we simply assume we have access to the true advantage function.

Our main result for model-free policy optimization is the following asymptotic risk lower bound on Algorithm 4.

Theorem 2.5 *Let $(A_\star, B_\star)$ be such that $A_\star$ is stable,* $\mathrm{range}(A_\star) \subseteq \mathrm{range}(B_\star)$*, and $B_\star$ has full column rank. Consider Algorithm 4 with $K_1 = 0_{d\times n}$, step-sizes $\alpha_i = [2(T-1)\sigma_w^2 \sigma_d(B_\star)^2 \cdot i]^{-1}$, and threshold $\zeta \in (\|K_\star\|, \infty)$. We have that the risk is lower bounded by:*

$$\liminf_{N\to\infty} N \cdot \mathbb{E}[J(\widehat{K}_{\mathrm{pg}}(N)) - J_\star] \geq \frac{1}{\sigma_d(B_\star)^2(1+\|B_\star\|^2)} \times \begin{cases} \Omega(T^2 d(n+\|B_\star\|_F^2)^3) + o_T(T^2) & \textit{(Simple baseline)} \\ \Omega(Td(n+\|B_\star\|_F^2)(n+\|B_\star^\mathsf{T} B_\star\|_F^2)) + o_T(T) & \textit{(Value function baseline)} \\ \Omega(d(n+\|B_\star\|_F^2)\|B_\star^\mathsf{T} B_\star\|_F^2) & \textit{(Advantage baseline)} \end{cases} .$$

Here, $\Omega(\cdot)$ hides constants depending only on σ_w^2, σ_u^2.

In order to interpret Theorem 2.5, we consider a restricted family of instances $(A_\star, B_\star)$. For a $\rho \in (0,1)$ and $1 \leq d \leq n$, we define the family $\mathscr{G}(\rho, d)$ over $(A_\star, B_\star)$ as:

$$\mathscr{G}(\rho, d) := \{(\rho U_\star U_\star^\mathsf{T}, \rho U_\star) : U_\star \in \mathbb{R}^{n\times d} \,,\ U_\star^\mathsf{T} U_\star = I_d\} \,.$$

This is a simple family where the $A_\star$ matrix is stable and contractive, and furthermore we have $\mathrm{range}(A_\star) = \mathrm{range}(B_\star)$. The optimal feedback is $K_\star = -U_\star^\mathsf{T}$ for each of these instances.

Theorem 2.5 states that for instances from $\mathscr{G}(\rho, d)$, the simple baseline has risk $\Omega(T^2 \cdot dn^3/N)$, the value function baseline has risk $\Omega(T \cdot dn^2/N)$, and the advantage baseline has risk $\Omega(d^2 n/N)$. On the other hand, Theorem 2.4 states that the model-based risk is upper bounded by $O(nd/N)$, which is less than the lower bound for all baselines considered in Theorem 2.5. For the simple and value function baselines, we see that the sample complexity of the model-free policy gradient method is several factors of n and T more than the model-based method. The extra factors of the horizon length appear due to the large variance of the policy gradient estimator without the variance reduction effects of the advantage baseline. The advantage baseline performs the best, only one factor of d more than the model-based method.

We note that we prove Theorem 2.5 with a specific choice of step size α_i. This step size corresponds to the standard $1/(mt)$ step sizes commonly found in proofs for SGD on strongly convex functions (see e.g. Rakhlin et al. (2012)), where m is the strong convexity parameter. We leave to future work extending our results to support Polyak-Ruppert averaging, which would yield asymptotic results that are more robust to specific step size choices.

Finally, we turn to our information-theoretic lower bound for any (possibly adaptive) method over the family $\mathscr{G}(\rho, d)$.

Theorem 2.6 *Fix a $d \leq n/2$ and suppose $d(n-d)$ is greater than an absolute constant. Consider the family $\mathscr{G}(\rho, d)$ as describe above. Fix a time horizon T and number of rollouts N. The risk over any algorithm $\mathcal{A}$ which plays (possibly adaptive) feedbacks of the form $u_t = K_i x_t + \eta_t$ with $\|K_i\| \leq 1$ and $\eta_t \sim \mathcal{N}(0, \sigma_u^2 I_d)$ is lower bounded by:*

$$\inf_{\mathcal{A}} \sup_{\substack{\rho \in (0,1/4), \\ (A_\star, B_\star) \in \mathscr{G}(d,\rho)}} \mathbb{E}[J(\mathcal{A}) - J_\star] \gtrsim \frac{\sigma_w^4}{\sigma_w^2 + \sigma_u^2} \frac{d(n-d)}{N} \,.$$

Observe that this bound is $\Omega(nd/N)$. Therefore, Theorem 2.6 tells us that asymptotically, the model-based method in Algorithm 3 is optimal in terms of its dependence on the state and input dimensions n and d over the family $\mathscr{G}(\rho, d)$.

3. Related Work

For general Markov Decision Processes (MDPs), the setting best understood theoretically is the tabular setting with finite state and action spaces. One definition (c.f. Strehl et al. (2006) and Jin et al. (2018)) of a tabular model-free algorithm is an algorithm that requires $o(S^2 AH)$ space, where S is the number of states, A is the number of actions, and H is the horizon length (this definition can also be modified for the infinite horizon). Under this definition, the best known regret bound in the model-based case is $\widetilde{O}(\sqrt{H^2 SAT})$ from Azar et al. (2017), which matches the known lower bound of $\Omega(\sqrt{H^2 SAT})$ from Jaksch et al. (2010); Jin et al. (2018) up to log factors. On the other hand, the best known regret bound in the model-free case is by Jin et al. (2018), who show that Q-learning achieves $\widetilde{O}(\sqrt{H^3 SAT})$ regret. It is open whether or not the gap in H is fundamental.

We now turn to the PAC setting for tabular MDPs. We specifically focus on the generative model assumption, where an oracle exists that allows one to query the state transition from any state/action pair at every timestep. For infinite horizon discounted MDPs, Azar et al. (2013) show that model-based policy iteration can find a ε optimal policy with $\widetilde{O}(SA/((1-\gamma)^3\varepsilon^2))$ samples as long as $\varepsilon \leq O(1/\sqrt{(1-\gamma)S})$, which matches the minimax lower bound given in the same work. Sidford et al. (2018) show that a model-free variance reduced value iteration algorithm also achieves $\widetilde{O}(SA/((1-\gamma)^3\varepsilon^2))$ sample complexity even beyond the small ε regime of the model-based method. Therefore, in this setting, there is no gap between model-based and model-free methods in the small ε regime, and the best upper bounds currently suggest that model-free methods actually outperform model-based methods in the moderate ε regime by a factor of $1/(1-\gamma)^2$ in sample complexity. It is still open if this gap can be resolved in the moderate ε regime.

Sun et al. (2019) present an information-theoretic definition of model-free algorithms which can be applied beyond the tabular setting. Under their definition, they construct a family of factored MDPs with horizon length H where any model-free algorithm incurs sample complexity $\Omega(2^H)$, whereas they exhibit a model-based algorithm that has sample complexity polynomial in H and other relevant quantities. We leave proving lower bounds for LQR under a rigorous definition of model-free algorithms to future work.

Turning our attention from tabular MDPs to LQR, the story is less complete. Unlike the tabular setting, the storage requirements of a model-based method are comparable to a model-free method. For instance, it takes $O(n(n+d))$ space to store the state transition model and $O((n+d)^2)$ space to store the Q-function. In presenting the known results of LQR, we will delineate between offline (one-shot) methods versus online (adaptive) methods.

In the offline setting, the first non-asymptotic result is from Fiechter (1997), who studied the sample complexity of the certainty equivalence controller on the *discounted* infinite horizon LQR problem. Later, Dean et al. (2017) study the average cost infinite horizon problem, using tools from robust control to quantify how the uncertainty in the model affects control performance in an interpretable way. Both works are model-based, since they both propose to first estimate the state transition matrices from sample trajectories and then use the estimated dynamics in a control synthesis procedure. For model-free methods on LQR, Tu and Recht (2018) study the performance of least-squares temporal difference learning (LSTD) Boyan (1999); Bradtke and Barto (1996) (c.f Section 2.1). They focus on the discounted cost setting and provide a non-asymptotic bound on the error of LSTD. Later, Abbasi-Yadkori et al. (2019) extend this result to the average cost setting. Most related to our analysis for policy gradient is Fazel et al. (2018), who study the performance of model-free policy gradient related methods on LQR. Unfortunately, their bounds do not give explicit dependence on the problem instance parameters and are therefore difficult to compare to. Furthermore, Fazel et al. study an infinite horizon setting where the only noise is in the initial state; all subsequence state transitions have no process noise. Other than our current work, we are unaware of any analysis (asymptotic or non-asymptotic) which explicitly studies the behavior of policy gradient on the finite horizon LQR problem. We also note that Fazel et al. analyze a policy optimization method which is more akin to derivative-free random search (e.g. Mania et al. (2018); Nesterov and Spokoiny (2017); Salimans et al. (2017)) than REINFORCE. Derivative-free random search for LQR is studied by Malik et al. (2019), who prove upper bounds that suggest that having two point evaluations is more sample efficient compared to single point evaluation. We leave analyzing these derivative-free algorithms under our framework to future work. Finally, note that all the results mentioned for LQR are only *upper bounds*; we are unaware of any *lower bounds* in the literature for LQR which give explicit dependence on the problem instance.

We now discuss known results for the online (adaptive) setting for LQR. For model-based algorithms, both *optimism in the face of uncertainty* (OFU) Abbasi-Yadkori and Szepesvári (2011); Faradonbeh et al. (2017); Ibrahimi et al. (2012) and *Thompson sampling* Abeille and Lazaric (2017, 2018); Ouyang et al. (2017) have been analyzed in the online learning literature. In both cases, the algorithms have been shown to achieve $\widetilde{O}(\sqrt{T})$ regret, which is known to be nearly optimal in the dependence on T. However, in nearly all the bounds the dependence on the problem instance parameters is hidden. Furthermore, it is currently unclear how to solve the OFU subproblem in polynomial time for LQR. In response to the computational issues with OFU, Dean et al. (2018) propose a polynomial time adaptive algorithm with sub-linear regret $\widetilde{O}(T^{2/3})$; their bounds make the dependence on the problem instance parameters explicit, but are quite conservative in this regard. For model-free algorithms, Abbasi-Yadkori et al. (2019) study the regret of a model-free algorithm based on follow the leader. They prove that their algorithm has regret $\widetilde{O}(T^{2/3+\varepsilon})$ for any $\varepsilon > 0$, nearly matching the bound given by Dean et al. in terms of the dependence on T. In terms of the dependence on the problem specific parameters, however, their bound is not directly comparable to that of Dean et al. Experimentally, Abbasi-Yadkori et al. observe that their model-free algorithm performs quite sub-optimally compared to model-based methods; these empirical observations are also consistent with experiments by Mania et al. (2018); Recht (2018); Tu and Recht (2018).

4. Proof Sketch

At a high level our proofs are relatively straightforward, relying on classical arguments from asymptotic statistics. However, various technical issues arise which make the arguments more involved. Below, we briefly outline the proof strategies that we use for our main results.

4.1. Policy Evaluation

Model-based (Algorithm 1 and Theorem 2.1). We first compute the limiting distribution of $\sqrt{T}\operatorname{vec}(\widehat{L}(T) - L_\star)$ (Lemma A.1) as a consequence of a standard Markov chain CLT (Theorem D.1). We then use the delta method to compute the limiting distribution of $\sqrt{T}\operatorname{svec}(\widehat{P}_{\mathrm{plug}}(T) - P_\star)$ by differentiating the map $L \mapsto \mathsf{dlyap}(L, Q + K^\mathsf{T} RK)$ at $L_\star$. We then prove that this sequence of random variables is uniformly integrable by controlling its higher order moments. Uniform integrability then implies (Lemma A.5) that the limit of the scaled risk $T \cdot \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|_F^2]$ is equal to the trace of the covariance of this limiting distribution, which yields the result. The details of this are worked out in Section B.1.

Model-free LSTD (Algorithm 2 and Theorem 2.2). This case is simpler since we can directly compute the limiting distribution of $\sqrt{T}(\widehat{w}_{\mathrm{lstd}} - w_\star)$ (Lemma A.3) using Markov chain CLTs. Then, the trace of the limiting distribution immediately lower bounds (Lemma A.5) the limit of the scaled risk $T \cdot \mathbb{E}[\|\widehat{P}_{\mathrm{lstd}} - P_\star\|_F^2]$ without having to establish uniform integrability. The details are worked out in Section B.2.

4.2. Policy Optimization

Model-based (Algorithm 3 and Theorem 2.4). As before, we start by computing the limiting distribution of $\sqrt{N}\mathrm{vec}(\widehat{\Theta}(N) - \Theta_\star)$ (Lemma A.2). Next, we use the delta method to compute the limiting distribution of $\sqrt{N}\mathrm{vec}(K(\widehat{\Theta}(N)) - K_\star)$, where $K(\Theta)$ is the optimal LQR controller designed with the model parameters Θ. This is done by differentiating the solution of the discrete algebraic Riccati equation with respect to the model parameters (Lemma C.2). Next, we make the observation that $\nabla J(K_\star) = 0$ and apply the second order delta method in order to compute the limiting distribution of $N \cdot (J(\widehat{K}(N)) - J_\star)$. We then show uniform integrability of this sequence by once again controlling its higher order moments. Again, by Lemma A.5 this yields an expression for the limit of the scaled risk $N \cdot \mathbb{E}[J(\widehat{K}(N)) - J_\star]$. The details are worked out in Section C.2.

Model-free policy gradients (Algorithm 4 and Theorem 2.5). This proof is the most involved, because it requires us to establish the limiting distribution of a first-order stochastic optimization algorithm with a convex projection step. In Section C.1, we show that the geometry of the smoothed policy gradient function satisfies restricted strong convexity for the particular dynamics we consider. Then in Section E, we compute the limiting distribution for SGD with projection on restricted strong convex functions, when the optimal solution lives in the interior of the domain. To do this, we build on the results of Toulis and Airoldi (2017) and Rakhlin et al. (2012). The remainder of the proof involves computing the variance of the various policy gradient estimators with different baselines. In general this calculation is not tractable, but our particular choice of models we study allows us to obtain very sharp estimates on this variance. The details are worked out in Section C.3.

5. Conclusion

We compared the asymptotic performance of both model-based and model-free methods for LQR. We showed that for policy evaluation, a simple plugin estimator is always more asymptotically sample efficient than the classical LSTD estimator. For policy optimization, we studied a family of instances where the convergence of policy gradient to the optimal solution is guaranteed, and showed that in this setting a simple plugin estimator is asymptotically at least a factor of state-dimension more efficient than policy gradient, depending on what specific baseline is used.

This work opens a variety of new directions for future research. The first is analyzing the asymptotic behavior of other model-free algorithms, such as derivative-free optimization, policy iteration, and Q-learning. Beyond algorithmic specific bounds, formulating a rigorous definition of model-free methods and proving lower bounds against the definition is of interest. As mentioned earlier, Sun et al. (2019) take an important first step in this direction. Another direction is to use asymptotic analysis to understand the effect of various baselines in policy gradient. Designing efficient baselines is still an open problem in RL, and understanding limiting behavior of different estimators could lead to new insights. Finally, extending the asymptotic analysis to the online setting would help us understand the difference between optimal $\sqrt{T}$ regret algorithms at a finer resolution.

Acknowledgments

We thank John Duchi and Daniel Russo, who both independently suggested studying LQR using asymptotic analysis; this paper is a direct result of their feedback. We also thank Horia Mania for many helpful discussions regarding policy gradient methods. Finally, we thank Nicolas Flammarion and Nilesh Tripuraneni for pointers regarding asymptotic analysis of stochastic gradient methods. ST is supported by a Google PhD fellowship. BR is generously supported in part by ONR awards N00014-17-1-2191, N00014-17-1-2401, and N00014-18-1-2833, the DARPA Assured Autonomy (FA8750-18-C-0101) and Lagrange (W911NF-16-1-0552) programs, and an Amazon AWS AI Research Award.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret Bounds for the Adaptive Control of Linear Quadratic Systems. In *Conference on Learning Theory*, 2011.

Yasin Abbasi-Yadkori, Nevena Lazić, and Csaba Szepesvári. Model-Free Linear Quadratic Control via Reduction to Expert Prediction. In *AISTATS*, 2019.

Marc Abeille and Alessandro Lazaric. Thompson Sampling for Linear-Quadratic Control Problems. In *AISTATS*, 2017.

Marc Abeille and Alessandro Lazaric. Improved Regret Bounds for Thompson Sampling in Linear Quadratic Control Problems. In *International Conference on Machine Learning*, 2018.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Neural Information Processing Systems*, 2017.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3): 325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *International Conference on Machine Learning*, 2017.

Patrick Billingsley. *Probability and Measure*. 1995.

Justin Boyan. Least-Squares Temporal Difference Learning. In *International Conference on Machine Learning*, 1999.

Steven J. Bradtke and Andrew G. Barto. Linear Least-Squares Algorithms for Temporal Difference Learning. *Machine Learning*, 22(1–3):33–57, 1996.

Ignasi Clavera, Jonas Rothfuss, John Schulman, Yasuhiro Fujita, Tamim Asfour, and Pieter Abbeel. Model-Based Reinforcement Learning via Meta-Policy Optimization. In *Conference on Robot Learning*, 2018.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the Sample Complexity of the Linear Quadratic Regulator. *arXiv:1710.01688*, 2017.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator. In *Neural Information Processing Systems*, 2018.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-Based Adaptive Regulation of Linear-Quadratic Systems. *arXiv:1711.07230*, 2017.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator. In *International Conference on Machine Learning*, 2018.

Claude-Nicolas Fiechter. PAC Adaptive Control of Linear Systems. In *Conference on Learning Theory*, 1997.

Morteza Ibrahimi, Adel Javanmard, and Benjamin Van Roy. Efficient Reinforcement Learning for High Dimensional Linear Quadratic Systems. In *Neural Information Processing Systems*, 2012.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Kevin G. Jamieson, Robert D. Nowak, and Benjamin Recht. Query Complexity of Derivative-Free Optimization. In *Neural Information Processing Systems*, 2012.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning Provably Efficient? In *Neural Information Processing Systems*, 2018.

Galin L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

Harold Kushner and George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. 2003.

Jan R. Magnus. The expectation of products of quadratic forms in normal variables: the practice. *Statistica Neerlandica*, 33(3):131–136, 1979.

Dhruv Malik, Kush Bhatia, Koulik Khamaru, Peter L. Bartlett, , and Martin J. Wainwright. Derivative-Free Methods for Policy Optimization: Guarantees for Linear Quadratic Systems. In *AISTATS*, 2019.

Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search provides a competitive approach to reinforcement learning. In *Neural Information Processing Systems*, 2018.

Henry B. Mann and Abraham Wald. On the Statistical Treatment of Linear Stochastic Difference Equations. *Econometrica*, 11(3–4):173–220, 1943.

Abdelkader Mokkadem. Mixing properties of ARMA processes. *Stochastic Processes and their Applications*, 29(2):309–315, 1988.

B. Molinari. The Stabilizing Solution of the Discrete Algebraic Riccati Equation. *IEEE Transactions on Automatic Control*, 20(3):396–399, 1975.

Anusha Nagabandi, Gregory Kahn, Ronald S. Fearing, and Sergey Levine. Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning. In *International Conference on Robotics and Automation*, 2018.

Yurii Nesterov and Vladimir Spokoiny. Random Gradient-Free Minimization of Convex Functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with Thompson sampling. In *55th Annual Allerton Conference on Communication, Control, and Computing*, 2017.

Alain Pajor. Metric Entropy of the Grassmann Manifold. *Convex Geometric Analysis*, 34:181–188, 1998.

Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008.

Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine. Temporal Difference Models: Model-Free Deep RL for Model-Based Control. In *International Conference on Learning Representations*, 2018.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *International Conference on Machine Learning*, 2012.

Benjamin Recht. A Tour of Reinforcement Learning: The View from Continuous Control. *arXiv:1806.09460*, 2018.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution Strategies as a Scalable Alternative to Reinforcement Learning. *arXiv:1703.03864*, 2017.

