MASTER'S THESIS IN MATHEMATICS

DEPARTMENT OF MATHEMATICS, STOCKHOLM UNIVERSITY

Asymptotic Analysis and Comparison of Model-Based and Model-Free Methods for the Linear Quadratic Regulator

by

Zahra Alimoradzadeh

2025 - No M10

Asymptotic Analysis and Comparison of Model-Based and Model-Free Methods for the Linear Quadratic Regulator

Zahra Alimoradzadeh

MASTER'S THESIS IN MATHEMATICS 30 Higher Education Credits (ECTS), Advanced Leve

Supervisor: Prof. Yishao Zhou

2025

Abstract

This thesis studies the asymptotic sample efficiency of model-based and model-free reinforcement learning algorithms in the Linear Quadratic Regulator (LQR) setting. We focus on the problem of policy evaluation under a fixed linear controller $u_t = Kx_t$, where the value function is quadratic and characterized by the unique solution $P^\star$ of a discrete-time Lyapunov equation.

Two estimators of $P^\star$ are analyzed:

1. A model-based plug-in estimator, which estimates the closed-loop dynamics via regularized least squares and substitutes the estimate into the Lyapunov operator, and
2. A model-free estimator based on Least-Squares Temporal Difference (LSTD) learning, which directly estimates the quadratic value function from trajectory data.

We analyze policy evaluation in infinite-horizon LQR under a fixed stabilizing controller, comparing a model-based plug-in estimator of the Lyapunov solution with a model-free LSTD estimator. Using Markov chain Central Limit Theorems (CLTs), the Delta Method, and uniform integrability, we establish that the model-based estimator attains strictly smaller asymptotic risk than LSTD.

Abstract

Denna avhandling studerar den asymptotiska sampleffektiviteten hos modellbaserade och modelfria förstärkningsinlärningsalgoritmer i den linjära kvadratiska regulatorn (LQR). Vi fokuserar på problemet med policyevakuering under en fixerad linjär regulator $u_t = Kx_t$, där värdefunktionen är kvadratisk och karakteriseras av den unika lösningen $P^\star$ till en diskret Lyapunov-ekvation.

Två estimatorer av $P^\star$ analyseras:

1. en modellbaserad plug-in-estimator som skattar det slutna systemets dynamik via regulariserad minsta kvadrat-metod och ersätter skattningen i Lyapunov-operatorn, och
2. en modelfri estimator baserad på Least-Squares Temporal Difference (LSTD), som direkt skattar den kvadratiska värdefunktionen från trajektoriedata.

Vi analyserar policyevakuering i en oändlighorisont-LQR under en fixerad stabiliserande regulator och jämför en modellbaserad plug-in-estimator av Lyapunov-lösningen med en modelfri LSTD-estimator. Med hjälp av centrala gränsvärdessatser för Markovkedjor, delta-metoden och uniform integrabilitet visar vi att den modellbaserade estimatorn uppnår strikt lägre asymptotisk risk än LSTD.

Acknowledgments

I would like to sincerely thank my supervisor, Yishao, for her invaluable guidance, support, and encouragement throughout this project. I am also deeply grateful to my husband and my dear mother for their constant love, patience, and support, without which this work would not have been possible.

Contents

1 Introduction

In recent years, reinforcement learning (RL) has achieved impressive results in fields such as robotics, game playing, and autonomous systems. These applications often involve systems that evolve over time, where making optimal decisions is both challenging and critical. Among the tools used in RL, there are two major families of methods, model-based and model-free. Each has its strengths and weaknesses, but the precise nature of their trade-offs, especially in continuous control settings, remains an open question.

To study these trade-offs, this thesis focuses on the Linear Quadratic Regulator (LQR), a classical control problem involving linear systems with quadratic cost. Despite its simplicity, LQR captures many of the essential difficulties of real-world control problems and serves as a common benchmark in both control theory and reinforcement learning research.

Although both model-based and model-free methods have been extensively studied individually, there has been limited theoretical work directly comparing their sample efficiency in continuous control tasks. The paper by Tu and Recht (2019) takes an important step toward filling this gap by providing an asymptotic analysis of both approaches on the LQR problem, revealing significant differences in how efficiently they learn from the data.

The goal of this thesis is to carefully study and reproduce the theoretical results of Tu and Recht (2019), while also providing accessible explanations and background to make the material understandable to other students in mathematics.

In the following sections, we begin by introducing the foundational mathematics underlying the LQR problem and reinforcement learning. The core theoretical results are then presented and analyzed.

2 Preliminary

2.1 Probability theory

This section provides the mathematical background necessary to understand the theoretical results and algorithms presented in this thesis. The focus is on linear dynamical systems, optimal control theory, particularly the Linear Quadratic Regulator (LQR),and essential concepts from reinforcement learning.[1][2][3][4]

Definition 2.1 (Expectation)**.** Let $X_1, X_2, \ldots, X_n$ be discrete random variables with joint probability mass function

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$

Then the expectation of a function $g(X_1, X_2, \ldots, X_n)$ is defined as

$$\mathbb{E}[g(X_1, X_2, \ldots, X_n)] = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} g(x_1, x_2, \ldots, x_n)\, p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n).$$

Definition 2.2 (Expected Value)**.** The *expected value* of a random variable is the long-run average of its outcomes, weighted by their probabilities.
Formally:

- If X is a discrete random variable with outcomes x_i and probabilities p_i, then

$$\mathbb{E}[X] = \sum_i x_i\, p_i.$$

- If X is a continuous random variable with probability density function $f(x)$, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

Definition 2.3 (Covariance)**.** The *covariance* of two random variables X and Y is defined as

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]\,.$$

Definition 2.4 (Variance)**.** Let X be a random variable with expected value operator denoted by $\mathbb{E}[\cdot]$. The *variance* of X is defined as

$$\operatorname{Var}[X] = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big]\,,$$

provided that the expected value exists.

Definition 2.5 (Markov Chain)**.** A Markov chain is a sequence of random variables $\{X_n\}_{n\geq 0}$ taking values in a countable state space $\mathcal{S}$, such that the probability of transitioning to the next state depends only on the current state. This property is known as the *Markov property*, and is formally stated as

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

for all $n \geq 0$ and all sequences $x_0, x_1, \ldots, x_{n+1} \in \mathcal{S}$.

The evolution of the chain is governed by a *transition probability matrix* $P \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, where each entry P_{ij} represents the probability of transitioning from state i to state j:

$$P_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i)$$

Definition 2.6 (Ergodic Markov chain)**.** A Markov chain $\{X_n\}$ is called *ergodic* if the limit

$$\pi(j) = \lim_{n\to\infty} \mathbb{P}^i\{X_n = j\}$$

exists for every state j and does not depend on the initial state i. The D-vector $\boldsymbol{\pi}$ is called the stationary probability.