Kathrin Schäcke. On the Kronecker Product. Master's thesis, University of Waterloo, 2004.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations*, 2016.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye. Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model. In *Neural Information Processes Systems*, 2018.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification. In *Conference on Learning Theory*, 2018.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC Model-Free Reinforcement Learning. In *International Conference on Machine Learning*, 2006.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-Based Reinforcement Learning in Contextual Decision Processes. In *Conference on Learning Theory*, 2019.

Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

Stephen Tu and Benjamin Recht. Least-Squares Temporal Difference Learning for the Linear Quadratic Regulator. In *International Conference on Machine Learning*, 2018.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–246, 1992.

Fuzhen Zhang. *The Schur Complement and its Applications*, volume 4 of *Numerical Methods and Algorithms*. 2005.

Appendix A. Asymptotic Toolbox

Our analysis relies heavily on computing limiting distributions for the various estimators we study. A crucial fact we use is that if the matrix $L_\star$ is stable, then the Markov chain $\{x_t\}$ given by $x_{t+1} = L_\star x_t + w_t$ with $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$ is geometrically ergodic. This allows us to apply well known limit theorems for ergodic Markov chains.

In what follows, we let $\xrightarrow{\text{a.s.}}$ denote almost sure convergence and $\overset{D}{\rightsquigarrow}$ denote convergence in distribution. We also let $\otimes$ denote the standard Kronecker product and $\otimes_s$ denote the *symmetric* Kronecker product; see e.g. Schäcke (2004) for a review of the basic properties of the Kronecker and symmetric Kronecker product which we will use extensively throughout the sequel. For a matrix M, the notation $\mathrm{vec}(M)$ denotes the vectorized version of M by stacking the columns. We will also let $\mathrm{svec}(\cdot)$ denote the operator that satisfies $\langle \mathrm{svec}(M_1), \mathrm{svec}(M_2)\rangle = \langle M_1, M_2 \rangle$ for all symmetric matrices $M_1, M_2 \in \mathbb{R}^{n\times n}$, where the first inner product is with respect to $\mathbb{R}^{n(n+1)/2}$ and the second is with respect to $\mathbb{R}^{n\times n}$. Finally, we let $\mathrm{mat}(\cdot)$ and $\mathrm{smat}(\cdot)$ denote the functional inverses of $\mathrm{vec}(\cdot)$ and $\mathrm{svec}(\cdot)$. The proofs of the results presented in this section are deferred to Section D.

We first state a well-known result that concerns the least-squares estimator of a stable dynamical system. In the scalar case, this result dates back to Mann and Wald (1943).

Lemma A.1 *Let $x_{t+1} = L_\star x_t + w_t$ be a dynamical system with $L_\star$ stable and $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$. Given a trajectory $\{x_t\}_{t=0}^T$, let $\widehat{L}(T)$ denote the least-squares estimator of $L_\star$ with regularization $\lambda \geq 0$:*

$$\widehat{L}(T) = \arg\min_{L \in \mathbb{R}^{n\times n}} \frac{1}{2}\sum_{t=0}^{T-1} \|x_{t+1} - L x_t\|_2^2 + \frac{\lambda}{2}\|L\|_F^2 .$$

Let P_∞ denote the stationary covariance matrix of the process $\{x_t\}_{t=0}^\infty$, i.e. $L_\star P_\infty L_\star^\mathsf{T} - P_\infty + \sigma_w^2 I_n = 0$. We have that $\widehat{L}(T) \xrightarrow{\text{a.s.}} L_\star$ and furthermore:

$$\sqrt{T}\mathrm{vec}(\widehat{L}(T) - L_\star) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \sigma_w^2 (P_\infty^{-1} \otimes I_n)) .$$

We now consider a slightly altered process where the system is no longer autonomous, and instead will be driven by white noise.

Lemma A.2 *Let $x_{t+1} = A_\star x_t + B_\star u_t + w_t$ be a stable dynamical system driven by $u_t \sim \mathcal{N}(0, \sigma_u^2 I_d)$ and $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$. Consider a least-squares estimator $\widehat{\Theta}$ of $\Theta_\star := (A_\star, B_\star) \in \mathbb{R}^{n \times (n+d)}$ based off of N independent trajectories of length T, i.e. given $\{\{z_t^{(i)} := (x_t^{(i)}, u_t^{(i)})\}_{t=0}^T\}_{i=1}^N$,*

$$\widehat{\Theta}(N) = \arg \min_{(A,B) \in \mathbb{R}^{n \times (n+d)}} \frac{1}{2} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \| x_{t+1}^{(i)} - A x_t^{(i)} - B u_t^{(i)} \|_2^2 + \frac{\lambda}{2} \| \begin{bmatrix} A & B \end{bmatrix} \|_F^2 \,.$$

Let P_∞ denote the stationary covariance of the process $\{x_t\}_{t=0}^\infty$, i.e. P_∞ solves

$$A_\star P_\infty A_\star^\mathsf{T} - P_\infty + \sigma_u^2 B_\star B_\star^\mathsf{T} + \sigma_w^2 I_n = 0 \,.$$

We have that $\widehat{\Theta}(N) \xrightarrow{\text{a.s.}} \Theta_\star$ and furthermore:

$$\sqrt{N} \mathrm{vec}(\widehat{\Theta}(N) - \Theta_\star) \overset{D}{\rightsquigarrow} \mathcal{N}\left(0, \frac{\sigma_w^2}{T} \begin{bmatrix} P_\infty^{-1} & 0 \\ 0 & (1/\sigma_u^2) I_d \end{bmatrix} \otimes I_n + o(1/T)\right) \,.$$

Next, we consider the asymptotic distribution of Least-Squares Temporal Difference Learning for LQR.

Lemma A.3 *Let $x_{t+1} = A_\star x_t + B_\star u_t + w_t$ be a linear system driven by $u_t = K x_t$ and $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$. Suppose the closed-loop matrix $A_\star + B_\star K$ is stable. Let ν_∞ denote the stationary distribution of the Markov chain $\{x_t\}_{t=0}^\infty$. Define the two matrices A_∞, B_∞, the mapping $\psi(x)$, and the vector $w_\star$ as*

$$\begin{aligned}
A_\infty &:= \mathop{\mathbb{E}}_{\substack{x \sim \nu_\infty, \\ x' \sim p(\cdot | x, \pi(x))}} [\phi(x)(\phi(x) - \phi(x'))^\mathsf{T}] \,, \\
B_\infty &:= \mathop{\mathbb{E}}_{\substack{x \sim \nu_\infty, \\ x' \sim p(\cdot | x, \pi(x))}} [((\phi(x') - \psi(x))^\mathsf{T} w_\star)^2 \phi(x) \phi(x)^\mathsf{T}] \,, \\
\psi(x) &:= \mathop{\mathbb{E}}_{x' \sim p(\cdot | x, \pi(x))} [\phi(x')] \,, \\
w_\star &:= \mathrm{svec}(P_\star) \,.
\end{aligned}$$

Let $\widehat{w}_{\mathrm{lstd}}(T)$ denote the LSTD estimator given by:

$$\widehat{w}_{\mathrm{lstd}}(T) = \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t) - \phi(x_{t+1}))^\mathsf{T} \right)^{-1} \left(\sum_{t=0}^{T-1} (c_t - \lambda_t) \phi(x_t) \right) \,.$$

Suppose that LSTD is run with the true $\lambda_t = \lambda_\star := \sigma_w^2 \operatorname{tr}(P_\star)$ and that the matrix A_∞ is invertible. We have that $\widehat{w}_{\mathrm{lstd}}(T) \xrightarrow{\text{a.s.}} w_\star$ and furthermore:

$$\sqrt{T}(\widehat{w}_{\mathrm{lstd}}(T) - w_\star) \overset{D}{\rightsquigarrow} \mathcal{N}(0, A_\infty^{-1} B_\infty A_\infty^{-\mathsf{T}}) \,.$$

As a corollary to Lemma A.3, we work out the formulas for A_∞ and B_∞ and a useful lower bound.

Corollary A.4 *In the setting of Lemma A.3, with $L_\star = A_\star + B_\star K$, we have that the matrix A_∞ is invertible, and:*

$$\begin{aligned}
A_\infty &= (P_\infty \otimes_s P_\infty) - (P_\infty L_\star^\mathsf{T} \otimes_s P_\infty L_\star^\mathsf{T}) \,, \\
B_\infty &= (\sigma_w^2 \langle P_\infty, L_\star^\mathsf{T} P_\star^2 L_\star \rangle + 2\sigma_w^4 \|P_\star\|_F^2)(2(P_\infty \otimes_s P_\infty) + \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T}) \\
&\quad + 2\sigma_w^2(\mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty L_\star^\mathsf{T} P_\star^2 L_\star P_\infty)^\mathsf{T} + \mathrm{svec}(P_\infty L_\star^\mathsf{T} P_\star^2 L_\star P_\infty)\mathrm{svec}(P_\infty)^\mathsf{T}) \\
&\quad + 8\sigma_w^2 (P_\infty L_\star^\mathsf{T} P_\star^2 L_\star P_\infty \otimes_s P_\infty) \,.
\end{aligned}$$

Furthermore, we can lower bound the matrix $A_\infty^{-1} B_\infty A_\infty^{-\mathsf{T}}$ by:

$$\begin{aligned}
A_\infty^{-1} B_\infty A_\infty^{-\mathsf{T}} &\succeq 8\sigma_w^2 \langle P_\infty, L_\star^\mathsf{T} P_\star^2 L_\star \rangle (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} (P_\infty^{-1} \otimes_s P_\infty^{-1}) (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}} \\
&\quad + 16\sigma_w^2 (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} (L_\star^\mathsf{T} P_\star^2 L_\star \otimes_s P_\infty^{-1}) (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}} \,. \qquad \text{(A.1)}
\end{aligned}$$

Next, we state a standard lemma which we will use to convert convergence in distribution guarantees to guarantees regarding the convergence of risk.

Lemma A.5 *Suppose that $\{X_n\}$ is a sequence of random vectors and $X_n \overset{D}{\rightsquigarrow} X$. Suppose that f is a non-negative continuous real-valued function such that $\mathbb{E}[f(X)] < \infty$. We have that:*

$$\liminf_{n \to \infty} \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)] \,.$$

If additionally we have $\sup_{n \geq 1} \mathbb{E}[f(X_n)^{1+\varepsilon}] < \infty$ holds for some $\varepsilon > 0$, then $\lim_{n\to\infty} \mathbb{E}[f(X_n)]$ exists and

$$\lim_{n \to \infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)] \,.$$

Proof Both facts are standard consequences of weak convergence of probability measures; see e.g. Chapter 5 of Billingsley (1995) for more details. ■

The next claim uniformly controls the p-th moments of the regularized least-squares estimate when T is large enough. This technical result will allow us to invoke Lemma A.5 to obtain convergence in L^p.

Lemma A.6 *Let $x_{t+1} = L_\star x_t + w_t$ with $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$ and $L_\star$ stable. Fix a regularization parameter $\lambda > 0$ and let $\widehat{L}(T)$ denote the LS estimator:*

$$\widehat{L}(T) = \arg\min_{L \in \mathbb{R}^{n \times n}} \frac{1}{2} \sum_{t=0}^{T-1} \|x_{t+1} - L x_t\|_2^2 + \frac{\lambda}{2}\|L\|_F^2 \,.$$

Fix a finite $p \geq 1$. Let $C_{L_\star,\lambda,n}$ and $C_{L_\star,\lambda,n,p}$ denote constants that depend only on $L_\star, \lambda, n$ (resp. $L_\star, \lambda, n, p$) and not on T, δ. Fix a $\delta \in (0,1)$. With probability at least $1-\delta$, as long as $T \geq C_{L_\star,\lambda,n} \log(1/\delta)$ we have:

$$\|\widehat{L}(T) - L_\star\| \leq C'_{L_\star,\lambda,n} \sqrt{\frac{\log(1/\delta)}{T}} \,.$$

Furthermore, as long as $T \geq C_{L_\star,\lambda,n,p}$, then:

$$\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] \leq C'_{L_\star,\lambda,n,p} \frac{1}{T^{p/2}} \,.$$

The next result is the analogue of Lemma A.6 for the non-autonomous system driven by white noise.

Lemma A.7 *Let $x_{t+1} = A_\star x_t + B_\star u_t + w_t$ with $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$, $u_t \sim \mathcal{N}(0, \sigma_u^2 I_d)$, and $A_\star$ stable. Fix a regularization parameter $\lambda > 0$ and let $\widehat{\Theta}(N)$ denote the LS estimator:*

$$\widehat{\Theta}(N) = \arg \min_{(A,B) \in \mathbb{R}^{n \times (n+d)}} \frac{1}{2} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \| x_{t+1}^{(i)} - A x_t^{(i)} - B u_t^{(i)} \|_2^2 + \frac{\lambda}{2} \| \begin{bmatrix} A & B \end{bmatrix} \|_F^2 \,.$$

Fix a finite $p \geq 1$. Let $C_{\Theta_\star, T, \lambda, n, d}$ and $C_{\Theta_\star, T, \lambda, n, d, p}$ denote constants that depend only on $\Theta_\star, T, \lambda, n, d$ (resp. $\Theta_\star, T, \lambda, n, d, p$) and not on N, δ. Fix a $\delta \in (0, 1)$. With probability at least $1 - \delta$, as long as $N \geq C_{\Theta_\star, T, \lambda, n, d} \log(1/\delta)$ we have:

$$\| \widehat{\Theta}(N) - \Theta_\star \| \leq C'_{\Theta_\star, T, \lambda, n, d} \sqrt{\frac{\log(1/\delta)}{N}} \,.$$

Furthermore, as long as $N \geq C_{\Theta_\star, T, \lambda, n, d, p}$, then:

$$\mathbb{E}[\| \widehat{\Theta}(N) - \Theta_\star \|^p] \leq C'_{\Theta_\star, T, \lambda, n, d, p} \frac{1}{N^{p/2}} \,.$$

Proof The proof is nearly identical to that of Lemma A.6, except we use the concentration result of Proposition 1.1 from Dean et al. (2017) instead of Theorem 2.4 of Simchowitz et al. (2018) to establish concentration over multiple independent rollouts. We omit the details as they very closely mimic that of Lemma A.6.

We note that in doing this we obtain a sub-optimal dependence on the horizon length T. This can be remedied by a more careful argument combining the concentration along each trajectory from Simchowitz et al. with the concentration across independent trajectories from Dean et al. However, as in our limit theorems only N the rollout length is being sent to infinity (e.g. T is considered a constant), a sub-optimal bound in T will suffice for our purpose. ■

Our final asymptotic result deals with the performance of stochastic gradient descent (SGD) with projection. This will be our key ingredient in analyzing policy gradient (Algorithm 4). While the asymptotic performance of SGD (and more generally stochastic approximation) is well-established (see e.g. Kushner and Yin (2003)), we consider a slight modification where the iterates are projected back into a compact convex set at every iteration. As long as the optimal solution is not on the boundary of the projection set, then one intuitively does not expect the asymptotic distribution to be affected by this projection, since eventually as SGD converges towards the optimal solution the projection step will effectively be inactive. Our result here makes this intuition rigorous. It follows by combining the asymptotic analysis of Toulis and Airoldi (2017) with the high probability bounds for SGD from Rakhlin et al. (2012).

To state the result, we need a few definitions. First, we say a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ satisfies *restricted strong convexity* (RSC) on a compact convex set $\Theta \subseteq \mathbb{R}^d$ if it has a unique minimizer $\theta_\star \in \mathrm{int}(\Theta)$ and for some $m > 0$, we have $\langle \nabla F(\theta), \theta - \theta_\star \rangle \geq m \| \theta - \theta_\star \|_2^2$ for all $\theta \in \Theta$. We denote this by $\mathsf{RSC}(m, \Theta)$.

Lemma A.8 *Let $F \in \mathcal{C}^3(\Theta)$ and suppose F satisfies $\mathsf{RSC}(m, \Theta)$. Let $\theta_\star \in \Theta$ denote the unique minimizer of F in Θ. Suppose we have a stochastic gradient oracle $g(\theta;\xi)$ such that g is continuous in both θ, ξ and $\nabla F(\theta) = \mathbb{E}_\xi[g(\theta;\xi)]$ for some distribution over ξ. Suppose that for some $G_1, G_2, L > 0$, for all $p \in [1,4]$ and $\delta \in (0,1)$, we have that*

$$\sup_{\theta \in \Theta} \mathbb{E}_\xi[\|g(\theta;\xi)\|_2^p] \le G_1^p \,, \tag{A.2}$$

$$\mathbb{P}_\xi\left(\sup_{\theta\in\Theta} \|g(\theta;\xi)\|_2 > G_2 \operatorname{polylog}(1/\delta) \right) \le \delta \,, \tag{A.3}$$

$$\mathbb{E}_\xi[\|g(\theta;\xi) - g(\theta_\star;\xi)\|_2^2] \le L\|\theta - \theta_\star\|_2^2 \;\; \forall \theta \in \Theta \,. \tag{A.4}$$

Given an sequence $\{\xi_t\}_{t=1}^\infty$ drawn i.i.d. from the law of ξ, consider the sequence of iterates $\{\theta_t\}_{t=1}^\infty$ starting with $\theta_1 \in \Theta$ and defined as:

$$\theta_{t+1} = \mathsf{Proj}_\Theta(\theta_t - \alpha_t g(\theta_t;\xi_t)) \,, \;\; \alpha_t = \frac{1}{mt} \,.$$

We have that:

$$\lim_{T\to\infty} mT \cdot \mathrm{Var}(\theta_T) = \Xi \,, \tag{A.5}$$

where $\Xi = \mathsf{lyap}(\frac{m}{2} I_d - \nabla^2 F(\theta_\star), \mathbb{E}_\xi[g(\theta_\star;\xi)g(\theta_\star;\xi)^\mathsf{T}])$ solves the continuous-time Lyapunov equation:

$$\left(\frac{m}{2} I_d - \nabla^2 F(\theta_\star)\right)\Xi + \Xi\left(\frac{m}{2} I_d - \nabla^2 F(\theta_\star)\right) + \mathbb{E}_\xi[g(\theta_\star;\xi)g(\theta_\star;\xi)^\mathsf{T}] = 0 \,. \tag{A.6}$$

We also have that for any $G \in \mathcal{C}^3(\Theta)$ with $\nabla G(\theta_\star) = 0$ and $\nabla^2 G(\theta_\star) \succ 0$,

$$\liminf_{T\to\infty} T \cdot \mathbb{E}[G(\theta_T) - G(\theta_\star)] \ge \frac{1}{2m} \operatorname{tr}(\nabla^2 G(\theta_\star) \cdot \Xi) \,. \tag{A.7}$$

We defer the proof of this lemma to Section E of the Appendix. We quickly comment on how the last inequality can be used. Taking trace of both sides from Equation A.6, we obtain:

$$\operatorname{tr}(\Xi \cdot (\nabla^2 F(\theta_\star) - \frac{m}{2} I_d)) = \frac{1}{2}\mathbb{E}_\xi[\|g(\theta_\star;\xi)\|_2^2] \,.$$

We now upper bound the LHS as:

$$\begin{aligned}
&\operatorname{tr}(\Xi \cdot (\nabla^2 F(\theta_\star) - \frac{m}{2} I_d)) \\
&= \operatorname{tr}(\Xi \cdot \nabla^2 G(\theta_\star)^{1/2} \cdot \nabla^2 G(\theta_\star)^{-1/2}(\nabla^2 F(\theta_\star) - \frac{m}{2} I_d)\nabla^2 G(\theta_\star)^{-1/2} \cdot \nabla^2 G(\theta_\star)^{1/2}) \\
&\le \operatorname{tr}(\Xi \cdot \nabla^2 G(\theta_\star))\lambda_{\max}(\nabla^2 G(\theta_\star)^{-1/2}(\nabla^2 F(\theta_\star) - \frac{m}{2} I_d)\nabla^2 G(\theta_\star)^{-1/2}) \\
&= \operatorname{tr}(\Xi \cdot \nabla^2 G(\theta_\star))\lambda_{\max}(\nabla^2 G(\theta_\star)^{-1}(\nabla^2 F(\theta_\star) - \frac{m}{2} I_d)) \,.
\end{aligned}$$

Combining the last two equations we obtain that:

$$\begin{aligned}
\liminf_{T\to\infty} T \cdot \mathbb{E}[G(\theta_T) - G(\theta_\star)] &\ge \frac{1}{2m} \operatorname{tr}(\Xi \cdot \nabla^2 G(\theta_\star)) \\
&\ge \frac{1}{4m\lambda_{\max}(\nabla^2 G(\theta_\star)^{-1}(\nabla^2 F(\theta_\star) - \frac{m}{2} I_d))}\mathbb{E}_\xi[\|g(\theta_\star;\xi)\|_2^2] \,.
\end{aligned} \tag{A.8}$$

We will use this last estimate in our analysis.