In other words, the probability $\boldsymbol{\pi}(j)$ of being on state j after a long time is *independent* of the initial state i. We can also write this as

$$\pi(j) = \lim_{n\to\infty} (\mathbf{P}^n)_{ij}.$$

Consider the following important implication. If a Markov chain is ergodic, then

$$\begin{aligned}\pi(j) &\triangleq \lim_{n\to\infty} (\mathbf{P}^n)_{ij} \overset{\star}{=} \lim_{n\to\infty} (\mathbf{P}^{n+1})_{ij} = \lim_{n\to\infty} (\mathbf{P}^n\mathbf{P})_{ij} \\ &= \lim_{n\to\infty} \sum_{d\in D} (\mathbf{P}^n)_{id}\mathbf{P}_{dj} = \sum_{d\in D} \pi(d)\mathbf{P}_{dj}.\end{aligned}$$

Step $\star$ holds because if the limit exists, the distinction between n and $n+1$ does not matter. So we can write this as

$$\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \mathbf{P},$$

where $\boldsymbol{\pi}$ is a column vector. Hence the name *stationary probability* for $\boldsymbol{\pi}$. It is a

distribution that does not change over time.[4]

Proposition 2.7 (Slutsky's theorem,[5]). *If $A_n \to_p a$, $B_n \to_p b$, and $Z_n \xrightarrow{d} Z$, then*

$$A_n Z_n + B_n \xrightarrow{d} aZ + b.$$

Proof. Following [5] $A_n \to_p a$, $B_n \to_p b$, and $Z_n \xrightarrow{d} Z$, where a, b are constants, implies that $(Z_n, A_n, B_n) \xrightarrow{d} (Z, a, b)$ in $\mathbb{R}^3$. Hence, by the $\mathbb{R}^3$ version of Skorokhod's theorem, there exists a sequence $(Z_n^*, A_n^*, B_n^*) \overset{d}{=} (Z_n, A_n, B_n)$ such that $(Z_n^*, A_n^*, B_n^*) \to (Z^*, a, b) \overset{d}{=} (Z, a, b)$. Hence

$$A_n Z_n + B_n \overset{d}{=} A_n^* Z_n^* + B_n^* \xrightarrow{a.s.} aZ^* + b \overset{d}{=} aZ + b.$$

Since $\xrightarrow{a.s.}$ implies $\xrightarrow{p}$, which in turn implies $\xrightarrow{d}$, the convergence above yields the desired conclusion. □

Lemma 2.8 ([6]). *Let $(X_n)_{n\geq 1}$ be random vectors with $X_n \Rightarrow X$. Let $f : \mathbb{R}^d \to [0, \infty)$ be continuous and assume $\mathbb{E}[f(X)] < \infty$.*

1. $\liminf_{n\to\infty} \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$.
2. *If moreover $\sup_{n\geq 1} \mathbb{E}[f(X_n)^{1+\varepsilon}] < \infty$ for some $\varepsilon > 0$, then $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$.*

For $M > 0$, define the bounded $f_M(x) := \min\{f(x), M\}$. Then f_M is bounded and continuous, hence by the Portmanteau theorem, $\mathbb{E}[f_M(X_n)] \to \mathbb{E}[f_M(X)]$ as $n \to \infty$. Since $f_M \leq f$,

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

Proof. Follwing [6] letting $M \uparrow \infty$ and using monotonic convergence (as $f \geq 0$ and $\mathbb{E}[f(X)] < \infty$), we get $\mathbb{E}[f_M(X)] \uparrow \mathbb{E}[f(X)]$, which yields $\liminf_{n\to\infty} \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$.

For any $M > 0$, consider the "tail" random variable

$$Y_n^{(M)} := \begin{cases} f(X_n), & \text{if } f(X_n) > M, \\ 0, & \text{otherwise.} \end{cases}$$

Then we have the bound

$$Y_n^{(M)} \;\le\; \frac{f(X_n)^{1+\varepsilon}}{M^{\varepsilon}}.$$

Taking expectations and the supremum over n yields

$$\sup_n \mathbb{E}[Y_n^{(M)}] \;\le\; \frac{C}{M^{\varepsilon}} \;\xrightarrow[M\to\infty]{}\; 0.$$

Thus the family $\{f(X_n)\}_n$ is uniformly integrable.

For any $M > 0$,

$$\begin{aligned}
&\Big|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]\Big| \\
&\le \underbrace{\Big|\mathbb{E}[f_M(X_n)] - \mathbb{E}[f_M(X)]\Big|}_{\to 0 \text{ by Portmanteau}} + \underbrace{\mathbb{E}[Y_n^{(M)}]}_{\to 0 \text{ as } M\to\infty \text{ uniformly in } n} + \underbrace{\mathbb{E}\Big[f(X)\,\mathbf{1}_{\{f(X)>M\}}\Big]}_{\to 0 \text{ as } M\to\infty}.
\end{aligned}$$

Given $\eta > 0$, choose M large enough so that the last two terms are below η uniformly in n. Then let $n \to \infty$ to eliminate the first term. Hence

$$\mathbb{E}[f(X_n)] \;\to\; \mathbb{E}[f(X)],$$

as desired. □

Central Limit Theorem (CLT)

Theorem 2.9 (Classical Central Limit Theorem,[7])**.** *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with finite mean $\mathbb{E}[X_i] = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2 > 0$. Then the normalized sum*

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

converges in distribution to the standard normal distribution:

$$Z_n \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty.$$

Proof. We use Characteristic Functions to proof the theorem. Define the normalized random variables:

$$Y_i := \frac{X_i - \mu}{\sigma}, \quad \text{so that } \mathbb{E}[Y_i] = 0, \quad \mathrm{Var}(Y_i) = 1$$

Then the normalized sum becomes:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

Let $\varphi_{Z_n}(t)$ be the characteristic function of Z_n. Since the Y_i are i.i.d., we have:

$$\varphi_{Z_n}(t) = \mathbb{E}\left[e^{itZ_n}\right] = \mathbb{E}\left[e^{it\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{itY_i/\sqrt{n}}\right] = \left(\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

We now use the Taylor expansion of the characteristic function around $t = 0$. Since $\mathbb{E}[Y_1] = 0$, $\mathbb{E}[Y_1^2] = 1$, we get:

$$\varphi_{Y_1}(t) = 1 - \frac{t^2}{2} + o(t^2) \quad \text{as } t \to 0$$

Then,

$$\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)$$

Substituting into $\varphi_{Z_n}(t)$:

$$\varphi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n$$

Now using the limit $(1 + \frac{a}{n})^n \to e^a$, we obtain:

$$\varphi_{Z_n}(t) \to \exp\left(-\frac{t^2}{2}\right)$$

This is the characteristic function of the standard normal distribution $\mathcal{N}(0, 1)$. By Lévy's continuity theorem, this implies:

$$Z_n \xrightarrow{d} \mathcal{N}(0, 1)$$

which completes the proof. □

Theorem 2.10 ([2])**.** *Let Y be a centered strictly stationary strongly mixing sequence. Suppose at least one of the following conditions holds:*

1. *There exists $B < \infty$ such that $|Y_n| < B$ almost surely and $\sum_n \alpha(n) < \infty$; or*

2. $\mathbb{E}|Y_n|^{2+\delta} < \infty$ *for some* $\delta > 0$ *and*

$$\sum_n \alpha(n)^{\delta/(2+\delta)} < \infty. \tag{10}$$

Then

$$\sigma^2 = \mathbb{E}(Y_0^2) + 2\sum_{j=1}^{\infty} \mathbb{E}(Y_0 Y_j) < \infty,$$

and if $\sigma^2 > 0$*, then as* $n \to \infty$*,*

$$n^{-1/2} S_n \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Remark 2.11. The assumption of stationarity is not an issue for Harris ergodic Markov chains, since if a Central Limit Theorem (CLT) holds for any one initial distribution, then it holds for every initial distribution.

Corollary 2.12 ([2])**.** *Let* $f : X \to \mathbb{R}$ *be a Borel function such that* $\mathbb{E}_\pi|f(x)|^{2+\delta} < \infty$ *for some* $\delta > 0$*, and suppose* X *is a Harris ergodic Markov chain with stationary distribution* π*. If (3) holds such that* $\mathbb{E}_\pi M < \infty$ *and* $\gamma(n)$ *satisfies*

$$\sum_n \gamma(n)^{\delta/(2+\delta)} < \infty, \tag{11}$$

then for any initial distribution, as $n \to \infty$*,*

$$\sqrt{n}(\bar{f}_n - \mathbb{E}_\pi f) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Proof. Let $\alpha(n)$ and $\alpha_f(n)$ denote the strong mixing coefficients for the Markov chain $X = \{X_n\}$ and the functional process $\{f(X_n)\}$, respectively. By an earlier remark, $\alpha_f(n) \leq \alpha(n)$ for all $n \geq 1$. Moreover, we have that $\alpha(n) \leq \gamma(n)\mathbb{E}_\pi M$, where $\gamma(n)$ and M are given in (3). Hence, (11) guarantees that

$$\sum_n \alpha_f(n)^{\delta/(2+\delta)} < \infty,$$

and the result follows from the Theorem 2.10 and Remark 2.11. □

Theorem 2.13 (Markov Chain Central Limit Theorem [6])**.** *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-measurable function, and suppose that

$$\mathbb{E}_\pi \left[|f|^{2+\delta} \right] < \infty \quad \textit{for some } \delta > 0.$$

Then, for any initial distribution, the following Central Limit Theorem holds:

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

where the asymptotic variance is given by

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

This result generalizes the classical Central Limit Theorem to dependent data generated by a Markov chain. In this setting, the observations $f(x_1), \ldots, f(x_n)$ are not independent and identically distributed but are correlated through the Markovian dynamics. Geometric ergodicity ensures that the chain mixes sufficiently fast so that long-run averages behave similarly to those obtained from i.i.d. samples. Although the Central Limit Theorem still applies, the asymptotic variance must incorporate the autocorrelation across time steps. Consequently, the asymptotic variance σ_f^2 reflects both the marginal variance of $f(x_0)$ and the covariances between $f(x_0)$ and its future values $f(x_i)$.

Proof of Theorem 2.13. [2] The proof follows from Theorem 2.10 and Corollary 2.12, which adapt the classical Central Limit Theorem for strongly mixing sequences to the Markov chain setting.

According to Theorem 2.10, if $\{Y_n\}$ is a strictly stationary sequence with strong-mixing coefficients $\alpha(k)$ satisfying

$$\sum_{k=1}^{\infty} \alpha(k)^{\delta/(2+\delta)} < \infty \quad \text{and} \quad \mathbb{E}[Y_0] = 0,\ \mathbb{E}[|Y_0|^{2+\delta}] < \infty,$$

then

$$\sqrt{n}\, \bar{Y}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \overset{d}{\Rightarrow} \mathcal{N}(0, \sigma_Y^2), \qquad \sigma_Y^2 = \mathrm{Var}(Y_0) + 2 \sum_{k=1}^{\infty} \mathrm{Cov}(Y_0, Y_k).$$

Let $Y_i = f(X_i)$, where $\{X_i\}$ is Harris ergodic with stationary distribution π. Under

Harris ergodicity and geometric convergence of the total-variation distance, there exist constants $M < \infty$ and $\rho \in (0, 1)$ such that

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq M\rho^n \qquad \text{for all } x \in \mathsf{X},\ n \geq 0.$$

By Theorem 2 of [2], this geometric ergodicity implies a bound on the strong-mixing coefficients of the stationary sequence $\{f(X_n)\}$:

$$\alpha(k) \leq C\,\rho^k$$

for some constant $C > 0$. Hence $\sum_{k\geq 1} \alpha(k)^{\delta/(2+\delta)} < \infty$.

Since $\pi(f^2) < \infty$ and $\mathbb{E}_\pi[|f|^{2+\delta}] < \infty$, we have $\mathrm{Var}_\pi(f(X_0)) < \infty$, and the absolute summability of the covariances $\sum_{k\geq 1} |\,\mathrm{Cov}_\pi(f(X_0), f(X_k))| < \infty$ follows from the strong-mixing bound and Hölder's inequality.

All assumptions of Theorem 3.5 are thus satisfied by the stationary sequence $\{Y_n = f(X_n)\}$. Consequently,

$$\sqrt{n}\,\bar{f}_n \overset{d}{\Rightarrow} \mathcal{N}(0, \sigma_f^2), \qquad \sigma_f^2 = \mathrm{Var}_\pi(f(X_0)) + 2\sum_{k=1}^{\infty} \mathrm{Cov}_\pi(f(X_0), f(X_k)).$$

Corollary 2.12 further shows that the same conclusion holds if the chain is polynomially ergodic of order $m > 1$, provided $\mathbb{E}_\pi[|f|^{2+\delta}] < \infty$ for some $\delta > 2/(m-1)$. No new proof is required for these cases; the result follows from the same application of Theorem 3.5 once the appropriate bound on the mixing coefficients is established. This completes the proof. □

Delta Method

Delta Method:[8] Let $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^d$ be an estimator such that:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(0, \Sigma),$$

and let $g : \mathbb{R}^d \to \mathbb{R}^k$ be a differentiable function at $\boldsymbol{\theta}$. Then:

$$\sqrt{n}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right) \xrightarrow{D} \mathcal{N}\left(0, Dg(\boldsymbol{\theta})\,\Sigma\,Dg(\boldsymbol{\theta})^\top\right),$$

where $Dg(\boldsymbol{\theta})$ is the Jacobian matrix (derivative) of g at $\boldsymbol{\theta}$

The Delta Method tells us that if you have an estimator that satisfies a Central Limit Theorem (CLT), then any smooth function of that estimator also satisfies a CLT — and the limiting variance is transformed by the derivative of the function.

Kullback–Leibler divergence

Definition 2.14 ([9])**.** (The *Kullback–Leibler divergence* (or simply, *KL divergence*)) is a measure of the difference between two probability distributions defined over the same random variable x. It originates from probability theory and information theory and is closely related to *relative entropy*, *information divergence*, and *information for discrimination*.

Let $p(x)$ and $q(x)$ be two probability distributions over a discrete random variable x, with $p(x) > 0$ and $q(x) > 0$ for all $x \in \mathcal{X}$ and $\sum_{x\in\mathcal{X}} p(x) = 1$. The KL divergence of $q(x)$ from $p(x)$, denoted $D_{\mathrm{KL}}(p(x)\|q(x))$, is defined as

$$D_{\mathrm{KL}}(p(x)\|q(x)) = \sum_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)}. \tag{1}$$

The continuous version is given by

$$D_{\mathrm{KL}}(p(x)\|q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)}\, dx. \tag{2}$$

The KL divergence measures the expected number of extra bits required to code samples drawn from $p(x)$ when using a code optimized for $q(x)$ instead of the true distribution $p(x)$. Typically, $p(x)$ represents the "true" or empirical distribution, while $q(x)$ represents a model or approximation.[9]

Properties. [9]

1. **Non-negativity:**

$$D_{\mathrm{KL}}(P\|Q) \geq 0, \qquad D_{\mathrm{KL}}(P\|Q) = 0 \iff P = Q.$$

2. **Asymmetry:** In general, $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$.