Appendix B. Analysis of Policy Evaluation Methods

In this section, recall that Q, R, K are fixed, and furthermore define $M := Q + K^\mathsf{T} R K$.

B.1. Proof of Theorem 2.1

The strategy is as follows. Recall that Lemma A.1 gives us the asymptotic distribution of the (regularized) least-squares estimator $\widehat{L}(T)$ of the true closed-loop matrix $L_\star$. For a stable matrix L, let $P(L) = \mathsf{dlyap}(L, M)$. Since the map $L \mapsto P(L)$ is differentiable, using the delta method we can recover the asymptotic distribution of $\sqrt{T}\mathrm{svec}(P(\widehat{L}(T)) - P_\star)$. Upper bounding the trace of the covariance matrix for this asymptotic distribution then yields Theorem 2.1.

Let $[DP(L)]$ denote the Fréchet derivative of the map $P(\cdot)$ evaluated at L, and let $[DP(L)](X)$ denote the action of the linear operator $[DP(L)]$ on X. By a straightforward application of the implicit function theorem, we have that:

$$[DP(L_\star)](X) = \mathsf{dlyap}(L_\star, X^\mathsf{T} P_\star L_\star + L_\star^\mathsf{T} P_\star X) \,.$$

Before we proceed, we introduce some notation surrounding Kronecker products. Let Γ denote the matrix such that $(A \otimes_s B) = \frac{1}{2}\Gamma^\mathsf{T}(A \otimes B + B \otimes A)\Gamma$ for any square matrices A, B. It is a fact that $\Gamma\mathrm{vec}(S) = \mathrm{svec}(S)$ for any symmetric matrix S. Also let Π be the orthonormal matrix such that $\Pi\mathrm{vec}(X) = \mathrm{vec}(X^\mathsf{T})$ for all square matrices X. It is not hard to verify that $\Pi^\mathsf{T}(A \otimes B)\Pi = (B \otimes A)$, a fact we will use later. With this notation, we proceed as follows:

$$\begin{aligned}
\mathrm{svec}([DP(L_\star)](X)) &= (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1}\mathrm{svec}(X^\mathsf{T} P_\star L_\star + L_\star^\mathsf{T} P_\star X) \\
&= (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1}\Gamma\mathrm{vec}(X^\mathsf{T} P_\star L_\star + L_\star^\mathsf{T} P_\star X) \\
&= (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1}\Gamma((L_\star^\mathsf{T} P_\star \otimes I_n)\Pi + (I_n \otimes L_\star^\mathsf{T} P_\star))\mathrm{vec}(X) \,.
\end{aligned}$$

Applying Lemma A.1 in conjunction with the delta method, we obtain:

$$\sqrt{T}\mathrm{svec}(P(\widehat{L}(T)) - P_\star) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \sigma_w^2 (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} V (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}}) \,,$$

where,

$$\begin{aligned}
V &:= \Gamma[((L_\star^\mathsf{T} P_\star \otimes I_n)\Pi + (I_n \otimes L_\star^\mathsf{T} P_\star))(P_\infty^{-1} \otimes I_n)((L_\star^\mathsf{T} P_\star \otimes I_n)\Pi + (I_n \otimes L_\star^\mathsf{T} P_\star))^\mathsf{T}]\Gamma^\mathsf{T} \\
&\overset{(a)}{\preceq} 2\Gamma[(L_\star^\mathsf{T} P_\star \otimes I_n)\Pi(P_\infty^{-1} \otimes I_n)\Pi^\mathsf{T}(P_\star L_\star \otimes I_n) + (I_n \otimes L_\star^\mathsf{T} P_\star)(P_\infty^{-1} \otimes I_n)(I_n \otimes P_\star L_\star)]\Gamma^\mathsf{T} \\
&= 2\Gamma[(L_\star^\mathsf{T} P_\star \otimes I_n)(I_n \otimes P_\infty^{-1})(P_\star L_\star \otimes I_n) + (I_n \otimes L_\star^\mathsf{T} P_\star)(P_\infty^{-1} \otimes I_n)(I_n \otimes P_\star L_\star)]\Gamma^\mathsf{T} \\
&= 2\Gamma[(L_\star^\mathsf{T} P_\star^2 L_\star \otimes P_\infty^{-1}) + (P_\infty^{-1} \otimes L_\star^\mathsf{T} P_\star^2 L_\star)]\Gamma^\mathsf{T} \\
&= 4(L_\star^\mathsf{T} P_\star^2 L_\star \otimes_s P_\infty^{-1}) \,.
\end{aligned}$$

In (a), we used the inequality for any matrices X, Y and positive definite matrices F, G, (see e.g. Chapter 3, page 94 of Zhang (2005)):

$$(X + Y)(F + G)^{-1}(X + Y)^\mathsf{T} \preceq X F^{-1} X^\mathsf{T} + Y G^{-1} Y^\mathsf{T} \,.$$

Suppose that the sequence $\{\|Z_T\|_F^2\}$ is uniformly integrable, where $Z_T := \sqrt{T}\mathrm{svec}(P(\widehat{L}(T)) - P_\star)$. Then:

$$\lim_{T\to\infty} T \cdot \mathbb{E}[\|P(\widehat{L}(T)) - P_\star\|_F^2] \leq 4\,\mathrm{tr}((I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1}(L_\star^\mathsf{T} P_\star^2 L_\star \otimes_s \sigma_w^2 P_\infty^{-1})(I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}}) \,,$$

which is the desired bound on the asymptotic risk.

We now show that the sequence $\{\|Z_T\|_F^2\}$ is uniformly integrable. To do this, we need a simple matrix stability perturbation bound.

Lemma B.1 *Let A be a stable matrix that satisfies $\|A^k\| \le C\rho^k$ for all $k \ge 0$ for some $C > 0$ and $\rho \in (0,1)$. Fix a $\gamma \in (\rho, 1)$. Suppose that Δ is a perturbation that satisfies:*

$$\|\Delta\| \le \frac{\gamma - \rho}{C} .$$

Then we have that (a) $A + \Delta$ is a stable matrix with $\rho(A+\Delta) \le \gamma$ and (b) $\|(A+\Delta)^k\| \le C\gamma^k$ for all $k \ge 0$.

Proof We start by proving (b). Fix an integer $k \ge 1$. Consider the expansion of $(A+\Delta)^k$ into 2^k terms. Label all these terms as $T_{i,j}$ for $i = 0, ..., k$ and $j = 1, ..., \binom{k}{i}$ where i denotes the degree of Δ in the term (hence there are $\binom{k}{i}$ terms with a degree of i for Δ). Using the fact that $\|A^k\| \le C\rho^k$ for all $k \ge 0$, we can bound $\|T_{i,j}\| \le C^{i+1}\rho^{k-i}\|\Delta\|^i$. Hence by triangle inequality:

$$\begin{aligned}
\|(A+\Delta)^k\| &\le \sum_{i=0}^{k} \sum_{j} \|T_{i,j}\| \\
&\le \sum_{i=0}^{k} \binom{k}{i} C^{i+1} \rho^{k-i} \|\Delta\|^i \\
&= C \sum_{i=0}^{k} \binom{k}{i} (C\|\Delta\|)^i \rho^{k-i} \\
&= C(C\|\Delta\| + \rho)^k \\
&\le C\gamma^k ,
\end{aligned}$$

where the last inequality uses the assumption $\|\Delta\| \le \frac{\gamma-\rho}{C}$. This gives the claim (b).

To derive the claim (a), we use the inequality that $\rho(A+\Delta) \le \|(A+\Delta)^k\|^{1/k} \le C^{1/k}\gamma$ for any $k \ge 1$. Since this holds for any $k \ge 1$, we can take the infimum over all $k \ge 1$ on the RHS, which yields the desired claim. ∎

Fix a finite $p \ge 1$. Since $L_\star$ is stable and $\zeta \in (\rho(L_\star), 1)$, there exists a $C_\star$ such that $\|L_\star^k\| \le C_\star \zeta^k$ for all $k \ge 0$. For the rest of the proof, $O(\cdot), \Omega(\cdot)$ will hide constants that depend on $L_\star, C_\star, n, p, \lambda, \zeta, \psi$, but not on T. Set $\delta_T = O(1/T^{p/2})$ and let T be large enough so that there exists an event $\mathcal{E}_{\mathsf{Bdd}}$ promised by Lemma A.6 such that $\mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}) \ge 1 - \delta_T$ and on $\mathcal{E}_{\mathsf{Bdd}}$ we have $\|\widehat{L}(T) - L_\star\| \le O(\sqrt{\log(1/\delta_T)/T})$. Let T also be large enough so that on $\mathcal{E}_{\mathsf{Bdd}}$, we have $\|\widehat{L}(T) - L_\star\| \le \min((\gamma - \rho_\star)/C_\star, \psi - \|L_\star\|)$. With this setting, we have that on $\mathcal{E}_{\mathsf{Bdd}}$, for any $\alpha \in (0,1)$,

$$\tilde{L}(\alpha) := \alpha\widehat{L}(T) + (1-\alpha)L_\star \in \left\{ L \in \mathbb{R}^{n\times n} : \rho(L) \le \zeta \,,\ \ \|L\| \le \min\left(\|L_\star\| + \frac{\gamma - \rho_\star}{C_\star}, \psi \right) \right\} =: \mathcal{G} .$$

Therefore on $\mathcal{E}_{\mathsf{Bdd}}$, for some $\alpha \in (0,1)$,

$$\|P(\widehat{L}(T)) - P_\star\| = \|[DP(\tilde{L}(\alpha))](\widehat{L}(T) - L_\star)\| \le \sup_{\tilde{L}\in\mathcal{G}} \|[DP(\tilde{L})]\| \|\widehat{L}(T) - L_\star\| := S\|\widehat{L}(T) - L_\star\| .$$

Here the norm $|||[H]||| := \sup_{\|X\|\leq 1} |||[H](X)|||$. We have that S is finite since $\mathcal{G}$ is a compact set. Next, define the set $\mathcal{G}_{\mathsf{Alg}}$ as:

$$\mathcal{G}_{\mathsf{Alg}} := \{ L \in \mathbb{R}^{n \times n} : \rho(L) \leq \zeta \, , \ \|L\| \leq \psi \} \, ,$$

and define the event $\mathcal{E}_{\mathsf{Alg}}$ as $\mathcal{E}_{\mathsf{Alg}} := \{ \widehat{L}(T) \in \mathcal{G}_{\mathsf{Alg}} \}$. Consider the decomposition:

$$\begin{aligned}
\mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p] &= \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c}] \\
&\leq \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \\
&\qquad + \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c}] \, .
\end{aligned}$$

In what follows we will assume that T is sufficiently large.

On $\mathcal{E}_{\mathsf{Bdd}}$. On this event, since we have $\mathcal{E}_{\mathsf{Bdd}} \subseteq \mathcal{E}_{\mathsf{Alg}}$, we can bound by Lemma A.6:

$$\mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] = \mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}} \cap \mathcal{E}_{\mathsf{Alg}}}] \leq S^p \mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] \leq O(1/T^{p/2}) \, .$$

On $\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}$. On this event, we use the fact that $\mathcal{G}_{\mathsf{Alg}}$ is compact to bound:

$$\begin{aligned}
\mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] &\leq \sup_{\widehat{L} \in \mathcal{G}_{\mathsf{Alg}}} \|\mathsf{dlyap}(\widehat{L}, Q + K^\mathsf{T} R K) - P_\star\|^p \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}) \\
&\leq \sup_{\widehat{L} \in \mathcal{G}_{\mathsf{Alg}}} \|\mathsf{dlyap}(\widehat{L}, Q + K^\mathsf{T} R K) - P_\star\|^p \delta_T \\
&\leq O(1/T^{p/2}) \, .
\end{aligned}$$

On $\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c$. On this event, we simply have:

$$\mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c}] = \|P_\star\|^p \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c) \leq \|P_\star\|^p \delta_T \leq O(1/T^{p/2}) \, .$$

Putting it together. Combining these bounds we obtain that $\mathbb{E}[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p] \leq O(1/T^{p/2})$. Recall that $Z_T = \mathrm{svec}(P(\widehat{L}(T)) - P_\star)$. We have that for any finite $\gamma > 0$ and $T \geq \Omega(1)$:

$$\begin{aligned}
\mathbb{E}[\|Z_T\|_F^{2+\gamma}] &= T^{(2+\gamma)/2} \mathbb{E}[\|P(\widehat{L}(T)) - P_\star\|_F^{2+\gamma}] \\
&\leq n^{(2+\gamma)/2} T^{(2+\gamma)/2} \mathbb{E}[\|P(\widehat{L}(T)) - P_\star\|^{2+\gamma}] \\
&\leq n^{(2+\gamma)/2} T^{(2+\gamma)/2} O(1/T^{(2+\gamma)/2}) \\
&\leq n^{(2+\gamma)/2} O(1) \, .
\end{aligned}$$

On the other hand, when $T \leq O(1)$ it is easy to see that $\mathbb{E}[\|Z_T\|_F^{2+\gamma}]$ is finite. Hence we have $\sup_{T \geq 1} \mathbb{E}[\|Z_T\|_F^{2+\gamma}] < \infty$ which shows the desired uniformly integrable condition. This concludes the proof of Theorem 2.1.

B.2. Proof of Theorem 2.2

Lemma A.3 (specifically (A.1)) combined with Lemma A.5 tells us that:

$$\begin{aligned}
&\liminf_{T \to \infty} T \cdot \mathbb{E}[\|\widehat{P}_{\mathrm{lstd}}(T) - P_\star\|_F^2] \geq \mathrm{tr}(A_\infty^{-1} B_\infty A_\infty^{-\mathsf{T}}) \\
&\quad \geq 8\sigma_w^2 \, \mathrm{tr}(\langle P_\infty, L_\star^\mathsf{T} P_\star^2 L_\star \rangle (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} (P_\infty^{-1} \otimes_s P_\infty^{-1}) (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}}) \\
&\qquad + 16\sigma_w^2 \, \mathrm{tr}((I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-1} (L_\star^\mathsf{T} P_\star^2 L_\star \otimes_s P_\infty^{-1}) (I - L_\star^\mathsf{T} \otimes_s L_\star^\mathsf{T})^{-\mathsf{T}}) \, .
\end{aligned}$$

The claim now follows by using the risk bound from Theorem 2.1.

B.3. Proof of Theorem 2.3

Let $E_1, ..., E_N$ be d-dimensional subspaces of $\mathbb{R}^n$ with $d \leq n/2$ such that $\|P_{E_i} - P_{E_j}\|_F \gtrsim \sqrt{d}$. By Proposition 8 of Pajor (1998), we can take $N \geq e^{n(n-d)}$. Now consider instances A_i with $A_i = \tau P_{E_i} + \gamma I_n$ for a $\tau, \gamma \in (0,1)$ to be determined. We will set $\tau + \gamma = \rho$ so that each A_i is contractive (i.e. $\|A_i\| < 1$) and hence stable. This means implicitly that we will require $\tau < \rho$. Let $\mathbb{P}_i$ denote the distribution over $(x_1, ..., x_T)$ induced by instance A_i. We have that:

$$
\begin{aligned}
\mathrm{KL}(\mathbb{P}_i, \mathbb{P}_j) &= \sum_{t=1}^{T} \mathbb{E}_{x_t \sim \mathbb{P}_i}[\mathrm{KL}(\mathcal{N}(A_i x_t, \sigma^2 I), \mathcal{N}(A_j x_t, \sigma^2 I))] \\
&= \frac{1}{2\sigma^2} \sum_{t=1}^{T} \mathbb{E}_{x_t \sim \mathbb{P}_i}[\|(A_i - A_j)x_t\|_2^2] \\
&\leq \frac{\|A_i - A_j\|^2}{2\sigma^2} \sum_{t=1}^{T} \mathrm{tr}(\mathbb{E}_{x_t \sim \mathbb{P}_i}[x_t x_t^\mathsf{T}]) \\
&\leq \frac{\tau^2}{\sigma^2} T \, \mathrm{tr}(P_\infty) \\
&= \tau^2 T \left(\frac{d}{1-\rho^2} + \frac{n-d}{1-\gamma^2} \right) \\
&\leq \tau^2 T \frac{n}{1-\rho^2} \, .
\end{aligned}
$$

Now if we choose $n(n-d) \geq 4 \log 2$ and $T \gtrsim n(1-\rho^2)/\rho^2$, we can set $\tau^2 \asymp \frac{n(1-\rho^2)}{T}$ and obtain that $\frac{I(V;X) + \log 2}{\log |V|} \leq 1/2$.

On the other hand, let $P_i = \mathsf{dlyap}(A_i, I_n)$. We have that for any integer $k \geq 0$:

$$
\begin{aligned}
(\tau P_{E_i} + \gamma I_n)^k - (\tau P_{E_j} + \gamma I_n)^k &= \sum_{\ell=0}^{k} \binom{k}{\ell} \gamma^{k-\ell} \tau^\ell (P_{E_i}^\ell - P_{E_j}^\ell) \\
&= k\gamma^{k-1}\tau(P_{E_i} - P_{E_j}) + \sum_{\ell=2}^{k} \binom{k}{\ell} \gamma^{k-\ell} \tau^\ell (P_{E_i}^\ell - P_{E_j}^\ell) \, .
\end{aligned}
$$

Hence,

$$
\begin{aligned}
P_i - P_j &= \sum_{k=1}^{\infty} ((A_i^k)^\mathsf{T} A_i^k - (A_j^k)^\mathsf{T} A_j^k) \\
&= \sum_{k=1}^{\infty} (A_i^{2k} - A_j^{2k}) \\
&= \left(\sum_{k=1}^{\infty} 2k\gamma^{2k-1}\tau + \sum_{k=2}^{\infty} \sum_{\ell=2}^{2k} \binom{2k}{\ell} \gamma^{2k-\ell} \tau^\ell \right) (P_{E_i} - P_{E_j}) \\
&= \left(\frac{2\gamma\tau}{(1-\gamma^2)^2} + \sum_{k=2}^{\infty} \sum_{\ell=2}^{k} \binom{k}{\ell} \gamma^{k-\ell} \tau^\ell \right) (P_{E_i} - P_{E_j}) \, .
\end{aligned}
$$

Therefore,

$$\|P_i - P_j\|_F \geq \frac{2\gamma\tau}{(1-\gamma^2)^2}\|P_{E_i} - P_{E_j}\|_F \gtrsim \frac{\gamma\tau}{(1-\gamma^2)^2}\sqrt{d}\,.$$

The claim now follows by Fano's inequality and setting $d = n/4$.