3. **Not a true metric:** The KL divergence does not satisfy the triangle inequality and hence is not a distance metric.

4. **Handling zero probabilities:** If $p(x) \neq 0$ but $q(x) = 0$, then $D_{\mathrm{KL}}(p\|q) = \infty$. In practical applications, smoothing techniques are used to avoid assigning zero probability to any event.

2.2 Some useful notations and facts from Linear Algebra

Definition 2.15 (Kronecker Product,[10])**.** Let $A \in \mathbb{R}^{q\times n}$ and $B \in \mathbb{R}^{p\times m}$. The *Kronecker product* of A and B, denoted $A\otimes B$, is the block matrix of size $(pq)\times(nm)$ defined as

$$A \otimes B = \begin{bmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{q1}B & \cdots & A_{qn}B \end{bmatrix},$$

where A_{ij} denotes the element in the i-th row and j-th column of A.

Definition 2.16 (Vec-operation, [10])**.** Let $X \in \mathbb{R}^{m\times n}$. The *vec-operator*, denoted $\mathrm{vec}(X)$, is defined as the $mn \times 1$ column vector obtained by stacking the columns of X on top of each other:

$$\mathrm{vec}(X) = \begin{bmatrix} X_{:,1} \\ X_{:,2} \\ \vdots \\ X_{:,n} \end{bmatrix},$$

where $X_{:,k}$ denotes the k-th column of X.

Note that the vec-operator is related to the Kronecker product as follows:

$$\mathrm{vec}(ab^T) = \mathrm{vec}\Big(\begin{bmatrix} ab_1 & ab_2 & \cdots & ab_n \end{bmatrix}\Big) = \begin{bmatrix} ab_1 \\ ab_2 \\ \vdots \\ ab_n \end{bmatrix} = \begin{bmatrix} b_1 a \\ b_2 a \\ \vdots \\ b_n a \end{bmatrix} = b \otimes a.$$

Thus, as a basic rule,

$$\mathrm{vec}(ab^T) = b \otimes a,$$

where a and b can be **any size vectors**.[10]

Given two $p \times q$ matrices A, B such that

$$A_{p\times q} := [a_{ij}] \quad \text{and} \quad B_{p\times q} := [b_{ij}],$$

the following relationship between the trace and the vec operator holds:

$$\begin{aligned}
\operatorname{tr}(A^T B) &= \operatorname{tr}(BA^T) && \text{(trace invariant with respect to a cyclic transformation)} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q} a_{ij}b_{ij} && \text{(first sum for trace, second for matrix product)} \\
&= (\operatorname{vec} A)^T \operatorname{vec} B \\
&= \sum_{i,j} a_{ij}b_{ij}, && \text{(multiplies corresponding elements)}
\end{aligned}$$

which finally allows us to write

$$\operatorname{tr}(A^T B) = (\operatorname{vec} A)^T \operatorname{vec} B.$$

Proposition 2.17. *Given matrices A of size $p \times q$, B of size $q \times r$, and C of size $s \times r$, we have the following important relationship, which connects the vec-operator and the Kronecker:*

$$\operatorname{vec}(ABC^T) = (C \otimes A)\operatorname{vec}(B), \qquad A \in \mathbb{R}^{p\times q},\ \ B \in \mathbb{R}^{q\times r},\ \ C \in \mathbb{R}^{s\times r}.$$

Definition 2.18 (Positive Semidefinite and Positive Definite [10])**.** Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix.

(a) M is called *positive semidefinite* (PSD, denoted $M \succeq 0$) if

$$x^\top M x \geq 0 \quad \forall x \in \mathbb{R}^n.$$

(b) M is called *positive definite* (PD, denoted $M \succ 0$) if

$$x^\top M x > 0 \quad \forall x \in \mathbb{R}^n \setminus \{0\}.$$

Definition 2.19 (Spectral radius)**.** Let $A \in \mathbb{R}^n \times n$. The spectral radius of A dedonted $\rho(A)$, is defined as

$$\rho(A) := \max_{1\leq i\leq n} |\lambda_i(A)|$$

where $\lambda_i(A)$,$i = 1, ..., n$ are eigenvalues of A counted by multiplicity.

Definition 2.20 (Operator norm)**.** Let $A \in \mathbb{R}^n \times n$. The operator norm of A is

$$\|A\| := \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

where $\|x\|^2 = x_1^2 + \cdots + x_n^2$.

In general, we have

$$\rho(A) \leq \|A\|.$$

However, if A is normal, i.e., $A^\top A = AA^\top$, then $\rho(A) = \|A\|$.

In the sequel, we also use the Frobenius norm:

$$\|A\|_F^2 := \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}]^2, \quad A \in \mathbb{R}^{m \times n}.$$

We see that

$$\|A\|_F^2 = \sigma_1 + \cdots + \sigma_r$$

where $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are the nonzero singularvalues of A and r is the rank of A.

Lemma 2.21 ([6])**.** *Let $A \in \mathbb{R}^{n \times n}$ be a stable matrix satisfying:*

$$\|A^k\| \leq C\rho^k \quad \textit{for all } k \geq 0,$$

for some constants $C > 0$ and $\rho \in (0, 1)$. Suppose $\Delta \in \mathbb{R}^{n \times n}$ is a perturbation such that:

$$\|\Delta\| \leq \frac{\gamma - \rho}{C}, \quad \textit{for some } \gamma \in (\rho, 1).$$

Then:

(a) $A + \Delta$ is stable with $\rho(A + \Delta) \leq \gamma$,

(b) $\|(A + \Delta)^k\| \leq C\gamma^k \quad$ for all $k \geq 0$.

Proof. Part(b):We aim to show:

$$\|(A + \Delta)^k\| \leq C\gamma^k.$$

This is proven using the binomial expansion for matrix powers:

$$(A + \Delta)^k = \sum_{i=0}^{k} \sum_{j} T_{i,j},$$

where each term $T_{i,j}$ in the expansion involves i powers of Δ, and there are $\binom{k}{i}$ such terms. Each term satisfies:

$$T_{i,j} = A^{k-i}\Delta^i \quad \Rightarrow \quad \|T_{i,j}\| \leq C\rho^{k-i}\|\Delta\|^i.$$

Hence, by the triangle inequality:

$$\|(A+\Delta)^k\| \leq \sum_{i=0}^{k} \binom{k}{i} C\rho^{k-i}\|\Delta\|^i.$$

Factor out C:

$$= C\sum_{i=0}^{k} \binom{k}{i} (\|\Delta\|)^i \rho^{k-i} = C(\rho + \|\Delta\|)^k.$$

Using the assumption:

$$\|\Delta\| \leq \frac{\gamma - \rho}{C} \quad \Rightarrow \quad C\|\Delta\| + \rho \leq \gamma,$$

we conclude:

$$\|(A+\Delta)^k\| \leq C\gamma^k.$$

Part(a): To show $A+\Delta$ is stable, we must show $\rho(A+\Delta) < 1$. From matrix analysis:

$$\rho(A+\Delta) \leq \|(A+\Delta)^k\|^{1/k}.$$

Using the bound from (b):

$$\|(A+\Delta)^k\|^{1/k} \leq (C\gamma^k)^{1/k} = C^{1/k}\gamma.$$

As $k \to \infty$, $C^{1/k} \to 1$, so:

$$\limsup_{k\to\infty} \|(A+\Delta)^k\|^{1/k} \leq \gamma.$$

Thus:

$$\rho(A+\Delta) \leq \gamma < 1.$$

□

3 Some topics from mathematical control theory

The most relevant topics from the mathematical control theory for this project are the Lyapunov stability theory and dynamical programming. Due to the mature of reinforcement learning our presentation will be focus on the discrete-time problem. We start with a brief discussion of the Lyapunov stability theory where we study particularly linear Lyapunov function which is an key point in the convergence analysis later. Then we give a more detailed presentation of dynamic programming in deterministic as well as stochastic derivation, in particular, the solution of the linear quadratic regulator problem (LQ).