Appendix C. Analysis of Policy Optimization Methods

C.1. Preliminary Calculations

Given $(A_\star, B_\star)$ with $\mathrm{range}(A_\star) \subseteq \mathrm{range}(B_\star)$ and $\mathrm{rank}(B_\star) = d$, let $J_\Sigma(K)$ for a $K \in \mathbb{R}^{d\times n}$ denote the following cost:

$$J_\Sigma(K) := \mathbb{E}\left[\sum_{t=1}^{T}\|x_t\|_2^2\right]\,,\;\; x_{t+1} = A_\star x_t + B_\star u_t + w_t\,,\;\; u_t = Kx_t\,,\;\; w_t \sim \mathcal{N}(0,\Sigma)\,.$$

Here we assume $T \geq 2$ and Σ is positive definite. We write $J(K) = J_{\sigma_w^2 I_n}(K)$ as shorthand. Under this feedback law, we have $x_t \sim \mathcal{N}(0, \sum_{\ell=0}^{t-1} L(K)^\ell \Sigma (L(K)^\ell)^\mathsf{T})$ with $L(K) := A_\star + B_\star K$. Letting L be shorthand for $L(K)$, the cost can be written as:

$$J_\Sigma(K) = \sum_{t=1}^{T}\sum_{\ell=0}^{t-1}\mathrm{tr}(L^\ell\Sigma(L^\ell)^\mathsf{T}) = T\,\mathrm{tr}(\Sigma) + \sum_{t=1}^{T}\sum_{\ell=1}^{t-1}\mathrm{tr}(L^\ell\Sigma(L^\ell)^\mathsf{T})\,.$$

Let $K_\star$ denote the minimizer of $J_\Sigma(K)$; under our assumptions we have that $K_\star = -B_\star^\dagger A_\star$. Furthermore, because of the range condition we can write $A_\star = B_\star B_\star^\dagger A_\star$. Therefore, $L(K) = B_\star(B_\star^\dagger A_\star + K)$. While the function $J_\Sigma(K)$ is not convex, it has many nice properties. First, $J_\Sigma(K)$ satisfies a *quadratic growth condition*:

$$\begin{aligned} J_\Sigma(K) - J_\Sigma(K_\star) &\geq (T-1)\,\mathrm{tr}(L\Sigma L^\mathsf{T}) \\ &= (T-1)\,\mathrm{tr}(B_\star(B_\star^\dagger A_\star + K)\Sigma(B_\star^\dagger A_\star + K)^\mathsf{T} B_\star^\mathsf{T}) \\ &= (T-1)\mathrm{vec}(B_\star^\dagger A_\star + K)^\mathsf{T}(\Sigma \otimes B_\star^\mathsf{T} B_\star)\mathrm{vec}(B_\star^\dagger A_\star + K) \\ &\geq (T-1)\lambda_{\min}(\Sigma)\sigma_{\min}(B_\star)^2\|K - K_\star\|_F^2\,. \end{aligned} \tag{C.1}$$

Next, we will see $J_\Sigma(K)$ satisfies restricted strong convexity. To do this, we first compute the gradient $\nabla J_\Sigma(K)$. Consider the function $M \mapsto M^\ell$ for any integer $\ell \geq 2$. We have that the derivatives are:

$$[DM^\ell](\Delta) = \sum_{k=0}^{\ell-1} M^k \Delta M^{\ell-k-1}\,,\;\; [D(M^\ell)^\mathsf{T}](\Delta) = \sum_{k=0}^{\ell-1}(M^{\ell-k-1})^\mathsf{T}\Delta^\mathsf{T}(M^k)^\mathsf{T}\,.$$

By the chain rule,

$$[DL(K)^\ell](\Delta) = \sum_{k=0}^{\ell-1} L(K)^k B_\star \Delta L(K)^{\ell-k-1}\,.$$

Hence by the chain rule again,

$$
\begin{aligned}
[D\,\mathrm{tr}(L(K)^\ell \Sigma (L^\ell)^\mathsf{T})](\Delta) &= \mathrm{tr}\left(L^\ell \Sigma \sum_{k=0}^{\ell-1} (L^{\ell-k-1})^\mathsf{T} \Delta^\mathsf{T} B_\star^\mathsf{T} (L^k)^\mathsf{T} \right) \\
&\quad + \mathrm{tr}\left(\sum_{k=0}^{\ell-1} L^k B_\star \Delta L^{\ell-k-1} \Sigma (L^\ell)^\mathsf{T} \right) \\
&= 2 \left\langle \sum_{k=0}^{\ell-1} B_\star^\mathsf{T} (L^k)^\mathsf{T} L^\ell \Sigma (L^{\ell-k-1})^\mathsf{T}, \Delta \right\rangle .
\end{aligned}
$$

We have shown that:

$$
\nabla_K \mathrm{tr}(L(K)^\ell \Sigma (L(K)^\ell)^\mathsf{T}) = 2 \sum_{k=0}^{\ell-1} B_\star^\mathsf{T} (L^k)^\mathsf{T} L^\ell \Sigma (L^{\ell-k-1})^\mathsf{T} .
$$

Therefore we can compute the gradient of $J_\Sigma(K)$ as:

$$
\nabla J_\Sigma(K) = 2(T-1) B_\star^\mathsf{T} L \Sigma + 2 \sum_{\ell=2}^{T-1} \sum_{k=0}^{\ell-1} (T-\ell) B_\star^\mathsf{T} (L^k)^\mathsf{T} L^\ell \Sigma (L^{\ell-k-1})^\mathsf{T} .
$$

Now observe that $L(K) = B_\star(K - K_\star)$ and therefore:

$$
\begin{aligned}
\langle \nabla J_\Sigma(K), K - K_\star \rangle &= \mathrm{tr}(\nabla J_\Sigma(K)(K-K_\star)^\mathsf{T}) \\
&= 2(T-1)\,\mathrm{tr}(L \Sigma L^\mathsf{T}) + 2 \sum_{\ell=2}^{T-1} \sum_{k=0}^{\ell-1} (T-\ell)\,\mathrm{tr}(L^\ell \Sigma (L^\ell)^\mathsf{T}) \\
&\overset{(a)}{\geq} 2(T-1)\,\mathrm{tr}(L \Sigma L^\mathsf{T}) \\
&\geq 2(T-1) \lambda_{\min}(\Sigma) \sigma_{\min}(B_\star)^2 \|K - K_\star\|_F^2 .
\end{aligned}
$$

Above, (a) follows since $\mathrm{tr}(AB) \geq 0$ for positive semi-definite matrices A, B. This condition proves that $K = K_\star$ is the unique stationary point, and establishes the restricted strong convexity $\mathsf{RSC}(m, \mathbb{R}^{d \times n})$ condition for $J_\Sigma(K)$ with constant $m = 2(T-1)\lambda_{\min}(\Sigma)\sigma_{\min}(B_\star)^2$.

Finally, we show that the Hessian of $J_\Sigma(K)$ evaluated at $K_\star$ is positive definite. Fix a test matrix $H \in \mathbb{R}^{d \times n}$, and define the function $g(t) := \langle H, \nabla J_\Sigma(K_\star + tH) \rangle$. By standard properties of the directional derivative, we have that $\mathrm{Hess} J_\Sigma(K_\star)[H, H] = g'(0)$. Observing that $L(K_\star + tH) = t \cdot B_\star H$, we have that:

$$
\begin{aligned}
g(t) &= 2(T-1) t\, \mathrm{tr}(\Sigma H^\mathsf{T} B_\star^\mathsf{T} B_\star H) \\
&\qquad + 2 \sum_{\ell=2}^{T-1} \sum_{k=0}^{\ell-1} (T-\ell) t^{2\ell-1}\, \mathrm{tr}(H^\mathsf{T} B_\star^\mathsf{T} (H^\mathsf{T} B_\star^\mathsf{T})^k (B_\star H)^\ell \Sigma (H^\mathsf{T} B_\star^\mathsf{T})^{\ell-k-1}) ,
\end{aligned}
$$

from which we conclude:

$$
\mathrm{Hess} J_\Sigma(K_\star)[H, H] = 2(T-1)\,\mathrm{tr}(\Sigma H^\mathsf{T} B_\star^\mathsf{T} B_\star H) = 2(T-1) \mathrm{vec}(H)^\mathsf{T} (\Sigma \otimes B_\star^\mathsf{T} B_\star) \mathrm{vec}(H) .
$$

C.2. Proof of Theorem 2.4

Recal that the pair (A, B) is stabilizable if there exists a feedback matrix K such that $\rho(A+BK) < 1$. We first state a result which gives a sufficient condition for the existence of a unique positive definite solution to the discrete algebraic Riccati equation.

Lemma C.1 (Theorem 2, Molinari (1975)) *Suppose that $Q \succ 0$, (A, B) is stabilizable, and B has full column rank. Then there exists a unique positive definite solution P to the DARE:*

$$P = A^\mathsf{T} PA - A^\mathsf{T} PB(B^\mathsf{T} PB)^{-1}B^\mathsf{T} PA + Q\,. \tag{C.2}$$

This P satisfies the lower bound $P \succeq Q$, and if A is contractive (i.e. $\|A\| < 1$) satisfies the upper bound $\|P\| \leq \frac{\|Q\|}{1-\|A\|^2}$.

Proof Define the map $\Psi(z; A) := B^\mathsf{T}(z^{-1}I_n - A)^{-\mathsf{T}} Q(zI_n - A)^{-1}B$. Let K be such that $A+BK$ is stable. We observe that for $|z| = 1$, we have that:

$$\Psi(z; A + BK) = B^*(zI_n - (A + BK))^{-*}Q(zI_n - (A + BK))^{-1}B \succ 0\,.$$

This is because $Q \succ 0$, $B^*B \succ 0$, and the matrix $zI_n-(A+BK)$ does not drop rank since $A+BK$ has no eigenvalues on the unit circle. Therefore by Theorem 2 of Molinari (1975), there exists a unique symmetric solution P that satisfies (C.2) with the additional constraint that $B^\mathsf{T} PB \succ 0$ and that $\rho(A_c) < 1$ with $A_c := A - B(B^\mathsf{T} PB)^{-1}B^\mathsf{T} PA$. But (C.2) means that:

$$\begin{aligned}
A_c^\mathsf{T} PA_c &= (A - B(B^\mathsf{T} PB)^{-1}B^\mathsf{T} PA)^\mathsf{T} P(A - B(B^\mathsf{T} PB)^{-1}B^\mathsf{T} PA) \\
&= A^\mathsf{T} PA - A^\mathsf{T} PB(B^\mathsf{T} PB)^{-1}B^\mathsf{T} PA - A^\mathsf{T} PB(B^\mathsf{T} PB)^{-1}B^\mathsf{T} PA \\
&\quad + A^\mathsf{T} PB(B^\mathsf{T} PB)^{-1}B^\mathsf{T} PB(B^\mathsf{T} PB)^{-1}B^\mathsf{T} PA \\
&= A^\mathsf{T} PA - A^\mathsf{T} PB(B^\mathsf{T} PB)^{-1}B^\mathsf{T} PA \\
&= P - Q\,.
\end{aligned}$$

Hence, we have $A_c^\mathsf{T} PA_c - P + Q = 0$, and since A_c is stable by Lyapunov theory we know that $P \succeq Q$. Furthermore, since $P \succeq 0$, (C.2) implies that $P \preceq A^\mathsf{T} PA + Q$ from which the upper bound on $\|P\|$ follows under the contractivity assumptions. ■

Next, we state a result which gives the derivative of the discrete algebraic Riccati equation.

Lemma C.2 (Section A.2 of Abeille and Lazaric (2017)) *Let (Q, R) be positive semidefinite matrices. Suppose that (A, B) are such that there exists a unique positive definite solution $P(A, B)$ to* $\mathsf{dare}(A, B, Q, R)$*. For a perturbation $\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix} \in \mathbb{R}^{n\times(n+d)}$, we have that the Fréchet derivative $[D_{(A,B)}P(A, B)]$ evaluated at the perturbation $\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}$ is given by:*

$$[D_{(A,B)}P(A, B)](\begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}) = \mathsf{dlyap}\left(A_c, A_c^\mathsf{T} P \begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix} \begin{bmatrix}I_n \\ K\end{bmatrix} + \begin{bmatrix}I_n \\ K\end{bmatrix}^\mathsf{T} \begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}^\mathsf{T} PA_c\right)\,,$$

where $P = P(A, B)$, $K = -(B^\mathsf{T} PB + R)^{-1}B^\mathsf{T} PA$, and $A_c = A + BK$.

With these two lemmas, we are ready to proceed. We differentiate the map:

$$h(A,B) := -(B^\mathsf{T} P(A,B) B + R)^{-1} B^\mathsf{T} P(A,B) A \,.$$

By the chain rule:

$$\begin{aligned}[D_{(A,B)} h(A,B)](\Delta) &= -(B^\mathsf{T} P B + R)^{-1}(B^\mathsf{T} P \Delta_A + \Delta_B^\mathsf{T} P A + B^\mathsf{T} [D_{(A,B)} P](\Delta) A) \\ &\quad + (B^\mathsf{T} P B + R)^{-1}(\Delta_B^\mathsf{T} P B + B^\mathsf{T} P \Delta_B + B^\mathsf{T}[D_{(A,B)} P](\Delta) B)(B^\mathsf{T} P B + R)^{-1} B^\mathsf{T} P A \,.\end{aligned}$$

We now evaluate this derivative at:

$$A = A_\star \,, B = B_\star \,, Q = I_n \,, R = 0 \,.$$

Note that $P(A,B) = I_n$ and also by Lemma C.2, we have that $[D_{(A,B)} P(A,B)] = 0$, since $A_c = 0$. Therefore the derivative $[D_{(A,B)} h(A,B)](\Delta)$ simplifies to:

$$\begin{aligned}[D_{(A,B)} h(A,B)](\Delta) &= -(B_\star^\mathsf{T} B_\star)^{-1}(B_\star^\mathsf{T} \Delta_A + \Delta_B^\mathsf{T} A_\star) + (B_\star^\mathsf{T} B_\star)^{-1}(\Delta_B^\mathsf{T} B_\star + B_\star^\mathsf{T} \Delta_B) B_\star^\dagger A_\star \\ &= -B_\star^\dagger \Delta_A + B_\star^\dagger \Delta_B B_\star^\dagger A_\star \,.\end{aligned}$$

Hence we have:

$$\mathrm{vec}([D_{(A,B)} h(A,B)](\Delta)) = \begin{bmatrix} -(I_n \otimes B_\star^\dagger) & (B_\star^\dagger A_\star)^\mathsf{T} \otimes B_\star^\dagger \end{bmatrix} \mathrm{vec}(\Delta) \,.$$

Now using the assumption that $A_\star$ is stable, from Lemma A.2 we have that by the delta method:

$$\sqrt{N} \mathrm{vec}(h(\widehat{A}(N), \widehat{B}(N)) - K_\star) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \Psi) =: \varphi \,,$$

where

$$\Psi := \frac{\sigma_w^2}{T} \begin{bmatrix} -(I_n \otimes B_\star^\dagger) & (B_\star^\dagger A_\star)^\mathsf{T} \otimes B_\star^\dagger \end{bmatrix} \left(\begin{bmatrix} P_\infty^{-1} & 0 \\ 0 & (1/\sigma_u^2) I_d \end{bmatrix} \otimes I_n \right) \begin{bmatrix} -(I_n \otimes (B_\star^\dagger)^\mathsf{T}) \\ B_\star^\dagger A_\star \otimes (B_\star^\dagger)^\mathsf{T} \end{bmatrix} + o(1/T) \,.$$

We now make use of the second order delta method. Recall that the Hessian of J at $K_\star$ is $\mathrm{Hess} J(K_\star)[H,H] = 2(T-1)\sigma_w^2 \langle H, B_\star^\mathsf{T} B_\star H \rangle$. If $\sqrt{N}\mathrm{vec}(\widehat{K}(N) - K_\star) \overset{D}{\rightsquigarrow} \varphi$, then by the second order delta method:

$$N \cdot (J(\widehat{K}(N)) - J_\star) \overset{D}{\rightsquigarrow} (T-1)\sigma_w^2 \varphi^\mathsf{T} (I_n \otimes B_\star^\mathsf{T} B_\star) \varphi \,.$$

Next we make an intermediate calculation:

$$\begin{aligned}&\begin{bmatrix} -(I_n \otimes (B_\star^\dagger)^\mathsf{T}) \\ B_\star^\dagger A_\star \otimes (B_\star^\dagger)^\mathsf{T} \end{bmatrix} (I_n \otimes B_\star^\mathsf{T} B_\star) \begin{bmatrix} -(I_n \otimes B_\star^\dagger) & (B_\star^\dagger A_\star)^\mathsf{T} \otimes B_\star^\dagger \end{bmatrix} \\ &= \begin{bmatrix} I_n \otimes B_\star B_\star^\dagger & -((B_\star^\dagger A_\star)^\mathsf{T} \otimes B_\star B_\star^\dagger) \\ -(B_\star^\dagger A_\star \otimes B_\star B_\star^\dagger) & B_\star^\dagger A_\star A_\star^\mathsf{T} (B_\star^\dagger)^\mathsf{T} \otimes B_\star B_\star^\dagger \end{bmatrix} \\ &= \begin{bmatrix} I_n & -(B_\star^\dagger A_\star)^\mathsf{T} \\ -B_\star^\dagger A_\star & B_\star^\dagger A_\star A_\star^\mathsf{T} (B_\star^\dagger)^\mathsf{T} \end{bmatrix} \otimes B_\star B_\star^\dagger \,.\end{aligned}$$

Let $Z_N := N \cdot (J(\widehat{K}(N)) - J_\star)$. To conclude the proof, we show that the sequence $\{Z_N\}$ is uniformly integrable. Once we have the uniform integrability in place, then by Lemma A.5:

$$\begin{aligned}
&\lim_{N\to\infty} N \cdot (J(\widehat{K}(N)) - J_\star) \\
&= \sigma_w^4 \frac{T-1}{T} \operatorname{tr}\left(\left(\begin{bmatrix} P_\infty^{-1} & 0 \\ 0 & (1/\sigma_u^2) I_d \end{bmatrix} \otimes I_n\right)\left(\begin{bmatrix} I_n & -(B_\star^\dagger A_\star)^\mathsf{T} \\ -B_\star^\dagger A_\star & B_\star^\dagger A_\star A_\star^\mathsf{T} (B_\star^\dagger)^\mathsf{T} \end{bmatrix} \otimes B_\star B_\star^\dagger\right)\right) + o_T(1) \\
&= \sigma_w^4 \frac{T-1}{T} \operatorname{tr}\left(\begin{bmatrix} P_\infty^{-1} & 0 \\ 0 & (1/\sigma_u^2) I_d \end{bmatrix}\begin{bmatrix} I_n & -(B_\star^\dagger A_\star)^\mathsf{T} \\ -B_\star^\dagger A_\star & B_\star^\dagger A_\star A_\star^\mathsf{T} (B_\star^\dagger)^\mathsf{T} \end{bmatrix}\right) \operatorname{tr}(B_\star B_\star^\dagger) + o_T(1) \\
&= \sigma_w^4 \frac{T-1}{T}\left(\operatorname{tr}(P_\infty^{-1}) + \frac{\|B_\star^\dagger A_\star\|_F^2}{\sigma_u^2}\right) d + o_T(1) \,.
\end{aligned}$$

To conclude the proof, let $C_\star, \rho_\star$ be such that $\|A_\star^k\| \leq C_\star \rho_\star^k$ with $\rho_\star \in (0,1)$: these constants exist because $A_\star$ is stable. Now define the events:

$$\begin{aligned}
\mathcal{E}_{\mathsf{Alg}} &:= \{\rho(\widehat{A}(N)) \leq \varrho \,,\ \|\widehat{A}(N)\| \leq \zeta \,,\ \|\widehat{B}(N)\| \leq \psi \,,\ \sigma_d(\widehat{B}(N)) \geq \gamma\} \,, \\
\mathcal{E}_{\mathsf{Bdd}} &:= \{\|\widehat{A}(N) - A_\star\| \leq \frac{1-\rho_\star}{2C_\star} \,,\ \|\widehat{B}(N) - B_\star\| \leq \sigma_d(B_\star)/2\} \,.
\end{aligned}$$

Fix a finite $p \geq 1$. We write:

$$\begin{aligned}
\mathbb{E}[Z_N^p] &= N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c}] \\
&= N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \\
&\quad + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c}] \\
&= N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \\
&\quad + N^p (J(0) - J_\star)^p \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c) \\
&\leq N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] + N^p \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \\
&\quad + N^p (J(0) - J_\star)^p \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c) \,.
\end{aligned}$$

We now consider what happens on these three events. For the remainder of the proof, we let C denote a constant that depends on $n, d, p, C_\star, \rho_\star, \varrho, \zeta, \psi, \gamma, A_\star, B_\star, T, \varepsilon, \sigma_w^2, \sigma_u^2$ but not on N, whose value can change from line to line.

On the event $\mathcal{E}_{\mathsf{Bdd}}$. By a Taylor expansion we write:

$$h(\widehat{A}(N), \widehat{B}(N)) - h(A_\star, B_\star) = [D_{(A,B)} h(\tilde{A}, \tilde{B})]\left(\begin{bmatrix} \widehat{A}(N) - A_\star & \widehat{B}(N) - B_\star \end{bmatrix}\right) \,,$$

where $\tilde{A} = tA_\star + (1-t)\widehat{A}(N)$ and $\tilde{B} = tB_\star + (1-t)\widehat{B}(N)$ for some $t \in [0,1]$. Observe that on $\mathcal{E}_{\mathsf{Bdd}}$, we have that

$$\tilde{A}, \tilde{B} \in \mathcal{G} := \left\{(A,B) : \|A\| \leq \|A_\star\| + \frac{1-\rho_\star}{2C_\star} \,, \|B\| \leq \|B_\star\| + \sigma_d(B_\star)/2 \,, \sigma_d(B) \geq \sigma_d(B_\star)/2\right\} \,.$$

By Lemma B.1, each $(A,B) \in \mathcal{G}$ is stabilizable (since A is stable) and B has full column rank. Therefore by Lemma C.1, for any $(A,B) \in \mathcal{G}$ we have that $\mathsf{dare}(A, B, I_n, 0)$ has a unique positive definite solution and its derivative is well defined. By the compactness of $\mathcal{G}$ and the continuity of h and its derivative, we define the finite constants

$$C_K := \sup_{A,B \in \mathcal{G}} \|h(A,B)\| \:, \:\: C_{\mathrm{deriv}} := \sup_{A,B \in \mathcal{G}} \|[D_{(A,B)} h(A,B)]\| \:.$$

We can now Taylor expand $J(K)$ around $K_\star$ and obtain:

$$\begin{aligned} J(\widehat{K}(N)) - J_\star &= \frac{1}{2} \mathrm{Hess} J(\tilde{K})[\widehat{K}(N) - K_\star, \widehat{K}(N) - K_\star] \\ &\leq \frac{1}{2} \left(\sup_{\|\widetilde{K}\| \leq C_K + \|K_\star\|} \|\mathrm{Hess} J(\tilde{K})\| \right) \|\widehat{K}(N) - K_\star\|_F^2 \\ &\leq \frac{d}{2} \left(\sup_{\|\widetilde{K}\| \leq C_K + \|K_\star\|} \|\mathrm{Hess} J(\tilde{K})\| \right) C_{\mathrm{deriv}}^2 (\|\widehat{A}(N) - A_\star\|^2 + \|\widehat{B}(N) - B_\star\|^2) \:. \end{aligned}$$

Hence for N sufficiently large, by Lemma A.7 we have

$$\begin{aligned} N^p \cdot \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}}] &\leq C N^p (\mathbb{E}[\|\widehat{A}(N) - A_\star\|^{2p}] + \mathbb{E}[\|\widehat{B}(N) - B_\star\|^{2p}]) \\ &\leq C N^p (\frac{1}{N^p}) = C \:. \end{aligned}$$

On the event $\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}$. In this case, we use the bounds given by $\mathcal{E}_{\mathsf{Alg}}$ to bound the controller $\widehat{K}(N)$. Lemma C.1 ensures that the solution $\widehat{P} = \mathsf{dare}(\widehat{A}(N), \widehat{B}(N), I_n, 0)$ exists and satisfies $\widehat{P} \succeq I_n$. Let the finite constant C_P be $C_P := \sup_{\rho(A) \leq \varrho, \|A\| \leq \zeta, \|B\| \leq \psi, \sigma_d(B) \geq \gamma} \|\mathsf{dare}(A, B, I_n, 0)\|$. We can then bound $\|\widehat{K}(N)\|$ as follows. Dropping the indexing of N,

$$\|\widehat{K}\| = \|(\widehat{B}^\mathsf{T} \widehat{P} \widehat{B})^{-1} \widehat{B}^\mathsf{T} \widehat{P} \widehat{A}\| \leq \frac{1}{\sigma_{\min}(\widehat{B}^\mathsf{T} \widehat{P} \widehat{B})} \|\widehat{B}^\mathsf{T} \widehat{P} \widehat{A}\| \leq \frac{C_P \psi \zeta}{\gamma^2} \:.$$

Therefore:

$$N^p \cdot \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \leq N^p \cdot \left(\sup_{\|K\| \leq \frac{C_P \psi \zeta}{\gamma^2}} (J(K) - J_\star)^p \right) \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c) \leq C N^p \mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c) \:.$$

By Lemma A.7, we can choose N large enough such that $\mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c) \leq 1/N^p$ so that $N^p \cdot \mathbb{E}[(J(\widehat{K}(N)) - J_\star)^p \mathbf{1}_{\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}}] \leq C$.

On the event $\mathcal{E}_{\mathsf{Bdd}}^c \cap \mathcal{E}_{\mathsf{Alg}}^c$. This case is simple. We simply invoke Lemma A.7 to choose an N large enough such that $\mathbb{P}(\mathcal{E}_{\mathsf{Bdd}}^c) \leq 1/(N(J(0) - J_\star))^p$.

Putting it together. If we take N as the maximum over the three cases described above, we have hence shown that for all N greater than this constant:

$$\mathbb{E}[Z_N^p] \leq C \:.$$

This shows the desired uniform integrability condition for Z_N. The asymptotic bound now follows from Lemma A.5.

C.3. Proof of Theorem 2.5

The proof works by applying Lemma A.8 with the function $F(\theta) = J_\Sigma(K)$ with $\Sigma = \sigma_u^2 B_\star B_\star^\mathsf{T} + \sigma_w^2 I_n$ and $G(\theta) = J(K)$. We first need to verify the hypothesis of the lemma. We define the convex domain Θ as $\Theta = \{K \in \mathbb{R}^{d\times n} : \|K\| \le \zeta\}$. Note that $K_\star$ is in the interior of Θ, since we assume that $\|K_\star\| < \zeta$. Recall that the policy gradient $g(K;\xi)$ is:

$$g(K;\xi) = \frac{1}{\sigma_u^2}\sum_{t=1}^{T-1} \eta_t x_t^\mathsf{T} \Psi_t \ , \ \ \xi = (\eta_0, w_0, \eta_1, w_1, ..., \eta_{T-1}, w_{T-1}) \ .$$

It is clear that x_t is a polynomial in (K,ξ). Furthermore, all three of the Ψ_t's we study are also polynomials in (K,ξ). Hence $[D_K g(K;\xi)]$ is a matrix with entries that are polynomial in (K,ξ). Therefore, for every ξ, for all fixed $K_1, K_2 \in \Theta$,

$$\|g(K_1;\xi) - g(K_2;\xi)\|_F \le \sup_{K\in\Theta} \|[D_K g(K;\xi)]\|_F \|K_1 - K_2\|_F \ .$$

Hence squaring and taking expectations,

$$\mathbb{E}_\xi[\|g(K_1;\xi) - g(K_2;\xi)\|_F^2] \le \mathbb{E}_\xi\left[\sup_{K\in\Theta} \|[D_K g(K;\xi)]\|_F^2\right] \|K_1 - K_2\|_F^2 \ .$$

We can now define the constant $L := \mathbb{E}_\xi\left[\sup_{K\in\Theta}\|[D_K g(K;\xi)]\|_F^2\right]$. To see that this quantity L is finite, observe that $\|[D_K g(K;\xi)]\|_F^2$ is a polynomial of ξ with coefficients given by K (and $A_\star, B_\star$). Since K lives in a compact set Θ, these coefficients are uniformly bounded and hence the their moments are bounded. In Section C.1, we showed that the function $J_\Sigma(K)$ satisfies the $\mathsf{RSC}(m,\Theta)$ condition with $m = 2(T-1)\sigma_w^2 \sigma_{\min}(B_\star)^2$. Also it is clear that the high probability bound on $\|g(K;\xi)\|_F$ can be achieved by standard Gaussian concentration results. Hence by Lemma A.8, and in particular Equation A.8,

$$\begin{aligned}\liminf_{N\to\infty} N \cdot \mathbb{E}[J(\widehat{K}) - J_\star] &\ge \frac{\mathbb{E}_\xi[\|g(K_\star;\xi)\|_F^2]}{8(T-1)\sigma_w^2\sigma_{\min}(B_\star)^2 \lambda_{\max}((\nabla^2 J(K_\star))^{-1}(\nabla^2 J_\Sigma(K_\star) - \frac{m}{2} I_{nd}))} \\ &= \frac{\mathbb{E}_\xi[\|g(K_\star;\xi)\|_F^2]}{8(T-1)\sigma_{\min}(B_\star)^2(\sigma_w^2 + \sigma_u^2\|B_\star\|^2)} \ . \end{aligned} \tag{C.3}$$

Above, the inequality holds since we have that,

$$\begin{aligned}&\nabla^2 J(K_\star) = 2(T-1)(\sigma_w^2 I_n \otimes B_\star^\mathsf{T} B_\star) \ , \\ &\nabla^2 J_\Sigma(K_\star) = 2(T-1)((\sigma_w^2 I_n + \sigma_u^2 B_\star B_\star^\mathsf{T}) \otimes B_\star^\mathsf{T} B_\star) = \nabla^2 J(K_\star) + 2(T-1)\sigma_u^2(B_\star B_\star^\mathsf{T} \otimes B_\star^\mathsf{T} B_\star) \ ,\end{aligned}$$

and therefore,

$$\begin{aligned}(\nabla^2 J(K_\star))^{-1}(\nabla^2 J_\Sigma(K_\star) - \frac{m}{2} I_{nd}) &= I_{nd} + \frac{\sigma_u^2}{\sigma_w^2}(B_\star B_\star^\mathsf{T} \otimes I_d) - \frac{\sigma_{\min}(B_\star)^2}{2}(I_n \otimes (B_\star^\mathsf{T} B_\star)^{-1}) \\ &\preceq I_{nd} + \frac{\sigma_u^2}{\sigma_w^2}(B_\star B_\star^\mathsf{T} \otimes I_d) \ .\end{aligned}$$

The remainder of the proof is to estimate the quantity $\mathbb{E}_\xi[\|g(K_\star;\xi)\|_F^2]$. Note that at $K = K_\star$, $x_t = B_\star \eta_{t-1} + w_{t-1}$ since the dynamics are cancelled out. Define $c_{t\to T} := \sum_{\ell=t}^T \|x_\ell\|_2^2$. At

$K = K_\star$, we have $c_{t \to T} = \sum_{\ell=t-1}^{T-1} \|B_\star \eta_\ell + w_\ell\|_2^2$. Observe that we have for $t_2 > t_1$, for any h that depends on only $(\eta_{t_1}, w_{t_1}, \eta_{t_1+1}, w_{t_1+1}, ...)$:

$$\begin{aligned}\mathbb{E}[\langle \eta_{t_1}, \eta_{t_2} \rangle \langle x_{t_1}, x_{t_2} \rangle h] &= \mathbb{E}[\langle \eta_{t_1}, \eta_{t_2} \rangle (\langle B_\star \eta_{t_1-1}, B_\star \eta_{t_2-1} \rangle + \langle w_{t_1-1}, w_{t_2-1} \rangle \\ &\quad + \langle B_\star \eta_{t_1-1}, w_{t_2-1} \rangle + \langle B_\star \eta_{t_2-1}, w_{t_1-1} \rangle) h] \\ &= 0\,.\end{aligned}$$

As a consequence, we have that as long as Ψ_t only depends on $(\eta_t, w_t, \eta_{t+1}, w_{t+1}, ...)$:

$$\begin{aligned}\mathbb{E}[\|g(K;\xi)\|_F^2] &= \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \mathbb{E}[\|\eta_t\|_2^2 \|x_t\|_2^2 \Psi_t^2] + \frac{2}{\sigma_u^4} \sum_{t_2 > t_1 = 1}^{T-1} \mathbb{E}[\langle \eta_{t_1}, \eta_{t_2} \rangle \langle x_{t_1}, x_{t_2} \rangle \Psi_{t_1} \Psi_{t_2}] \\ &= \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \mathbb{E}[\|\eta_t\|_2^2 \|x_t\|_2^2 \Psi_t^2]\,.\end{aligned}$$

C.3.1. SIMPLE BASELINE

Recall that the simple baseline is to set $b_t(x_t; K) = \|x_t\|_2^2$. Hence, the policy gradient estimate simplifies to:

$$g(K;\xi) = \frac{1}{\sigma_u^2} \sum_{t=1}^{T-1} \eta_t x_t^\mathsf{T} c_{t+1 \to T}\,.$$

Since we have that $c_{t+1 \to T}$ at optimality only depends only on $(\eta_t, w_t, \eta_{t+1}, w_{t+1}, ...)$, we compute $\mathbb{E}[\|g(K_\star;\xi)\|_F^2]$ as follows:

$$\begin{aligned}&\mathbb{E}[\|g(K_\star;\xi)\|_F^2] = \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \mathbb{E}[\|\eta_t\|_2^2 \|x_t\|_2^2 c_{t+1 \to T}^2] \\ &= \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \mathbb{E}\Bigg[\|\eta_t\|_2^2 \|B_\star \eta_{t-1} + w_{t-1}\|_2^2 \\ &\qquad \times \left(\sum_{\ell=t}^{T-1} \|B_\star \eta_\ell + w_\ell\|_2^4 + 2 \sum_{\ell_2 > \ell_1 = t}^{T-1} \|B_\star \eta_{\ell_1} + w_{\ell_1}\|_2^2 \|B_\star \eta_{\ell_2} + w_{\ell_2}\|_2^2 \right) \Bigg] \\ &= \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \mathbb{E}[\|B_\star \eta_{t-1} + w_{t-1}\|_2^2 \|\eta_t\|_2^2 \|B_\star \eta_t + w_t\|_2^4] \\ &\qquad + \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \sum_{\ell=t+1}^{T-1} \mathbb{E}[\|B_\star \eta_{t-1} + w_{t-1}\|_2^2 \|\eta_t\|_2^2 \|B_\star \eta_\ell + w_\ell\|_2^4] \\ &\qquad + \frac{2}{\sigma_u^4} \sum_{t=1}^{T-1} \sum_{\ell_2 > t}^{T-1} \mathbb{E}[\|B_\star \eta_{t-1} + w_{t-1}\|_2^2 \|\eta_t\|_2^2 \|B_\star \eta_t + w_t\|_2^2 \|B_\star \eta_{\ell_2} + w_{\ell_2}\|_2^2] \\ &\qquad + \frac{2}{\sigma_u^4} \sum_{t=1}^{T-1} \sum_{\ell_2 > \ell_1 = t+1}^{T-1} \mathbb{E}[\|B_\star \eta_{t-1} + w_{t-1}\|_2^2 \|\eta_t\|_2^2 \|B_\star \eta_{\ell_1} + w_{\ell_1}\|_2^2 \|B_\star \eta_{\ell_2} + w_{\ell_2}\|_2^2]\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{\sigma_u^4} \sum_{t=1}^{T-1} \sum_{\ell_2 > \ell_1 = t+1}^{T-1} \mathbb{E}[\|B_\star \eta_{t-1} + w_{t-1}\|_2^2 \|\eta_t\|_2^2 \|B_\star \eta_{\ell_1} + w_{\ell_1}\|_2^2 \|B_\star \eta_{\ell_2} + w_{\ell_2}\|_2^2] + o(T^3) \\
&= \frac{2}{\sigma_u^4} \sum_{t=1}^{T-1} \sum_{\ell_2 > \ell_1 = t+1}^{T-1} \sigma_u^2 d (\mathbb{E}[\|B_\star \eta_0 + w_0\|_2^2])^3 + o(T^3) \\
&\asymp T^3 \frac{1}{\sigma_u^2} d (\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n)^3 + o(T^3) \,.
\end{aligned}
$$

C.3.2. VALUE FUNCTION BASELINE

Recall that the value function at time t for a particular policy K is defined as:

$$
V_t^K(x) = \mathbb{E}\left[\sum_{\ell=t}^{T} \|x_\ell\|_2^2 \middle| x_t = x \right] \,.
$$

We now consider policy gradient with the value function baseline $b_t(x_t; K) = V_t^K(x_t)$:

$$
g(K;\xi) = \frac{1}{\sigma_u^2} \sum_{t=1}^{T-1} \eta_t x_t^\mathsf{T} (c_{t\to T} - V_t^K(x_t)) \,.
$$

Recalling that under $K_\star$ the dynamics are cancelled out, we readily compute:

$$
V_t^{K_\star}(x) = \|x\|_2^2 + (T-t)(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n) \,.
$$

Therefore:

$$
g(K_\star;\xi) = \frac{1}{\sigma_u^2} \sum_{t=1}^{T-1} \eta_t x_t^\mathsf{T} (c_{t+1\to T} - (T-t)(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n)) \,.
$$

Define $\beta := \sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n$. We compute the variance as:

$$
\begin{aligned}
\mathbb{E}[\|g(K_\star;\xi)\|_F^2] &= \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \mathbb{E}\Bigg[\|\eta_t\|_2^2 \|B_\star \eta_{t-1} + w_{t-1}\|_2^2 \\
&\quad \times \left(\sum_{\ell=t}^{T-1} (\|B_\star \eta_\ell + w_\ell\|_2^2 - \beta)^2 + 2 \sum_{\ell_2 > \ell_1 = t}^{T-1} (\|B_\star \eta_{\ell_1} + w_{\ell_1}\|_2^2 - \beta)(\|B_\star \eta_{\ell_2} + w_{\ell_2}\|_2^2 - \beta) \right) \Bigg] \\
&= \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \mathbb{E}[\|\eta_t\|_2^2 \|B_\star \eta_{t-1} + w_{t-1}\|_2^2 (\|B_\star \eta_t + w_t\|_2^2 - \beta)^2] \\
&\quad + \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \sum_{\ell=t+1}^{T-1} \mathbb{E}[\|\eta_t\|_2^2 \|B_\star \eta_{t-1} + w_{t-1}\|_2^2 (\|B_\star \eta_\ell + w_\ell\|_2^2 - \beta)^2] \\
&= \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \sum_{\ell=t+1}^{T-1} \mathbb{E}[\|\eta_t\|_2^2 \|B_\star \eta_{t-1} + w_{t-1}\|_2^2 (\|B_\star \eta_\ell + w_\ell\|_2^2 - \beta)^2] + o(T^2)
\end{aligned}
$$

$$\asymp T^2 \frac{d}{\sigma_u^2}(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n)(\mathbb{E}[\|B_\star \eta_\ell + w_\ell\|_2^4] - \beta^2) + o(T^2)$$
$$\overset{(a)}{\asymp} T^2 \frac{d}{\sigma_u^2}(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n)(\sigma_u^4 \|B_\star^\mathsf{T} B_\star\|_F^2 + \sigma_w^4 n + \sigma_w^2 \sigma_u^2 \|B_\star\|_F^2) + o(T^2) \,,$$

Above, (a) follows because:

$$\mathbb{E}[\|B_\star \eta_\ell + w_\ell\|_2^4] = 2(\sigma_u^4 \|B_\star^\mathsf{T} B_\star\|_F^2 + \sigma_w^4 n + 2\sigma_w^2 \sigma_u^2 \|B_\star\|_F^2) + (\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n)^2 \,.$$

C.3.3. IDEAL ADVANTAGE BASELINE

Let us first compute $Q_t^{K_\star}(x_t, u_t)$. Under $K_\star$, $x_{\ell+1} = B_\star \eta_\ell + w_\ell$. So we have:

$$\begin{aligned} Q_t^{K_\star}(x_t, u_t) &= \|x_t\|_2^2 + \mathbb{E}_{w_t}[\|A_\star x_t + B_\star u_t + w_t\|_2^2] + (T-t-1)(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n) \\ &= \|x_t\|_2^2 + \|A_\star x_t + B_\star u_t\|_2^2 + \sigma_w^2 n + (T-t-1)(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n) \,. \end{aligned}$$

Recalling that $V_t^{K_\star}(x) = \|x\|_2^2 + (T-t)(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n)$,

$$A_t^{K_\star}(x_t, u_t) = Q_t^{K_\star}(x_t, u_t) - V_t^{K_\star}(x_t) = \|A_\star x_t + B_\star u_t\|_2^2 - \sigma_u^2 \|B_\star\|_F^2 \,.$$

Therefore, if $u_t = K_\star x_t + \eta_t$, we have $A_t^{K_\star}(x_t, u_t) = \|B_\star \eta_t\|_2^2 - \sigma_u^2 \|B_\star\|_F^2$. Since $A_t^{K_\star}(x_t, u_t)$ depends only on η_t,

$$\begin{aligned} \mathbb{E}[\|g(K_\star;\xi)\|_F^2] &= \frac{1}{\sigma_u^4} \sum_{t=1}^{T-1} \mathbb{E}[\|\eta_t\|_2^2 \|x_t\|_2^2 (\|B_\star \eta_t\|_2^2 - \sigma_u^2 \|B_\star\|_F^2)^2] \\ &= \frac{1}{\sigma_u^4}(T-1)(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n) \mathbb{E}[\|\eta_1\|_2^2 (\|B_\star \eta_1\|_2^2 - \sigma_u^2 \|B_\star\|_F^2)^2] \,. \end{aligned}$$

We have that $\mathbb{E}[\|\eta_1\|_2^2] = \sigma_u^2 d$, $\mathbb{E}[\|B_\star \eta_1\|_2^2 \|\eta_1\|_2^2] = \sigma_u^4 (d+2) \|B_\star\|_F^2$, and $\mathbb{E}[\|B_\star \eta_1\|_2^4 \|\eta_1\|_2^2] = \sigma_u^6((d+4)\|B_\star\|_F^4 + (2d+8)\|B_\star^\mathsf{T} B_\star\|_F^2)$ (this can be computed using Lemma D.2). Hence,

$$\begin{aligned} &\mathbb{E}[\|\eta_1\|_2^2 (\|B_\star \eta_1\|_2^2 - \sigma_u^2 \|B_\star\|_F^2)^2] \\ &\quad= \mathbb{E}[\|B_\star \eta_1\|_2^4 \|\eta_1\|_2^2 + \sigma_u^4 \|B_\star\|_F^4 \|\eta_1\|_2^2 - 2\sigma_u^2 \|B_\star\|_F^2 \|B_\star \eta_1\|_2^2 \|\eta_1\|_2^2] \\ &\quad= \sigma_u^6((d+4)\|B_\star\|_F^4 + (2d+8)\|B_\star^\mathsf{T} B_\star\|_F^2) + \sigma_u^6 \|B_\star\|_F^4 d - 2\sigma_u^6 \|B_\star\|_F^4 (d+2) \\ &\quad= \sigma_u^6 (2d+8) \|B_\star^\mathsf{T} B_\star\|_F^2 \,. \end{aligned}$$

Therefore,

$$\mathbb{E}[\|g(K_\star;\xi)\|_F^2] \asymp T(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n) \sigma_u^2 d \|B_\star^\mathsf{T} B_\star\|_F^2 \,.$$

C.3.4. PUTTING IT TOGETHER

Combining Equation (C.3) with the calculations for $\mathbb{E}_\xi[\|g(K_\star;\xi)\|_F^2]$, we obtain:

$$\liminf_{N\to\infty} N \cdot \mathbb{E}[J(\widehat{K}_{\mathrm{pg}}(N)) - J_\star] \gtrsim \frac{1}{\sigma_d(B_\star)^2 (\sigma_w^2 + \sigma_u^2 \|B_\star\|^2)} \times$$
$$\begin{cases} T^2 \frac{d}{\sigma_u^2}(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n)^3 + o(T^2) & \text{(Simple baseline)} \\ T \frac{d}{\sigma_u^2}(\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n)(\sigma_u^4 \|B_\star^\mathsf{T} B_\star\|_F^2 + \sigma_w^4 n + \sigma_w^2 \sigma_u^2 \|B_\star\|_F^2) + o(T) & \text{(Value function baseline)} \\ (\sigma_u^2 \|B_\star\|_F^2 + \sigma_w^2 n) \sigma_u^2 d \|B_\star^\mathsf{T} B_\star\|_F^2 & \text{(Advantage baseline)} \end{cases} ,$$

from which Theorem 2.5 follows.