The dynamical system we consider is the form

$$x_{k+1} = f(x_k),\ x_0 \text{ is given}, \tag{3}$$

with the time $t \in \mathbb{Z}_+$ and the state $x \in \mathbb{R}^n$ and f a sufficiently smooth function from $\mathbb{R}^n$ to $\mathbb{R}^n$. Note that any iterative numerical algorithm can be thought of as a discrete-time dynamical system formulated here. For example the gradient decent method for finding the optimum of a (convex) differential function g over $x \in \mathbb{R}^n$:

$$x_{k+1} = x_k - \alpha_k \nabla g(x_k), \qquad k = 0, 1, 2, \ldots$$

with a proper choice of the initial point x_0, where $\alpha_k > 0$ is the step size, either constant or determined by a line search. The main question is for which choice of x_0 and the step size α_k, the algorithm converges to the optimum x^* of g, or equivalently, $\nabla g(x_k) \to \nabla G(x^*)$, as $k \to \infty$. This is, in fact, a question of stability in the term of a dynamical system as shown in the sequel.

3.1 The Lyapunov stability theory

Definition 3.1 (Equilibrium/fixed point [11])**.** If the point x^{eq} satisfies $x = f(x)$ then x^{eq} is called an equilibrium point, or a fixed point, of the dynamical system (3).

We will use equilibrium or equilibrium point and fixed point exchangeably in this report. The following stability concepts characterize the system behaviour around an equilibrium point.

Definition 3.2 (Stability [11])**.** An equilibrium x^{eq} of (3) is

- *stable* if for every $\varepsilon > 0$, there exists $\delta > 0$ (possibly dependent on ε) such that

$$\|x_0 - x^{\text{eq}}\| \leq \delta \Rightarrow \|x_t - x^{\text{eq}}\| \leq \varepsilon \quad \text{for all } t \geq 0;$$

- *asymptotically stable*, if it is satble and δ can be chosen such that

$$\|x_0 - x^{\text{eq}}\| \leq \delta \Rightarrow (x_t \to x^{\text{eq}}) \text{ as } t \to \infty;$$

- *unstable*, if it is not stable.

Note that these are local notions of stability. It is apparent to see that the convergence question of the gradient descent method is the problem of asymptotical stability. And the choice of initial condition x_0 is a problem of global convergence, to which we cane make use Lyapunov function.

Definition 3.3 (Positive (semi)definite function, [12])**.** A continuous function $V : D \to \mathbb{R}$ is said to be a positive semi-definite function if

(a) $V(x) \geq 0, \forall x$.

V is positive definite if it satisfies

(a) $V(x) \geq \forall x$,

(b) $V(0) = 0$ if and only if $x = 0$, and

(c) $V(x) \to \infty$ as $\|x\| \to \infty$.

Definition 3.4 (Lyapunov function, [12])**.** Let $x = 0$ be an equilibrium point for (3) with local Lipschitz f in $D \subset \mathbb{R}^n$ and $0 \in D$. A continuous function $V : D \to \mathbb{R}$ is said to be a *Lyapunov function* if

(a) $V(0) = 0$ and $V(x) > 0$, $\forall x \in D \setminus \{0\}$

(b) $V(f(x)) - V(x) \leq 0$, $\forall x \in D$.

Note that the condition in (a) means 0 is a local minimum of V and the condition in (b) means "energy" decreases along system trajectories.

Theorem 3.5 (Boundedness [11]). *If there exists a continuous function $V(x)$ whose sublevel sets*

$$\mathcal{L}_V(\alpha) := \{x : V(x) \leq \alpha\}$$

are bounded for every value of α and

$$\Delta V(x) = V(f(x)) - V(x) \leq 0, \forall x$$

then every trajectory of (3) is bounded

Proof. Note that, $V(f(x_t)) = V(x_{t+1})$ by (3) which together with the assumption yields

$$\Delta V(x_k) = V(f(x_k)) - V(x_k) = V(x_{k+1}) - V(x_t) \leq 0,$$

Then

$$V(x_t) = V(x_0) - V(x_0) + V(x_1) - V(x_1) + \cdots + V(x_{t-1}) - V(x_{t-1}) + V(x_t)$$

$$= V(x_0) + \sum_{k=0}^{t-1} \Delta V(x_k) \leq V(x_0)$$

for every possible value of x_k, and every trajectory lies in the set

$$\mathcal{L}_V(V(x_0))) = \{x : V(x) \leq V(x_0)\}$$

which is bounded by the assumption. □

Notice that the proof indicates that if V is decreasing along system trajectories, then once the state enters a level set of V it never leaves this set. The assumption of bounded level sets ensures that the state remains bounded, and hence that the system is stable.

Theorem 3.6 ([11]). *Let $x = 0$ be an equilibrium point for (3). If there exists a continuous function $V : \mathbb{R}^n \to \mathbb{R}$ such that*

(a) $V(x)$ is positive definite, and

(b) for some positive semi-definite function $l(x)$,

$$V(f(x)) - V(x) \leq -l(x),$$

then along all trajectories $\{x_t\}$ *generated by* (3) *it holds that* $l(x_t) \to 0$ *as* $t \to \infty$. *If, in addition,* $l(x)$ *is positive definite, then* $x_t \to 0$ *as* $t \to \infty$.

Proof. By (b) and (3) we obtain

$$V(t_{k+1}) - V(x_t) \leq -l(x_t), \ \forall x_t$$

Then

$$\sum_{t=0}^{T}(V(t_{k+1}) - V(x_t)) \leq -\sum_{t=0}^{T} l(x_t).$$

Hence

$$\sum_{t=0}^{T} l(x_t) \leq V(x_0) - V(x_T).$$

Since $V(x_t) \geq 0$ by (a), and $\{V(x_t)\}$ is nondecreasing by (b), the right hand side will converge to a finite limit as $T \to \infty$. Hence, $\sum_{t=0}^{T} l(x_t)$ is bounded, which together with $l(x)$ is a positive semidefinite function (thus $\sum_{t=0}^{T} l(x_t)$ is nondecreasing), yields that $\lim_{T\to\infty} \sum_{t=0}^{T} l(x_t)$ exists. Therefore, $l(x_t$ must have the limit 0 as $T \to \infty$.

If $l(x)$ is positive definite, $l(x_t) \to 0$ implies that $x_t \to 0$ and asymptotic stability follows. The proof is complete. □

This theorems tells us that the existence of a Lyapunov function implies stability. Next we study the Lyapunov stability of linear systems in the form

$$x_{t+1} = Ax_t. \tag{4}$$

Theorem 3.7 ([11])**.** *The linear system described in* (4) *is asymptotically stable if and only if, for any positive definite matrix* Q, *the Lyapunov equation*

$$A^\top PA - P + Q = 0 \tag{5}$$

has a positive definite solusion P. *In addition, for any given* $Q > 0$ *the solution* P *is unique.*

Proof. Assume that (4) is asymptotically stable. Note that we have an explicit expression of the solution of (4): $x_t = A^t x_0$. It is not so hard to show that if (4) is asymptotically stable, i.e. $x_t \to 0$ if and only if $\rho(A) < 1$. Hence

$$P = \sum_{k=0}^{\infty} (A^\top)^k Q A^k$$

exists and satisfies (4).

Now we prove the other direction. Let Q be an arbitrary positive definite matrix and assume (5) has a positive definite solution P. We aim to finding a Lyapunov function. We calim that $V(x) = x^\top Px$ is sucha a function. This follows by the following computation:

$$V(x_{t+1}) - V(x_t) = x_{t+1}^\top P x_{t+1} - x_t^\top P x_t = x_t^\top (A^\top PA - P)x_t = -x_t^\top Q x_t$$

Using $l(x) = x^\top Qx$ in the previous theorem, we obtain the asymptotic stability.

Finally we prove that P is a unique solution. Assume (5) has two solutions P_1 and P_2. Then

$$A^\top (P_1 - P_2)A - (P_1 - P_2) = 0$$

Inductively

$$P_1 - P_2 = A^\top (P_1 - P_2)A = \cdots = \lim_{k\to\infty} (A^\top)^k (P_1 - P_2)A^k = 0$$

where the last equality follows from stability of A. □

3.2 Deterministic dynamic programming for discrete-time finite-horizon

The basic optimal control problem we consider in this section is in the following form

$$\begin{aligned} \text{minimize} \quad & J(u) := \sum_{t=0}^{T-1} g_t(x_t, u_t) + g_T(x_T) \\ \text{subject to} \quad & x_{t+1} = f_t(x_t, u_t), t = 0, 1, ..., T-1 \\ & u_t \in \mathcal{U}_t(x_t), t = 0, 1, ..., T-1. \end{aligned}$$

We want to find an optimal policy $\mu^* = \{\mu_0^*, \mu_1^*, ..., \mu_{T-1}^*\}$ such that $u_t = \mu_t^*(x_t)$ defines the optimal action from state x_t in stage t. With the optimal policy the optimal cost is

$$J_0(x_0) = \sum_{t=0}^{T-1} g_t(x_t, \mu_t(x_t)) + g_T(x_T)$$

where $x_{t+1} = f_t(x_t, \mu_t^*(x_t))$. Note that J_0 is a function, mapping each initial state x_0 into the cost accumulated along the corresponding optimal trajectory.

A key concept in dynamic programming is called the *Bellman principle of opti-*

mality:

> *Any optimal policy has the property that, whatever the current state and decision, the remaining decisions must constitute an optimal policy with regard to the state resulting from the current decision.*

In simple terms, if you are on the best path, every step that you take along that path must also be the best choice. Applying this principle, dynamic programming solves problems by starting at the end and working backwards.

To apply the Bellman optimality principle we define the cost-to-go function: at an arbitrary stage t,

$$J_t(x_t) = \sum_{k=t}^{T-1} g_k(x_k, \mu_k(x_k)) + g_T(x_T).$$

Its optimal value

$$J_t^*(x_t) = \min_{\mu_t,\dots,\mu_{T-1}} \sum_{k=t}^{T-1} g_k(x_k, \mu_k(x_k)) + g_T(x_T).$$

is called value function, which is often denoted $V_t(x_t)(= J_t^*(x_t))$.

By the Bellman Optimality principle the truncated policy $\{\mu_{t+1}^*, ..., \mu_{T-1}^*\}$ is optimal from stage $t+1$ and onward. We can therefore express the value function as

$$\begin{aligned} V_t(x_t) &= \min_{\mu_t}[g_t(x_t, \mu_t(x_t)) + \sum_{k=t+1}^{T-1} g_k(x_k, \mu_k^*(x_k)) + g_T(x_T)] \\ &= \min_{\mu_t}[g_t(x_t, \mu_t(x_t)) + V_{t+1}(x_{t+1})] \\ &= \min_{\mu_t}[g_t(x_t, \mu_t(x_t)) + V_{t+1}(f_t(x_t, \mu_t(x_t)))]. \end{aligned}$$

Hence, given the value function V_{t+1} at stage $t+1$, we can compute the valuefunction at stage t by optimizing the sum of the cost of stage t and the cost-to-go from the resulting state. Noticing that $V_T(x_T) = g_T(x_T)$ leads to the following result.

Theorem 3.8. *[DP algorithm [11]] For every initial state x_0, the optimal cost of the basic optimal control problem is equal to $V_0(x_0)$, given by the last step of the following algorithm, which proceeds backwards from stage $T-1$ to stage 0:*

$$V_T(x) = g_T(x) \tag{6}$$

$$V_t(x) = \min_{u \in \mathcal{U}_t(x_t)} [g_t(x, u) + V_{t+1}(f_t(x, u))], \quad t = T-1, T-2, ..., 0. \tag{7}$$

If $u_t^* = \mu_t^*(x)$ minimizes the right hand-side of (7) for each x and each t, then the policy $\{\mu_0^*, \mu_1^*, ..., \mu_{T-1}^*\}$ is optimal.

3.3 Deterministic finite-horizon linear-quadratic optimal control

As an example of using DP, we consider the discrete-time system $x_{t+1} = Ax_t + Bu_t$ where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$. We want to find $\{u_0, u_1, ..., u_{T-1}\}$ so that the quadratic cost function

$$\sum_{t=0}^{T-1} (x_t^\top Q x_t + u_t^\top R u_t) + x_T^\top Q_T x_T$$

is minimized, where $Q = Q^\top \succeq 0$, $Q_T = Q_T^\top \succeq 0$ and $R = R^\top \succ 0$ are given matrices of appropriate dimensions.

The first term in the cost function measures state deviation. The second term measures input size or actuator authority. And last term measures final state deviation. The matrices Q, R set relative weights of state deviation and input usage. The above problem is called linear quadratic regulator problem.

The stage cost captures the trade-off between making the state vector converge quickly to zero and using control inputs with small energy. In particular, choosing a Q that is large relative to R makes state deviations more costly, and leads to an optimal controller that steers the states quickly to zero. Conversely, increasing the size of R shifts the focus of the stage cost to the control signal, leading to a lower energy input and state trajectories that tend to be closer to the open loop system?s natural response. The terminal cost matrix Q_T plays a more subtle role, but larger values of Q_T will reinforce a desire to drive the terminal state to rest at the origin.

Theorem 3.9 ([11]). *The finite-horizon linear quadratic optimal control problem*

$$\begin{aligned} \textit{minimize} \quad & = \sum_{t=0}^{T-1} (x_t^\top Q x_t + u_t^\top R u_t) + x_T^\top Q_T x_T \\ \textit{subject tov} \quad & x_{t+1} = Ax_t + Bu_t, \quad t = 0, 1, ..., T-1 \end{aligned}$$

with $Q = Q^\top \succeq 0$, $Q_f = Q_f^\top \succeq 0$ and $R = R^\top \succ 0$ has the optimal solution $u_t = -L_t x_t$ where

$$L_t = (B^\top P_{t+1} B + R)^{-1} B^\top P_{t+1} A \tag{8}$$

and P_t satisfies the Riccati recursion

$$P_t = Q + A^\top P_{t+1} A - A^\top P_{t+1} B (B^\top P_{t+1} B + R)^{-1} B^\top P_{t+1} A \tag{9}$$

with boundary condition $P_T = Q_T$. The minimal value of the cost function is $x_0^\top P_0 x_0$.