C.4. Proof of Theorem 2.6

Our proof is inspired by lower bounds for the query complexity of derivative-free optimization of stochastic optimization (see e.g. Jamieson et al. (2012)).

Recall from (C.1) that the function $J(K)$ satisfies the quadratic growth condition $J(K) - J_\star \geq (T-1)\rho^2\sigma_w^2\|K - K_\star\|_F^2$. Therefore for any $\vartheta > 0$,

$$\begin{aligned}
&\inf_{\widehat{K}} \sup_{(A_\star,B_\star)\in\mathscr{G}(\rho,d)} \mathbb{E}[J(\widehat{K}) - J_\star] \\
&\geq \inf_{\widehat{K}} \sup_{(A_\star,B_\star)\in\mathscr{G}(\rho,d)} (T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \mathbb{P}(J(\widehat{K}) - J_\star \geq (T-1)\rho^2\sigma_w^2\vartheta^2) \\
&\geq \inf_{\widehat{K}} \sup_{(A_\star,B_\star)\in\mathscr{G}(\rho,d)} (T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \mathbb{P}((T-1)\rho^2\sigma_w^2\|(-U_\star^\mathsf{T}) - \widehat{K}\|_F^2 \geq (T-1)\rho^2\sigma_w^2\vartheta^2) \\
&= \inf_{\widehat{K}} \sup_{(A_\star,B_\star)\in\mathscr{G}(\rho,d)} (T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \mathbb{P}(\|(-U_\star^\mathsf{T}) - \widehat{K}\|_F \geq \vartheta) \,.
\end{aligned}$$

Above, the first inequality is Markov's inequality and the second is the quadratic growth condition.

We first state a result regarding the packing number of $O(n,d)$, which we define as:

$$O(n,d) := \{U \in \mathbb{R}^{n\times d} : U^\mathsf{T} U = I_d\} \,.$$

Lemma C.3 *Let $\delta > 0$, and suppose that $d \leq n/2$. We have that the packing number M of $O(n,d)$ in the Frobenius norm $\|\cdot\|_F$ satisfies*

$$M(O(n,d), \|\cdot\|_F, \delta d^{1/2}) \geq \left(\frac{c}{\delta}\right)^{d(n-d)} \,,$$

where $c > 0$ is a universal constant.

Proof Let $G_{n,d}$ denote the Grassman manifold of d-dimensional subspaces of $\mathbb{R}^n$. For two subspaces $E, F \in G_{n,d}$, equip $G_{n,d}$ with the metric $\rho(E,F) = \|P_E - P_F\|_F$, where P_E, P_F are the projection matrices onto E, F respectively. Proposition 8 of Pajor (1998) tells us that the covering number $N(G_{n,d}, \rho, \delta d^{1/2}) \geq \left(\frac{c}{\delta}\right)^{d(n-d)}$. But since $M(G_{n,d}, \rho, \delta d^{1/2}) \geq N(G_{n,d}, \rho, \delta d^{1/2})$, this gives us a lower bound on the packing number of $G_{n,d}$. Now for every $E \in G_{n,d}$ we can associate a matrix $E_1 \in O(n,d)$ such that $\operatorname{span}(E_1) = E$. The projector P_E is simply $P_E = E_1 E_1^\mathsf{T}$. Now let $E, F \in G_{n,d}$ and observe the inequality,

$$\|P_E - P_F\|_F = \|E_1 E_1^\mathsf{T} - F_1 F_1^\mathsf{T}\|_F \leq 2\|E_1 - F_1\|_F \,.$$

Hence a packing of $G_{n,d}$ also yields a packing of $O(n,d)$ up to constant factors. ■

Now letting $U_1, ..., U_M$ be a 2ϑ-separated set we have by the standard reduction to multiple hypothesis testing that that the risk is lower bounded by:

$$(T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \inf_{\widehat{V}} \mathbb{P}(\widehat{V} \neq V) \geq (T-1)\rho^2\sigma_w^2\vartheta^2 \cdot \left(1 - \frac{I(V;Z) + \log 2}{\log M}\right) \,. \tag{C.4}$$

where V is a uniform index over $\{1, ..., M\}$ and the inequality is Fano's inequality.

Now we can proceed as follows. First, we let $U_1, ..., U_M$ be elements of $O(n,d)$ that form a $2\vartheta \asymp \sqrt{d}$ packing in the $\|\cdot\|_F$ norm. We know we can let $M \geq e^{d(n-d)}$ by Lemma C.3. Each U_i induces a covariance $\Sigma_i = \sigma_w^2 I_n + \rho^2 \sigma_u^2 U_i U_i^\mathsf{T} \preceq (\sigma_w^2 + \rho^2\sigma_u^2) I_n$. Furthermore, the closed-loop L_i given by playing a feedback matrix K that satisfies $\|K\| \leq 1$ is:

$$L_i = \rho U_i (U_i + K^\mathsf{T})^\mathsf{T} \,.$$

It is clear that $\|L_i\| \leq 2\rho$ and hence if $\rho < 1/2$ then this system is stable. Furthermore, we have that $\mathrm{rank}(L_i) \leq d$. With this, we can control:

$$\begin{aligned}
\mathrm{tr}(\mathbb{E}[x_t x_t^\mathsf{T}]) &= \mathrm{tr}\left(\sum_{\ell=0}^{t-1} L_i^\ell \Sigma_i (L_i^\ell)^\mathsf{T}\right) \leq (\sigma_w^2 + \rho^2\sigma_u^2) \sum_{\ell=0}^{t-1} \|L_i^\ell\|_F^2 \\
&\leq d(\sigma_w^2 + \rho^2\sigma_u^2) \sum_{\ell=0}^{t-1} \|L_i^\ell\|^2 \leq \frac{d(\sigma_w^2+\rho^2\sigma_u^2)}{1-(2\rho)^2} \,.
\end{aligned}$$

Hence for one trajectory $Z = (x_0, u_0, x_1, u_1, ..., x_{T-1}, u_{T-1}, x_T)$, conditioned on a particular K,

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{i|K}, \mathbb{P}_{j|K}) &\leq \sum_{t=0}^{T-1} \frac{1}{2\sigma_w^2} \mathbb{E}_{x_t \sim \mathbb{P}_{i|K}}[\|(L_i - L_j)x_t\|^2] \\
&\leq \frac{8\rho^2}{\sigma_w^2} \sum_{t=0}^{T-1} \mathrm{tr}(\mathbb{E}[x_t x_t^\mathsf{T}]) \\
&\leq \frac{8(\sigma_w^2 + \rho^2\sigma_u^2)\rho^2 T d}{\sigma_w^2(1-(2\rho)^2)} \,.
\end{aligned}$$

This allows us to bound the KL between the distributions involving all the iterations as:

$$\mathrm{KL}(\mathbb{P}_i, \mathbb{P}_j) = \sum_{\ell=1}^{N} \mathbb{E}_{K_\ell \sim \mathbb{P}_i}[\mathrm{KL}(\mathbb{P}_{i|K_\ell}, \mathbb{P}_{j|K_\ell})] \leq \frac{8(\sigma_w^2 + \rho^2\sigma_u^2)\rho^2 NTd}{\sigma_w^2(1-(2\rho)^2)} \,.$$

Assuming $d(n-d)$ is greater than an absolute constant, we can set ρ to be (recall we have N different rollouts):

$$\rho^2 \asymp \frac{\sigma_w^2}{\sigma_w^2 + \sigma_u^2} \frac{n-d}{TN} \,,$$

and bound $\frac{I(V;Z)+\log 2}{\log M} \leq 1/2$. The result now follows from plugging in our choice of ρ into (C.4).

Appendix D. Deferred Proofs for Asymptotic Toolbox

Our main limit theorem is the following CLT for ergodic Markov chains.

Theorem D.1 (Corollary 2 of Jones (2004)) *Suppose that $\{x_t\}_{t=0}^\infty \subseteq X$ is a geometrically ergodic (Harris) Markov chain with stationary distribution π. Let $f : X \to \mathbb{R}$ be a Borel function. Suppose that $\mathbb{E}_\pi[|f|^{2+\delta}] < \infty$ for some $\delta > 0$. Then for any initial distribution, we have:*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_\pi[f(x)]\right) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \sigma_f^2) \,,$$

where

$$\sigma_f^2 := \mathrm{Var}_\pi(f(x_0)) + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\pi(f(x_0), f(x_i)) \,.$$

D.1. Proof of Lemma A.1

Proof Let $X \in \mathbb{R}^{T\times n}$ be the data matrix with rows $(x_0, ..., x_{T-1})$ and $W \in \mathbb{R}^{T\times n}$ be the noise matrix with rows $(w_0, ..., w_{T-1})$. We write:

$$\widehat{L}(T) - L_\star = -\lambda L_\star (X^\mathsf{T} X + \lambda I_n)^{-1} + W^\mathsf{T} X (X^\mathsf{T} X + \lambda I_n)^{-1} \,.$$

Using the fact that $\mathrm{vec}(AXB) = (B^\mathsf{T} \otimes A)\mathrm{vec}(X)$,

$$\sqrt{T}\mathrm{vec}(\widehat{L}(T) - L_\star) = -\sqrt{T}\mathrm{vec}(\lambda L_\star (X^\mathsf{T} X + \lambda I_n)^{-1}) + ((T^{-1} X^\mathsf{T} X)^{-1} \otimes I_n)\mathrm{vec}(T^{-1/2} W^\mathsf{T} X) \,.$$

It is well-known that $\{x_t\}$ is geometrically ergodic (see e.g. Mokkadem (1988)), and therefore the augmented Markov chain $\{(x_t, w_t)\}$ is geometrically ergodic as well. By Theorem D.1 combined with the Cramér-Wold theorem we conclude:

$$\mathrm{vec}(T^{-1/2} W^\mathsf{T} X) = T^{-1/2} \sum_{t=1}^{T} \mathrm{vec}(w_t x_t^\mathsf{T}) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \mathbb{E}_{x\sim\nu_\infty, w}[\mathrm{vec}(wx^\mathsf{T})\mathrm{vec}(wx^\mathsf{T})^\mathsf{T}]) \,.$$

Above, we let ν_∞ denote the stationary distribution of $\{x_t\}$. We note that the cross-correlation terms disappear in the asymptotic covariance due to the martingale difference property of $\sum_{t=0}^{T-1} w_t x_t^\mathsf{T}$. We now use the identity $\mathrm{vec}(wx^\mathsf{T}) = (x \otimes I_n)w$ and compute

$$\begin{aligned} \mathbb{E}_{x\sim\nu_\infty, w}[\mathrm{vec}(wx^\mathsf{T})\mathrm{vec}(wx^\mathsf{T})^\mathsf{T}] &= \mathbb{E}_{x\sim\nu_\infty, w}[(x \otimes I_n) w w^\mathsf{T} (x^\mathsf{T} \otimes I_n)] \\ &= \sigma_w^2 \mathbb{E}_{x\sim\nu_\infty}[(x \otimes I_n)(x^\mathsf{T} \otimes I_n)] \\ &= \sigma_w^2 \mathbb{E}_{x\sim\nu_\infty}[(x x^\mathsf{T} \otimes I_n)] \\ &= \sigma_w^2 (P_\infty \otimes I_n) \,. \end{aligned}$$

We have that $T^{-1} X^\mathsf{T} X \xrightarrow{\text{a.s.}} P_\infty$ by the ergodic theorem. Therefore by the continuous mapping theorem followed by Slutsky's theorem, we have that

$$((T^{-1} X^\mathsf{T} X)^{-1} \otimes I_n)\mathrm{vec}(T^{-1/2} W^\mathsf{T} X) \overset{D}{\rightsquigarrow} \mathcal{N}(0, \sigma_w^2 (P_\infty^{-1} \otimes I_n)) \,.$$

On the other hand, we have:

$$\sqrt{T}\mathrm{vec}(\lambda L_\star (X^\mathsf{T} X + \lambda I_n)^{-1}) = \frac{1}{\sqrt{T}} \mathrm{vec}(\lambda L_\star (T^{-1} X^\mathsf{T} X + T^{-1}\lambda I_n)^{-1}) \xrightarrow{\text{a.s.}} 0 \,.$$

The claim now follows by another application of Slutsky's theorem. ■

D.2. Proof of Lemma A.2

Proof Let $Z^{(i)} \in \mathbb{R}^{T\times(n+d)}$ be a data matrix with the rows $(z_0^{(i)}, ..., z_{T-1}^{(i)})$, and let $W^{(i)} \in \mathbb{R}^{T\times n}$ be the noise matrix with the rows $(w_0^{(i)}, ..., w_{T-1}^{(i)})$. With this notation we write:

$$
\begin{aligned}
\widehat{\Theta}(N) - \Theta_\star &= \left(\sum_{i=1}^N \frac{1}{T}\sum_{t=0}^{T-1} z_{t+1}^{(i)}(z_t^{(i)})^\mathsf{T}\right)\left(\sum_{i=1}^N \frac{1}{T}\sum_{t=0}^{T-1} z_t^{(i)}(z_t^{(i)})^\mathsf{T} + \lambda I_{n+d}\right)^{-1} - \Theta_\star \\
&= \Theta_\star\left(\sum_{i=1}^N \frac{1}{T}(Z^{(i)})^\mathsf{T} Z^{(i)}\right)\left(\sum_{i=1}^N \frac{1}{T}(Z^{(i)})^\mathsf{T} Z^{(i)} + \lambda I_{n+d}\right)^{-1} - \Theta_\star \\
&\quad + \left(\sum_{i=1}^N \frac{1}{T}(W^{(i)})^\mathsf{T} Z^{(i)}\right)\left(\sum_{i=1}^N \frac{1}{T}(Z^{(i)})^\mathsf{T} Z^{(i)} + \lambda I_{n+d}\right)^{-1} \\
&= -\lambda\Theta_\star\left(\sum_{i=1}^N \frac{1}{T}(Z^{(i)})^\mathsf{T} Z^{(i)} + \lambda I_{n+d}\right)^{-1} \\
&\quad + \left(\sum_{i=1}^N \frac{1}{T}(W^{(i)})^\mathsf{T} Z^{(i)}\right)\left(\sum_{i=1}^N \frac{1}{T}(Z^{(i)})^\mathsf{T} Z^{(i)} + \lambda I_{n+d}\right)^{-1} \\
&=: G_1(N) + G_2(N)\,.
\end{aligned}
$$

Taking vec of $G_2(N)$:

$$
\mathrm{vec}(G_2(N)) = \left(\left(\frac{1}{N}\sum_{i=1}^N \frac{1}{T}(Z^{(i)})^\mathsf{T} Z^{(i)} + \frac{\lambda}{N} I_{n+d}\right)^{-1} \otimes I_n\right)\mathrm{vec}\left(\frac{1}{N}\sum_{i=1}^N \frac{1}{T}\sum_{t=0}^{T-1} w_t^{(i)}(z_t^{(i)})^\mathsf{T}\right)\,.
$$

Now we write $\mathrm{vec}(w_t z_t^\mathsf{T}) = (z_t \otimes I_n)w_t$ and hence

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{vec}\left(\frac{1}{T}\sum_{t=0}^{T-1} w_t z_t^\mathsf{T}\right)\mathrm{vec}\left(\frac{1}{T}\sum_{t=0}^{T-1} w_t z_t^\mathsf{T}\right)^\mathsf{T}\right] &= \frac{1}{T^2}\sum_{t_1,t_2=0}^{T-1} \mathbb{E}[(z_{t_1}\otimes I_n)w_{t_1}w_{t_2}^\mathsf{T}(z_{t_2}^\mathsf{T}\otimes I_n)] \\
&= \frac{\sigma_w^2}{T^2}\sum_{t=0}^{T-1}\mathbb{E}[z_t z_t^\mathsf{T}] \otimes I_n\,.
\end{aligned}
$$

We have that:

$$
\frac{1}{N}\sum_{i=1}^N \frac{1}{T}(Z^{(i)})^\mathsf{T} Z^{(i)} + \frac{\lambda}{N} I_{n+d} \xrightarrow{\text{a.s.}} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[z_t z_t^\mathsf{T}]\,.
$$

Hence by the central limit theorem combined with the continuous mapping theorem and Slutsky's theorem,

$$\begin{aligned}
\sqrt{N}\mathrm{vec}(G_1(N)) &\xrightarrow{\text{a.s.}} 0 \,, \\
\sqrt{N}\mathrm{vec}(G_2(N)) &\overset{D}{\rightsquigarrow} \mathcal{N}\left(0, \frac{\sigma_w^2}{T}\left[\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[z_t z_t^\mathsf{T}]\right]^{-1} \otimes I_n\right) \\
&= \mathcal{N}\left(0, \frac{\sigma_w^2}{T}\begin{bmatrix} [\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[x_t x_t^\mathsf{T}]]^{-1} & 0 \\ 0 & (1/\sigma_u^2) I_d \end{bmatrix} \otimes I_n\right) \,.
\end{aligned}$$

To finish the proof, we note that $\mathbb{E}[x_t x_t^\mathsf{T}] = \sum_{\ell=0}^{t-1} A_\star^\ell M (A_\star^\ell)^\mathsf{T} := P_t$ with $M := \sigma_u^2 B_\star B_\star^\mathsf{T} + \sigma_w^2 I_n$ and $P_0 = 0$ (since $x_0 = 0$). Since $A_\star$ is stable, there exists a $\rho \in (0,1)$ and $C > 0$ such that $\|A_\star^k\| \le C\rho^k$ for all $k \ge 0$. Hence,

$$\|P_\infty - P_t\| = \left\|\sum_{\ell=t}^{\infty} A_\star^\ell M (A_\star^\ell)^\mathsf{T}\right\| \le C^2\|M\|\sum_{\ell=t}^{\infty}\rho^{2\ell} = C^2\|M\|\frac{\rho^{2t}}{1-\rho^2} \,.$$

Therefore,

$$\begin{aligned}
\left\|\frac{1}{T}\sum_{t=0}^{T-1} P_t - P_\infty\right\| &= \left\|\frac{1}{T}\sum_{t=1}^{T-1}(P_t - P_\infty) + \frac{1}{T}P_\infty\right\| \\
&\le \frac{1}{T}\sum_{t=1}^{T-1}\|P_\infty - P_t\| + \frac{1}{T}\|P_\infty\| \\
&\le \frac{C^2\|M\|}{T(1-\rho^2)}\sum_{t=1}^{T-1}\rho^{2t} + \frac{1}{T}\|P_\infty\| \\
&\le \frac{C^2\|M\|}{T(1-\rho^2)^2} + \frac{1}{T}\|P_\infty\| = O(1/T) \,.
\end{aligned}$$

Therefore, $[\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[x_t x_t^\mathsf{T}]]^{-1} = P_\infty^{-1} + O(1/T)$ from which the claim follows. ■

D.3. Proof of Lemma A.3

Proof Let $c_t = x_t^\mathsf{T}(Q + K^\mathsf{T} RK)x_t$. From Bellman's equation, we have $c_t - \lambda_\star = (\phi(x_t) - \psi(x_t))^\mathsf{T} w_\star$. We write:

$$\begin{aligned}
\widehat{w}_{\mathrm{lstd}}(T) - w_\star &= \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t) - \phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\sum_{t=0}^{T-1}(c_t - \lambda_\star)\phi(x_t)\right) - w_\star \\
&= \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t) - \phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\sum_{t=0}^{T-1}\phi(x_t)(\phi(x_t) - \psi(x_t))^\mathsf{T}\right) w_\star - w_\star \\
&= \left(\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t) - \phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\sum_{t=0}^{T-1}\phi(x_t)(\phi(x_{t+1}) - \psi(x_t))^\mathsf{T} w_\star\right) \\
&= \left(\frac{1}{T}\sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t) - \phi(x_{t+1}))^\mathsf{T}\right)^{-1} \left(\frac{1}{T}\sum_{t=0}^{T-1}\phi(x_t)(\phi(x_{t+1}) - \psi(x_t))^\mathsf{T} w_\star\right) .
\end{aligned}$$

We now proceed by considering the Markov chain $\{z_t := (x_t, w_t)\}$. Observe that x_{t+1} is z_t-measurable, and furthermore the stationary distribution of this chain is $\nu_\infty \times \mathcal{N}(0, \sigma_w^2 I_n)$. From this we conclude two things. First, we conclude by the ergodic theorem that the term inside the inverse converges a.s. to A_∞ and hence the inverse converges a.s. to A_∞^{-1} by the continuous mapping theorem. Next, Theorem D.1 combined with the Cramér-Wold theorem allows us to conclude that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \phi(x_t)(\phi(x_{t+1}) - \psi(x_t))^\mathsf{T} w_\star \overset{D}{\rightsquigarrow} \mathcal{N}(0, B_\infty) .$$

The final claim now follows by Slutsky's theorem. ■

D.4. Proof of Corollary A.4

Proof In the proof we write $\Sigma = \sigma_w^2 I_n$. First, we note that a quick computation shows that $\psi(x) = \mathrm{svec}(Lxx^\mathsf{T} L^\mathsf{T} + \Sigma)$.

Matrix A_∞. We have

$$\begin{aligned}
\phi(x) - \phi(x') &= \mathrm{svec}(xx^\mathsf{T} - (Lx + w)(Lx + w)^\mathsf{T}) \\
&= \mathrm{svec}(xx^\mathsf{T} - Lxx^\mathsf{T} L^\mathsf{T} - Lxw^\mathsf{T} - wx^\mathsf{T} L^\mathsf{T} - ww^\mathsf{T}) .
\end{aligned}$$

Hence, conditioning on x and iterating expectations, we have

$$A_\infty = \mathbb{E}_{x\sim\nu_\infty}[\phi(x)\mathrm{svec}(xx^\mathsf{T} - Lxx^\mathsf{T} L^\mathsf{T} - \Sigma)^\mathsf{T}] .$$

Now let m, n be two test vectors and $M = \mathrm{smat}(m), N = \mathrm{smat}(n)$. We have that,

$$\begin{aligned} m^\mathsf{T} A_\infty n &= \mathbb{E}_{x \sim \nu_\infty}[x^\mathsf{T} M x \langle x x^\mathsf{T} - L x x^\mathsf{T} L^\mathsf{T} - \Sigma, N \rangle] \\ &= \mathbb{E}_{x \sim \nu_\infty}[x^\mathsf{T} M x (x^\mathsf{T} (N - L^\mathsf{T} N L) x - \langle \Sigma, N \rangle)] \\ &= \mathbb{E}_{x \sim \nu_\infty}[x^\mathsf{T} M x x^\mathsf{T} (N - L^\mathsf{T} N L) x] - \langle \Sigma, N \rangle \mathbb{E}_{x \sim \nu_\infty}[x^\mathsf{T} M x] \\ &= \mathbb{E}_g[g^\mathsf{T} P_\infty^{1/2} M P_\infty^{1/2} g g^\mathsf{T} P_\infty^{1/2} (N - L^\mathsf{T} N L) P_\infty^{1/2} g] - \langle \Sigma, N \rangle \langle M, P_\infty \rangle \\ &= 2 \langle P_\infty^{1/2} M P_\infty^{1/2}, P_\infty^{1/2} (N - L^\mathsf{T} N L) P_\infty^{1/2} \rangle + \langle M, P_\infty \rangle \langle N - L^\mathsf{T} N L, P_\infty \rangle - \langle \Sigma, N \rangle \langle M, P_\infty \rangle \\ &= 2 \langle P_\infty^{1/2} M P_\infty^{1/2}, P_\infty^{1/2} (N - L^\mathsf{T} N L) P_\infty^{1/2} \rangle \,, \end{aligned}$$

where the last identity follows since $L P_\infty L^\mathsf{T} - P_\infty + \Sigma = 0$. We therefore have:

$$\begin{aligned} A_\infty &= (P_\infty \otimes_s P_\infty) - (P_\infty L^\mathsf{T} \otimes_s P_\infty L^\mathsf{T}) \\ &= (P_\infty \otimes_s P_\infty)(I - L^\mathsf{T} \otimes_s L^\mathsf{T}) \,. \end{aligned}$$

Note that this writes A_∞ as the product of two invertible matrices and hence A_∞ is invertible.

Matrix B_∞. We have

$$\begin{aligned} \langle \phi(x') - \psi(x), w_\star \rangle &= \mathrm{svec}(L x w^\mathsf{T} + w x^\mathsf{T} L^\mathsf{T} + w w^\mathsf{T} - \Sigma)^\mathsf{T} w_\star \\ &= 2 x^\mathsf{T} L^\mathsf{T} P_\star w + \langle w w^\mathsf{T} - \Sigma, P_\star \rangle \,. \end{aligned}$$

Hence,

$$\begin{aligned} \langle \phi(x') - \psi(x), w_\star \rangle^2 &= 4 (x^\mathsf{T} L^\mathsf{T} P_\star w)^2 + \langle w w^\mathsf{T} - \Sigma, P_\star \rangle^2 + 4 x^\mathsf{T} L^\mathsf{T} P_\star w \langle w w^\mathsf{T} - \Sigma, P_\star \rangle \\ &=: T_1 + T_2 + T_3 \,. \end{aligned}$$

Now we have that $m^\mathsf{T} B_\infty n$ is

$$m^\mathsf{T} B_\infty n = \mathbb{E}[T_1 x^\mathsf{T} M x x^\mathsf{T} N x] + \mathbb{E}[T_2 x^\mathsf{T} M x x^\mathsf{T} N x] + \mathbb{E}[T_3 x^\mathsf{T} M x x^\mathsf{T} N x] \,. \tag{D.1}$$

First, we have

$$\begin{aligned} \mathbb{E}[T_1 x^\mathsf{T} M x x^\mathsf{T} N x] &= 4 \mathbb{E}[(x^\mathsf{T} L^\mathsf{T} P_\star w)^2 x^\mathsf{T} M x x^\mathsf{T} N x] \\ &= 4 \mathbb{E}[x^\mathsf{T} L^\mathsf{T} P_\star w w^\mathsf{T} P_\star L x x^\mathsf{T} M x x^\mathsf{T} N x] \\ &= 4 \mathbb{E}[x^\mathsf{T} L^\mathsf{T} P_\star \Sigma P_\star L x x^\mathsf{T} M x x^\mathsf{T} N x] \\ &= 4 \mathbb{E}_g[g^\mathsf{T} (P_\infty^{1/2} L^\mathsf{T} P_\star \Sigma P_\star L P_\infty^{1/2}) g g^\mathsf{T} (P_\infty^{1/2} M P_\infty^{1/2}) g g^\mathsf{T} (P_\infty^{1/2} N P_\infty^{1/2}) g] \end{aligned}$$

Now we state a result from Magnus to compute the expectation of the product of three quadratic forms of Gaussians.

Lemma D.2 (See e.g. Magnus (1979)) *Let $g \sim \mathcal{N}(0, I)$ and A_1, A_2, A_3 be symmetric matrices. Then,*

$$\begin{aligned} \mathbb{E}[g^\mathsf{T} A_1 g g^\mathsf{T} A_2 g g^\mathsf{T} A_3 g] &= \mathrm{tr}(A_1) \, \mathrm{tr}(A_2) \, \mathrm{tr}(A_3) \\ &\quad + 2 (\mathrm{tr}(A_1) \, \mathrm{tr}(A_2 A_3) + \mathrm{tr}(A_2) \, \mathrm{tr}(A_1 A_3) + \mathrm{tr}(A_3) \, \mathrm{tr}(A_1 A_2)) \\ &\quad + 8 \, \mathrm{tr}(A_1 A_2 A_3) \,. \end{aligned}$$

Now by setting

$$A_1 = P_\infty^{1/2} L^\top P_\star \Sigma P_\star L P_\infty^{1/2} \,,$$
$$A_2 = P_\infty^{1/2} M P_\infty^{1/2} \,,$$
$$A_3 = P_\infty^{1/2} N P_\infty^{1/2} \,,$$

we can compute the expectation $\mathbb{E}[T_1 x^\top M x x^\top N x]$ using Lemma D.2. In particular,

$$\begin{aligned}
\mathrm{tr}(A_1)\,\mathrm{tr}(A_2)\,\mathrm{tr}(A_3) &= \langle P_\infty, L^\top P_\star \Sigma P_\star L \rangle m^\top \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\top n \,, \\
\mathrm{tr}(A_1)\,\mathrm{tr}(A_2 A_3) &= \langle P_\infty, L^\top P_\star \Sigma P_\star L \rangle m^\top (P_\infty \otimes_s P_\infty) n \,, \\
\mathrm{tr}(A_2)\,\mathrm{tr}(A_1 A_3) &= m^\top \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty L^\top P_\star \Sigma P_\star L P_\infty)^\top n \,, \\
\mathrm{tr}(A_3)\,\mathrm{tr}(A_1 A_2) &= m^\top \mathrm{svec}(P_\infty L^\top P_\star \Sigma P_\star L P_\infty)\mathrm{svec}(P_\infty)^\top n \,, \\
\mathrm{tr}(A_1 A_2 A_3) &= m^\top (P_\infty L^\top P_\star \Sigma P_\star L P_\infty \otimes_s P_\infty) n \,.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}[g^\top A_1 g g^\top A_2 g g^\top A_3 g] \\
&\quad = m^\top (\langle P_\infty, L^\top P_\star \Sigma P_\star L \rangle (2(P_\infty \otimes_s P_\infty) + \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\top) \\
&\qquad + 2\mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty L^\top P_\star \Sigma P_\star L P_\infty)^\top + 2\mathrm{svec}(P_\infty L^\top P_\star \Sigma P_\star L P_\infty)\mathrm{svec}(P_\infty)^\top \\
&\qquad + 8(P_\infty L^\top P_\star \Sigma P_\star L P_\infty \otimes_s P_\infty)) n
\end{aligned}$$

Next, we compute

$$\begin{aligned}
\mathbb{E}[T_2 x^\top M x x^\top N x] &= \mathbb{E}[\langle w w^\top - \Sigma, P_\star \rangle^2 x^\top M x x^\top N x] \\
&= \mathbb{E}[\langle w w^\top - \Sigma, P_\star \rangle^2] \mathbb{E}[x^\top M x x^\top N x] \,.
\end{aligned}$$

First, we have

$$\begin{aligned}
\mathbb{E}[\langle w w^\top - \Sigma, P_\star \rangle^2] &= \mathbb{E}[(w^\top P_\star w)^2] - 2\langle \Sigma, P_\star \rangle \mathbb{E}[w^\top P_\star w] + \langle \Sigma, P_\star \rangle^2 \\
&= 2\|\Sigma^{1/2} P_\star \Sigma^{1/2}\|_F^2 + \langle P_\star, \Sigma \rangle^2 - 2\langle \Sigma, P_\star \rangle^2 + \langle P_\star, \Sigma \rangle^2 \\
&= 2\|\Sigma^{1/2} P_\star \Sigma^{1/2}\|_F^2 \,.
\end{aligned}$$

On the other hand,

$$\mathbb{E}[x^\top M x x^\top N x] = 2\langle P_\infty^{1/2} M P_\infty^{1/2}, P_\infty^{1/2} N P_\infty^{1/2} \rangle + \langle M, P_\infty \rangle \langle N, P_\infty \rangle \,.$$

Combining these calculations,

$$\begin{aligned}
\mathbb{E}[T_2 x^\top M x x^\top N x] &= 2\|\Sigma^{1/2} P_\star \Sigma^{1/2}\|_F^2 (2\langle P_\infty^{1/2} M P_\infty^{1/2}, P_\infty^{1/2} N P_\infty^{1/2} \rangle + \langle M, P_\infty \rangle \langle N, P_\infty \rangle) \\
&= 2\|\Sigma^{1/2} P_\star \Sigma^{1/2}\|_F^2 m^\top (2(P_\infty \otimes_s P_\infty) + \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\top) n
\end{aligned}$$

Finally, we have $\mathbb{E}[T_3 x^\top M x x^\top N x] = 0$, which is easy to see because it involves odd powers of w. This gives us that B_∞ is:

$$\begin{aligned}
B_\infty = (&\langle P_\infty, L^\top P_\star \Sigma P_\star L \rangle + 2\|\Sigma^{1/2} P_\star \Sigma^{1/2}\|_F^2)(2(P_\infty \otimes_s P_\infty) + \mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty)^\top) \\
&+ 2\mathrm{svec}(P_\infty)\mathrm{svec}(P_\infty L^\top P_\star \Sigma P_\star L P_\infty)^\top + 2\mathrm{svec}(P_\infty L^\top P_\star \Sigma P_\star L P_\infty)\mathrm{svec}(P_\infty)^\top \\
&+ 8(P_\infty L^\top P_\star \Sigma P_\star L P_\infty \otimes_s P_\infty) \,.
\end{aligned}$$

This completes the proof of the formulas for A_∞ and B_∞.

To obtain the lower bound, we need the following lemma which gives a useful lower bound to Lemma D.2.

Lemma D.3 *Let A_1 be positive semi-definite and let A_2 be symmetric. Let $g \sim \mathcal{N}(0, I)$. We have that:*

$$\mathbb{E}[g^\mathsf{T} A_1 g (g^\mathsf{T} A_2 g)^2] \geq 2\operatorname{tr}(A_1)\operatorname{tr}(A_2^2) + 4\operatorname{tr}(A_1 A_2^2) .$$

Proof Suppose that $A_1 \neq 0$, otherwise the bound holds vacuously. From Lemma D.2,

$$\mathbb{E}[g^\mathsf{T} A_1 g (g^\mathsf{T} A_2 g)^2] = \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 2\operatorname{tr}(A_1)\operatorname{tr}(A_2^2) + 4\operatorname{tr}(A_2)\operatorname{tr}(A_1A_2) + 8\operatorname{tr}(A_1A_2^2) .$$

Since A_1 is PSD and non-zero, this means that $\operatorname{tr}(A_1) > 0$. We proceed as follows:

$$\begin{aligned}
4|\operatorname{tr}(A_2)\operatorname{tr}(A_1A_2)| &= 2|\operatorname{tr}(A_2)\operatorname{tr}(A_1)^{1/2}| \left| 2\frac{\operatorname{tr}(A_1A_2)}{\operatorname{tr}(A_1)^{1/2}} \right| \\
&\overset{(a)}{\leq} \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 4\frac{\operatorname{tr}(A_1A_2)^2}{\operatorname{tr}(A_1)} \\
&= \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 4\frac{\operatorname{tr}(A_1^{1/2}A_1^{1/2}A_2)^2}{\operatorname{tr}(A_1)} \\
&\overset{(b)}{\leq} \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 4\frac{\|A_1^{1/2}\|_F^2\|A_1^{1/2}A_2\|_F^2}{\operatorname{tr}(A_1)} \\
&= \operatorname{tr}(A_1)\operatorname{tr}(A_2)^2 + 4\operatorname{tr}(A_1A_2^2) ,
\end{aligned}$$

where in (a) we used Young's inequality and in (b) we used Cauchy-Schwarz. The claim now follows. ■

We now start from the decomposition (D.1) for B_∞, with $m = n$ and noting that $\mathbb{E}[T_2(x^\mathsf{T} M x)^2] \geq 0$ and $\mathbb{E}[T_3(x^\mathsf{T} M x)^3] = 0$:

$$\begin{aligned}
&m^\mathsf{T} B_\infty m \geq \mathbb{E}[T_1(x^\mathsf{T} M x)^2] \\
&\quad \overset{(a)}{\geq} 8\langle P_\infty, L^\mathsf{T} P_\star \Sigma P_\star L\rangle m^\mathsf{T}(P_\infty \otimes_s P_\infty)m + 16 m^\mathsf{T}(P_\infty L^\mathsf{T} P_\star \Sigma P_\star L P_\infty \otimes_s P_\infty)m .
\end{aligned}$$

Above in (a) we applied the lower bound from Lemma D.3. Hence since m is arbitrary,

$$B_\infty \succeq 8\langle P_\infty, L^\mathsf{T} P_\star \Sigma P_\star L\rangle (P_\infty \otimes_s P_\infty) + 16(P_\infty L^\mathsf{T} P_\star \Sigma P_\star L P_\infty \otimes_s P_\infty) .$$

We also have that $A_\infty = (P_\infty \otimes_s P_\infty)(I - L^\mathsf{T} \otimes L^\mathsf{T})$, and hence $A_\infty^{-1} = (I - L^\mathsf{T} \otimes L^\mathsf{T})^{-1}(P_\infty^{-1} \otimes_s P_\infty^{-1})$. Therefore,

$$\begin{aligned}
A_\infty^{-1} B_\infty A_\infty^{-\mathsf{T}} \succeq{}& 8\langle P_\infty, L^\mathsf{T} P_\star \Sigma P_\star L\rangle (I - L^\mathsf{T} \otimes_s L^\mathsf{T})^{-1}(P_\infty^{-1} \otimes_s P_\infty^{-1})(I - L^\mathsf{T} \otimes_s L^\mathsf{T})^{-\mathsf{T}} \\
&+ 16(I - L^\mathsf{T} \otimes_s L^\mathsf{T})^{-1}(L^\mathsf{T} P_\star \Sigma P_\star L \otimes_s P_\infty^{-1})(I - L^\mathsf{T} \otimes_s L^\mathsf{T})^{-\mathsf{T}} .
\end{aligned}$$

■

D.5. Proof of Lemma A.6

Proof Recall in the notation of the proof of Lemma A.1,

$$\widehat{L}(T) - L_\star = -\lambda L_\star (X^\mathsf{T} X + \lambda I_n)^{-1} + W^\mathsf{T} X (X^\mathsf{T} X + \lambda I_n)^{-1} .$$

Now let us suppose that we are on an event where $X^\mathsf{T} X$ is invertible. Let $X = U\Sigma V^\mathsf{T}$ denote the compact SVD of X. We have:

$$\begin{aligned}\|\widehat{L}(T) - L_\star\| &\leq \lambda \frac{\|L_\star\|}{\lambda_{\min}(X^\mathsf{T} X + \lambda I_n)} + \|W^\mathsf{T} X (X^\mathsf{T} X + \lambda I_n)^{-1}\| \\ &\overset{(a)}{\leq} \lambda \frac{\|L_\star\|}{\lambda_{\min}(X^\mathsf{T} X + \lambda I_n)} + \|W^\mathsf{T} X (X^\mathsf{T} X)^{-1}\| .\end{aligned}$$

The inequality (a) holds due to the following. First observe that $(X^\mathsf{T} X + \lambda I_n)^{-2} \preceq (X^\mathsf{T} X)^{-2}$. Therefore with $M = W^\mathsf{T} X$, conjugating both sides by M, we have $M(X^\mathsf{T} X + \lambda I_n)^{-2} M^\mathsf{T} \preceq M(X^\mathsf{T} X)^{-2} M^\mathsf{T}$. Hence,

$$\begin{aligned}\|M(X^\mathsf{T} X + \lambda I_n)^{-1}\| &= \sqrt{\lambda_{\max}(M(X^\mathsf{T} X + \lambda I_n)^{-2} M^\mathsf{T})} \\ &\leq \sqrt{\lambda_{\max}(M(X^\mathsf{T} X)^{-2} M^\mathsf{T})} \\ &= \|M(X^\mathsf{T} X)^{-1}\| .\end{aligned}$$

By Theorem 2.4 of Simchowitz et al. (2018) for $T \geq C_{L_\star,n} \log(1/\delta)$, there exists an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ such that on $\mathcal{E}$ we have:

$$\|\widehat{L}_{\mathrm{ols}}(T) - L_\star\| \leq C'_{L_\star,n} \sqrt{\log(1/\delta)/T} , \quad X^\mathsf{T} X \succeq C''_{L_\star,n} T \cdot I_n .$$

Hence on this event we have $\|\widehat{L}(T) - L_\star\| \leq C'_{L_\star,n,\lambda} \sqrt{\log(1/\delta)/T}$.