Proof. We apply DP to derive the optimal control and show that the value function is quadratic $V_t(x_t) = x_t^\top P_t x_t$ by induction. Note that at $t = T$, $P_T = Q_T$. Thus $V_T(x_T) = x_T^\top Q_t x_t$ is quadratic. Now assume that the value function at $t+1$ is $V_{t+1}(x_{t+1}) = x_{t+1}^\top P_{t+1} x_{t+1}$ with P_{t+1} satisfying the Riccati recursion (9). We want to show that $V_t(x_t) = x_t^\top P_t x_t$. At this stage, the dynamic programming algorithm (7) gives

$$\begin{aligned} V_t(x_t) &= \min_{u_t} [x_t^\top Q x_t + u_t^\top R u_t + V_{t+1}(Ax_t + Bu_t)] \\ &= \min_{u_t} [x_t^\top Q x_t + u_t^\top R u_t + (Ax_t + Bu_t)^\top P_{t+1} (Ax_t + Bu_t)] \\ &= \min_{u_t} [u_t^\top (B^\top P_{t+1} B + R) u_t + 2 u_t^\top B^\top P_{t+1} A x_t + x_t^\top (A^\top P_{t+1} A + Q) x_t] \end{aligned}$$

Since $B^\top P_{t+1} B + R \succ 0$ V_t is a convex quadratic function in u_t and the minimizer can be obtained by first order optimality condition which gives

$$u_t^*(x_t) = -(B^\top P_{t+1} B + R) B^\top P_{t+1} A x_t =: L_t x_t$$

and the associated value function is

$$V_t(x_t) = x_t^\top (Q + A^\top P_{t+1} A - A^\top P_{t+1} B (B^\top P_{t+1} B + R)^{-1} B^\top P_{t+1} A) x_t =: x^\top P_t x_t.$$

Since both stage costs and terminal costs are non-negative, the value function must also be non-negative and P_t must be a positive semidefinite matrix. Hence, by induction, the value function remains quadratic and positive semidefinite for $t = T, T-1, ..., 1, 0$. Observe that the expressions for L_t and P_t derived above are the same as in (8) and (9), respectively, completing the proof □

3.4 Infinite-horizon linear-quadratic control: optimality and stability

It is natural to ask if the solution presented in the previous theorem remains valid when we are interested in the behavior of the closed-loop system over an infinite

horizon, that is, we consider the limiting behavior when $T \to \infty$. The following questions arise.

1. When does there exist a *bounded limiting* solution $P(0) = P_\infty$ to the Riccati recursion (9) for all choices of $P_T = Q_T \succeq 0$?

2. When does there exist a *unique limiting* solution $P(0) = P_\infty$ to the Riccati recursion (9) regardless of the choice of $P_T = Q_T \succeq 0$?

3. When does the *limiting* solution $P(0) = P_\infty$ to the Riccati recursion (9) yield an asymptotically stable closed loop system? That is $A_c = A - BK_\infty$ is Schur (all eigenvalues inside the unit cicle) with $K_\infty = (R + B^\top P_\infty B)^{-1} B^\top P_\infty A$.

It turns out that the convergence of the Riccati recursion in general is more complicated, [13]. However we present the following answers,[14] without proofs. To this end we need some more concepts from control theory.

Definition 3.10. Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$. We say that (A, B) is controllable if the matrix $(B\ AB\ \cdots\ A^{n-1}B)$ has full rank and (C, A) is observable if the matrix $(C; CA; \cdots; CA^{n-1})$ has full rank, where we borrowed the Matlab syntax ";" for writing a block column matrix.

Theorem 3.11. *Consider the infinite-horizon linear-quadratic regulator problem*

$$\begin{aligned} \textit{minimize} \quad & \sum_{t=0}^{\infty} (x_t^\top Q x_t + u_t^\top R u_t) \\ \textit{subject to} \quad & x_{t+1} = Ax_t + Bu_t, \quad t = 0, 1, ..., T-1 \end{aligned}$$

with $Q = Q^\top \succeq 0$, and $R = R^\top \succ 0$. If (A, B) is controllable then the optimal cost is bounded. If, in addition, (C, A) is observable, where C is a full rank factorizatoin of $Q \succeq 0$ such that $Q = CC^\top$ then the discrete algebraic Riccati equation

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

has a unique positive semi-definite solution, and the optimal control policy $u_t = -Lx_t$ with

$$L = (R + B^\top PB)^{-1} B^\top PA$$

results in an asymptotically stable closed-loop system.

Briefly speaking the controllability guarantees the finiteness of the objective function and the observability ensures the stability.

These results are also very important in practice of the reason that we often use the steady state regulator from the infinite-horizon LQ control problem for finite-horizon feedback because P_t usually converges rapidly as t decreases below T.

3.5 Stochastic dynamic programming for discrete-time finite-horizon

Following [15] we derive the dynamic programming for optimal control problem. The problem setting is to minimize the *expected total cost*

$$J^{\pi}(x_0) = \mathbb{E}^{\pi}\left[g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k) \,\middle|\, x_0 \right]$$

along the stochastic dynamical system

$$x_{k+1} = f_k(x_k, u_k, w_k), \qquad k = 0, 1, \ldots, N-1,$$

for a given initial state x_0 and an admissible policy π according to the description below, where

- x_k is the state of the system at time k, belonging to a state space $\mathcal{S}_k$,
- u_k is the control input at time k, taking values in a control space $\mathcal{U}_k$,
- w_k is a random disturbance affecting the system at time k, characterized by a conditional probability distribution $P_k(\cdot \,|\, x_k, u_k)$, and
- $f_k(\cdot)$ is the state transition function that may depend explicitly on x_k, u_k, and w_k, but not on past disturbances $w_{k-1}, \ldots, w_0$.
- the expectation $\mathbb{E}^{\pi}[\cdot]$ is taken with respect to the joint probability distribution of the random disturbances $\{w_k\}$ and the state trajectory $\{x_k\}$ generated by the policy π and the system dynamics.

At each stage k, the control u_k is chosen from an admissible set $U(x_k)$ that depends on the current state x_k; and $g_k(x_k, u_k, w_k)$ denote the stage cost incurred at time k, and $g_N(x_N)$ is the terminal cost. The control u_k is chosen according to a *policy*

$\pi = \{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$, where each μ_k maps states to controls, $u_k = \mu_k(x_k)$, and satisfies the constraint $\mu_k(x_k) \in U_k(x_k)$ for all x_k. Such policies are called *admissible*.

The goal of stochastic dynamic programming is to find an *optimal policy* π^* that minimizes the expected total cost over all admissible policies, i.e.,

$$J^*(x_0) = \min_{\pi \in \Pi} J^\pi(x_0),$$

where Π denotes the set of all admissible policies.