For the remainder of the proof, $O(\cdot)$ will hide constants that depend on $L_\star, n, p, \lambda$ but not on T or δ. We bound the p-th moment as follows. We decompose:

$$\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] = \mathbb{E}[\|\widehat{L}(T) - L_\star\|^p \mathbf{1}_{\mathcal{E}}] + \mathbb{E}[\|\widehat{L}(T) - L_\star\|^p \mathbf{1}_{\mathcal{E}^c}] .$$

On $\mathcal{E}$ we have by the inequality $(a+b)^p \leq 2^{p-1}(a^p + b^p)$ for non-negative a, b,

$$\|\widehat{L}(T) - L_\star\|^p \leq 2^{p-1}(O(\lambda^p/T^p) + O((\log(1/\delta)/T)^{p/2})) .$$

On the other hand, we always have:

$$\|\widehat{L}(T) - L_\star\|^p \leq 2^{p-1}(\|L_\star\|^p + (\|W^\mathsf{T} X\|/\lambda)^p) .$$

Hence:

$$\begin{aligned}\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p \mathbf{1}_{\mathcal{E}^c}] &\leq 2^{p-1}\|L_\star\|^p \mathbb{P}(\mathcal{E}^c) + \frac{2^{p-1}}{\lambda^p} \mathbb{E}[\|W^\mathsf{T} X\|^p \mathbf{1}_{\mathcal{E}^c}] \\ &\leq 2^{p-1}\|L_\star\|^p \delta + \frac{2^{p-1}}{\lambda^p} \sqrt{\mathbb{E}[\|W^\mathsf{T} X\|^{2p}]\delta} .\end{aligned}$$

We will now compute a very crude bound on $\mathbb{E}[\|W^\mathsf{T} X\|^{2p}]$ which will suffice. For non-negative a_t, we have $(a_1 + ... + a_T)^{2p} \le T^{2p-1}(\sum_{t=1}^T a_i^{2p})$ by Hölder's inequality. Hence

$$\begin{aligned}
\mathbb{E}[\|W^\mathsf{T} X\|^{2p}] &= \mathbb{E}\left[\left\| \sum_{t=0}^{T-1} w_i x_i^\mathsf{T} \right\|^{2p} \right] \\
&\le T^{2p-1} \mathbb{E}\left[\sum_{t=1}^{T} \|w_t\|^{2p} \|x_t\|^{2p} \right] \\
&= T^{2p-1} \mathbb{E}[\|w_1\|^{2p}] \sum_{t=1}^{T} \mathbb{E}[\|x_t\|^{2p}] \\
&\le T^{2p} \mathbb{E}[\|w_1\|^{2p}] \|P_\infty\|^p \mathbb{E}_{g \sim \mathcal{N}(0,I)}[\|g\|^{2p}] \\
&= O(T^{2p}) \,.
\end{aligned}$$

Above, P_∞ denotes the covariance of the stationary distribution of $\{x_t\}$. Continuing from above:

$$\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p \mathbf{1}_{\mathcal{E}^c}] = 2^{p-1} \|L_\star\|^p \delta + \frac{2^{p-1}}{\lambda^p} \sqrt{O(T^{2p})\delta} \,.$$

We now set $\delta = O(1/T^{3p})$ so that the term above is $O(1/T^{p/2})$. Doing this we obtain that for T sufficiently large (as a function of only $L_\star, p, \lambda$),

$$\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] \le O(1/T^{p/2}) \,.$$

■

Appendix E. Proof of Lemma A.8

We now state a high probability bound for SGD. This is a straightforward modification of Lemma 6 from Rakhlin et al. (2012) (modifications are needed to deal with the lack of almost surely bounded gradients), and hence we omit the proof.

Lemma E.1 (Lemma 6, Rakhlin et al. (2012)) *Let the assumptions of Lemma A.8 hold. Define two constants:*

$$M := \sup_{\theta \in \Theta} \|\theta\|_2 \,, \;\; G_3 := \sup_{\theta \in \Theta} \|\nabla F(\theta)\|_2 \,.$$

Note that since Θ is compact, both M and G_3 are finite. Fix a $T \ge 4$ and $\delta \in (0, 1/e)$. We have that with probability at least $1 - \delta$, for all $t \le T$,

$$\|\theta_t - \theta_\star\|_2^2 \lesssim \frac{\mathrm{polylog}(T/\delta)}{t} \left(\frac{G_1^2 + G_2^2}{m^2} + \frac{M(G_2 + G_3)}{m} \right) \,.$$

We are now in a position to analyze the asymptotic variance of SGD with projection. As mentioned previously, our argument follows closely that of Toulis and Airoldi (2017). For the remainder of the proof, $O(\cdot)$ and $\Omega(\cdot)$ will hide all constants except those depending on t and δ. Introduce the notation:

$$\begin{aligned}\tilde{\theta}_{t+1} &= \theta_t - \alpha_t g(\theta_t; \xi_t) \,, \\ \theta_{t+1} &= \mathsf{Proj}_\Theta(\tilde{\theta}_{t+1}) \,.\end{aligned}$$

Let $\mathcal{E}_t := \{\tilde{\theta}_t = \theta_t\}$ be the event that the projection step is inactive at time t. Recall that we assumed that $\theta_\star$ is in the interior of Θ. This means there exists a radius $R > 0$ such that $\{\theta : \|\theta - \theta_\star\|_2 \leq R\} \subseteq \Theta$. Therefore, the event $\{\|\tilde{\theta}_t - \theta_\star\|_2 \leq R\} \subseteq \mathcal{E}_t$. We now decompose,

$$\begin{aligned}\mathrm{Var}(\theta_{t+1}) &= \mathrm{Var}(\theta_{t+1} - \tilde{\theta}_{t+1} + \tilde{\theta}_{t+1}) \\ &= \mathrm{Var}(\tilde{\theta}_{t+1}) + \mathrm{Var}(\theta_{t+1} - \tilde{\theta}_{t+1}) + \mathrm{Cov}(\theta_{t+1} - \tilde{\theta}_{t+1}, \tilde{\theta}_{t+1}) + \mathrm{Cov}(\tilde{\theta}_{t+1}, \theta_{t+1} - \tilde{\theta}_{t+1}) \,.\end{aligned}$$

We have that,

$$\theta_{t+1} - \tilde{\theta}_{t+1} = (\theta_{t+1} - \tilde{\theta}_{t+1}) \mathbf{1}_{\mathcal{E}^c_{t+1}} \,.$$

Hence,

$$\begin{aligned}\|\mathrm{Var}(\theta_{t+1} - \tilde{\theta}_{t+1})\| &\leq \mathbb{E}[\|\tilde{\theta}_{t+1} \mathbf{1}_{\mathcal{E}^c_{t+1}} - \theta_{t+1} \mathbf{1}_{\mathcal{E}^c_{t+1}}\|_2^2] \\ &\leq 2(\mathbb{E}[\|\tilde{\theta}_{t+1}\|_2^2 \mathbf{1}_{\mathcal{E}^c_{t+1}}] + \mathbb{E}[\|\theta_{t+1}\|_2^2 \mathbf{1}_{\mathcal{E}^c_{t+1}}]) \\ &\leq 2(\sqrt{\mathbb{E}[\|\tilde{\theta}_{t+1}\|_2^4] \mathbb{E}[\mathbf{1}_{\mathcal{E}^c_{t+1}}]} + M^2 \mathbb{E}[\mathbf{1}_{\mathcal{E}^c_{t+1}}]) \,.\end{aligned}$$

We can bound $\mathbb{E}[\|\tilde{\theta}_{t+1}\|_2^4]$ by a constant for all t using our assumption (A.2). On the other hand,

$$\mathbb{E}[\mathbf{1}_{\mathcal{E}^c_{t+1}}] \leq \mathbb{P}(\|\tilde{\theta}_{t+1} - \theta_\star\|_2 > R) \,.$$

By triangle inequality,

$$\|\tilde{\theta}_{t+1} - \theta_\star\|_2 \leq \|\theta_t - \theta_\star\|_2 + \alpha_t \|g_t\|_2 \,.$$

By Lemma E.1 and the concentration bound on $\|g_t\|_2$ from our assumption (A.3), with probability at least $1 - \delta$,

$$\|\tilde{\theta}_{t+1} - \theta_\star\|_2 \leq O(\mathrm{polylog}(t/\delta)/\sqrt{t}) \,.$$

Hence for t large enough, $\mathbb{E}[\mathbf{1}_{\mathcal{E}^c_{t+1}}] \leq O(\exp(-t^\alpha))$ for some $\alpha > 0$. This shows that $\|\mathrm{Var}(\theta_{t+1} - \tilde{\theta}_{t+1})\| \leq O(\exp(-t^\alpha))$. Similar arguments show that

$$\max\{\|\mathrm{Cov}(\theta_{t+1} - \tilde{\theta}_{t+1}, \tilde{\theta}_{t+1})\|, \|\mathrm{Cov}(\tilde{\theta}_{t+1}, \theta_{t+1} - \tilde{\theta}_{t+1})\|\} \leq O(\exp(-t^\alpha)) \,.$$

Hence:

$$\mathrm{Var}(\theta_{t+1}) = \mathrm{Var}(\tilde{\theta}_{t+1}) + O(\exp(-t^\alpha)) \,.$$

Therefore,

$$\begin{aligned}\mathrm{Var}(\theta_{t+1}) &= \mathrm{Var}(\tilde{\theta}_{t+1}) + O(\exp(-t^\alpha)) \\ &= \mathrm{Var}(\theta_t - \alpha_t g(\theta_t;\xi_t)) + O(\exp(-t^\alpha)) \\ &= \mathrm{Var}(\theta_t) + \alpha_t^2 \mathrm{Var}(g(\theta_t;\xi_t)) - \alpha_t \mathrm{Cov}(\theta_t, g(\theta_t;\xi_t)) - \alpha_t \mathrm{Cov}(g(\theta_t;\xi_t),\theta_t) \\ &\quad + O(\exp(-t^\alpha)) \\ &= \mathrm{Var}(\theta_t) + \alpha_t^2 \mathrm{Var}(g(\theta_t;\xi_t)) - \alpha_t \mathrm{Cov}(\theta_t, \nabla F(\theta_t)) - \alpha_t \mathrm{Cov}(\nabla F(\theta_t),\theta_t) \\ &\quad + O(\exp(-t^\alpha))\,. \end{aligned} \tag{E.1}$$

Now we write:

$$\begin{aligned}\mathrm{Var}(g(\theta_t;\xi_t)) &= \mathrm{Var}(g(\theta_\star;\xi_t) + (g(\theta_t;\xi_t) - g(\theta_\star;\xi_t))) \\ &= \mathrm{Var}(g(\theta_\star;\xi_t)) + \mathrm{Var}(g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)) \\ &\quad + \mathrm{Cov}(g(\theta_\star;\xi_t), g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)) + \mathrm{Cov}(g(\theta_t;\xi_t) - g(\theta_\star;\xi_t), g(\theta_\star;\xi_t))\,. \end{aligned}$$

We have by our assumption (A.4),

$$\begin{aligned}\|\mathrm{Var}(g(\theta_t;\xi_t) - g(\theta_\star;\xi_t))\| &\le \mathbb{E}[\|g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)\|_2^2] \\ &= \mathbb{E}_{\theta_t}\mathbb{E}_\xi[\|g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)\|_2^2] \\ &\le L\mathbb{E}[\|\theta_t - \theta_\star\|_2^2]\,. \end{aligned}$$

On the other hand,

$$\begin{aligned}\|\mathrm{Cov}(g(\theta_\star;\xi_t), g(\theta_t;\xi_t) - g(\theta_\star;\xi_t))\| &\le 2\mathbb{E}[\|g(\theta_\star;\xi_t)\|_2 \|g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)\|_2] \\ &\le 2\sqrt{\mathbb{E}[\|g(\theta_\star;\xi_t)\|_2^2]\mathbb{E}[\|g(\theta_t;\xi_t) - g(\theta_\star;\xi_t)\|_2^2]} \\ &\le 2\sqrt{LG_1^2\mathbb{E}[\|\theta_t - \theta_\star\|_2^2]}\,. \end{aligned}$$

The same bound also holds for $\|\mathrm{Cov}(g(\theta_t;\xi_t) - g(\theta_\star;\xi_t), g(\theta_\star;\xi_t))\|$. Since we know that $\mathbb{E}[\|\theta_t - \theta_\star\|_2^2] \le O(1/t)$, this shows that:

$$\mathrm{Var}(g(\theta_t;\xi_t)) = \mathrm{Var}(g(\theta_\star;\xi)) + o_t(1)\,.$$

Next, by a Taylor expansion of $\nabla F(\theta_t)$ around $\theta_\star$, we have that:

$$\nabla F(\theta_t) = \nabla^2 F(\theta_\star)(\theta_t - \theta_\star) + \mathsf{Rem}(\theta_t - \theta_\star)\,,$$

where $\|\mathsf{Rem}(\theta_t - \theta_\star)\| \le O(\|\theta_t - \theta_\star\|_2^2)$. Therefore, utilizing the fact that adding a non-random vector does not change the covariance,

$$\begin{aligned}\mathrm{Cov}(\theta_t, \nabla F(\theta_t)) &= \mathrm{Cov}(\theta_t, \nabla^2 F(\theta_\star)(\theta_t - \theta_\star) + \mathsf{Rem}(\theta_t - \theta_\star)) \\ &= \mathrm{Cov}(\theta_t, \nabla^2 F(\theta_\star)(\theta_t - \theta_\star)) + \mathrm{Cov}(\theta_t, \mathsf{Rem}(\theta_t - \theta_\star)) \\ &= \mathrm{Cov}(\theta_t, \nabla^2 F(\theta_\star)\theta_t) + \mathrm{Cov}(\theta_t - \theta_\star, \mathsf{Rem}(\theta_t - \theta_\star)) \\ &= \mathrm{Var}(\theta_t)\nabla^2 F(\theta_\star) + \mathrm{Cov}(\theta_t - \theta_\star, \mathsf{Rem}(\theta_t - \theta_\star))\,. \end{aligned}$$

We now bound $\mathrm{Cov}(\theta_t - \theta_\star, \mathsf{Rem}(\theta_t - \theta_\star))$ as:

$$\|\mathrm{Cov}(\theta_t - \theta_\star, \mathsf{Rem}(\theta_t - \theta_\star))\| \le O(\mathbb{E}[\|\theta_t - \theta_\star\|_2^3]) \le O(\mathrm{polylog}(t)/t^{3/2}) \ .$$

Above, the last inequality comes from the high probability bound given in Lemma E.1. Observing that $\mathrm{Cov}(\theta_t, \nabla F(\theta_t))^\mathsf{T} = \mathrm{Cov}(\nabla F(\theta_t), \theta_t)$, combining our calculations and continuing from Equation (E.1),

$$\begin{aligned}\mathrm{Var}(\theta_{t+1}) &= \mathrm{Var}(\theta_t) + \alpha_t^2(\mathrm{Var}(g(\theta_\star;\xi)) + o_t(1)) - \alpha_t(\mathrm{Var}(\theta_t)\nabla^2 F(\theta_\star) + \nabla^2 F(\theta_\star)\mathrm{Var}(\theta_t)) \\ &\quad + \alpha_t O(\mathrm{polylog}(t)/t^{3/2}) + O(\exp(-t^\alpha)) \ .\end{aligned}$$

We now make two observations. Recall that $\alpha_t = 1/(mt)$. Hence we have $O(\exp(-t^\alpha)) = \alpha_t^2 O(t^2 \exp(-t^\alpha)) = \alpha_t^2 o_t(1)$. Similarly, $\alpha_t O(\mathrm{polylog}(t)/t^{3/2}) = \alpha_t^2 O(\mathrm{polylog}(t)/t^{1/2}) = \alpha_t^2 o_t(1)$. Therefore,

$$\mathrm{Var}(\theta_{t+1}) = \mathrm{Var}(\theta_t) - \alpha_t(\mathrm{Var}(\theta_t)\nabla^2 F(\theta_\star) + \nabla^2 F(\theta_\star)\mathrm{Var}(\theta_t)) + \alpha_t^2(\mathrm{Var}(g(\theta_\star;\xi)) + o_t(1)) \ .$$

This matrix recursion can be solved by Corollary C.1 of Toulis and Airoldi (2017), yielding (A.5).

To complete the proof, by a Taylor expansion we have:

$$T \cdot \mathbb{E}[F(\theta_T) - F(\theta_\star)] = \frac{T}{2}\,\mathrm{tr}(\nabla^2 F(\theta_\star)\mathbb{E}[(\theta_T - \theta_\star)(\theta_T - \theta_\star)^\mathsf{T}]) + \frac{T}{6}\mathbb{E}[\nabla^3 f(\hat{\theta})(\theta_T - \theta_\star)^{\otimes 3}] \ .$$

As above, we can bound $|\mathbb{E}[\nabla^3 f(\hat{\theta})(\Theta_T - \theta_\star)^{\otimes 3}]| \le O(\mathbb{E}[\|\theta_T - \theta_\star\|_2^3]) \le O(\mathrm{polylog}(T)/T^{3/2})$, and hence $T \cdot |\mathbb{E}[\nabla^3 f(\hat{\theta})(\Theta_T - \theta_\star)^{\otimes 3}]| \to 0$. On the other hand, letting $\mu_T := \mathbb{E}[\theta_T]$, by a bias-variance decomposition,

$$\begin{aligned}\mathbb{E}[(\theta_T - \theta_\star)(\theta_T - \theta_\star)^\mathsf{T}] &= \mathbb{E}[(\theta_T - \mu_T)(\theta_T - \mu_T)^\mathsf{T}] + (\mu_T - \theta_\star)(\mu_T - \theta_\star)^\mathsf{T} \\ &\succeq \mathbb{E}[(\theta_T - \mu_T)(\theta_T - \mu_T)^\mathsf{T}] = \mathrm{Var}(\theta_T) \ .\end{aligned}$$

Therefore,

$$T \cdot \mathbb{E}[F(\theta_T) - F(\theta_\star)] \ge \frac{1}{2m}\,\mathrm{tr}(\nabla^2 F(\theta_\star)(mT)\mathrm{Var}(\theta_T)) - \frac{T}{6}|\mathbb{E}[\nabla^3 f(\hat{\theta})(\theta_T - \theta_\star)^{\otimes 3}]| \ .$$

Taking limits on both sides yields (A.7). This concludes the proof of Lemma A.8.