Theorem 3.12 (DP Algorithm for Stochastic Finite Horizon Problems)**.** *Start with*

$$J_N^*(x_N) = g_N(x_N), \tag{1.12}$$

and for $k = 0, 1, \ldots, N-1$, let

$$J_k^*(x_k) = \min_{u_k \in U_k(x_k)} \mathbb{E}\Big\{ g_k(x_k, u_k, w_k) + J_{k+1}^*\Big(f_k(x_k, u_k, w_k)\Big)\Big\}. \tag{1.13}$$

If $u_k^* = \mu_k^*(x_k)$ minimizes the right-hand side of this equation for each x_k and k, then the policy

$$\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*\}$$

is optimal.

Moreover, for every initial state x_0, the optimal cost equals the DP output:

$$J^*(x_0) = \min_{\pi} J^\pi(x_0) = J_0^*(x_0),$$

and any selector $\mu_k^*(x_k)$ attaining the inner minima yields an optimal policy $\pi^* = \{\mu_0^*, \ldots, \mu_{N-1}^*\}$.

Proof. We proceed by backward induction. In thebase case ($k = N$): At the terminal stage, the optimal cost-to-go is the terminal cost: $J_N^*(x_N) = g_N(x_N)$, so the claim holds.

Induction hypothesis: Assume for some $k+1 \leq N$ that the statement holds for all stages $t \in \{k+1, \ldots, N\}$, i.e., for every state x_t,

$$J_t^*(x_t) = \min_{u_t \in U_t(x_t)} \mathbb{E}\Big[g_t(x_t, u_t, w_t) + J_{t+1}^*\Big(f_t(x_t, u_t, w_t)\Big) \,\Big|\, x_t, u_t\Big],$$

and the minimal expected tail cost from stage t is $J_t^*(x_t)$.

Inductive step ($k+1 \Rightarrow k$). Assume that the statement holds for all stages $t = k+1, \ldots, N$. We now show that it also holds for stage k. For any admissible policy $\pi = \{\mu_k, \mu_{k+1}, \ldots, \mu_{N-1}\}$ starting at x_k, the expected tail cost satisfies (law of iterated expectations)

$$\begin{aligned} J_k^\pi(x_k) &= \mathbb{E}^\pi\left[\sum_{m=k}^{N-1} g_m(x_m, \mu_m(x_m), w_m) + g_N(x_N) \,\middle|\, x_k\right] \\ &= \mathbb{E}\left[g_k(x_k, \mu_k(x_k), w_k) + \mathbb{E}^\pi\left[\sum_{m=k+1}^{N-1} g_m(\cdot) + g_N(x_N) \,\middle|\, x_{k+1}\right] \,\middle|\, x_k\right] \\ &= \mathbb{E}\Big[g_k(x_k, \mu_k(x_k), w_k) + J_{k+1}^\pi(x_{k+1}) \,\Big|\, x_k\Big]. \end{aligned}$$

By the induction hypothesis, $J_{k+1}^\pi(x_{k+1}) \geq J_{k+1}^*(x_{k+1})$ almost surely, hence

$$J_k^\pi(x_k) \geq \mathbb{E}\Big[g_k(x_k, \mu_k(x_k), w_k) + J_{k+1}^*(x_{k+1}) \,\Big|\, x_k\Big].$$

Minimizing first over the tail policy (i.e., replacing J_{k+1}^π by J_{k+1}^*), and then over the current control $u_k = \mu_k(x_k) \in U_k(x_k)$, yields

$$\inf_\pi J_k^\pi(x_k) = \min_{u_k \in U_k(x_k)} \mathbb{E}\Big[g_k(x_k, u_k, w_k) + J_{k+1}^*\Big(f_k(x_k, u_k, w_k)\Big) \,\Big|\, x_k, u_k\Big] = J_k^*(x_k).$$

Selecting at stage k a minimizer $\mu_k^*(x_k)$ achieves this infimum and leaves a tail problem from $k+1$ that is solved optimally by the induction hypothesis. Therefore $J_k^*(x_k)$ is the minimal expected cost from stage k.

Applying the argument down to $k = 0$ gives $J^*(x_0) = J_0^*(x_0)$ and shows that the policy π^* formed by the stagewise minimizers is optimal. □

3.6 Stochastic linear-quadratic optimal control problem

The model is similar to the one used with LQR, except we now have process noise.

$$x_{t+1} = Ax_t + Bu_t + w_t, \ t = 0, 1, ..., T-1$$

where w_t is the process noise or disturbance at time t normally distributed with mean zero and variance:

$$w_t \sim \mathcal{N}(0, \Sigma_w).$$

For simplicity we consider the case where x_0 is deterministic (independent of w_t).

The stochastic LQ control problem can be formulated as

$$J = \min_{u_0,\ldots,u_{T-1}} \mathbb{E}\left[\sum_{t=0}^{T-1}(x_t^\top Q x_t + u_t^\top R u_t) + x_T^\top Q_T x_T\right]$$

subject to $x_{t+1} = Ax_t + Bu_t + w_t,\ t = 0, 1, ..., T-1$.

Let $V_t(z)$ be optimal value of objective from t on starting at $x_t = z$

$$V_t(z) = \min_{u_t,\ldots,u_{T-1}} \mathbb{E}\left[\sum_{k=t}^{T-1}(x_t^\top Q x_t + u_t^\top R u_t) + x_T^\top Q_T x_T\right]$$

subject to $x_{k+1} = Ax_k + Bu_k + w_k\ t = t, 1, ..., T-1$. Then we have $V_T(z) = z^\top Q_T z$ and $J^* = \mathbb{E}V_0(x_0)$. By DP algorithm in the previous theorem V_t can be found by backward recursion as done for deterministic LQ. In other words, we will prove that $V(z) = z^\top P_t z + r_t$. At $t = T$, we have $P_T = Q_T$ and $r_T = 0$.

We use induction to prove that Vt has the required form. It holds for T, so assume it holds for $t+1$ and we will show that it holds for t. Substitute the formula for V_{t+1} into

$$V_t(x) = \min_v [z^\top Q z + u^\top R u + \mathbb{E}_w V_{t+1}(Az + Bu + w_t)]$$

we obtain

$$\begin{aligned}
V_t(x) &= \min_u [z^\top Q z + u^\top R u + \mathbb{E}_w[(Az + Bu + w_t)^\top P_{t+1}(Az + Bu + w_t) + r_{t+1}] \\
&= \min_u [z^\top Q z + u^\top R u + (Az + Bu)^\top P_{t+1}(Az + Bu) + \mathrm{tr}(P_{t+1}\Sigma_w) + r_{t+1}] \\
&= \underbrace{\min_u \begin{pmatrix} z \\ u \end{pmatrix}^\top \begin{pmatrix} A^\top P_{t+1} A + Q & A^\top P_{t+1} B \\ B^\top P_{t+1} A & B^\top P_{t+1} B + R \end{pmatrix} \begin{pmatrix} z \\ u \end{pmatrix}}_{\text{same as deterministic LQR}} + \underbrace{\mathrm{tr}(P_{t+1}\Sigma_w) + r_{t+1}}_{\text{constant term}}
\end{aligned}$$

Therefore we can write that $V_t(z) = z^\top P_t z + r_t$ with

$$\begin{aligned}
P_T &= Q_T \\
P_t &= Q + A^\top P_{t+1} A - A^\top P_{t+1} B(B^\top P_{t+1} B + R)^{-1} B^\top P_{t+1} A \\
r_T &= 0 \\
r_t &= \mathrm{tr}(P_{t+1}\Sigma_w) + r_{t+1} \\
L_t &= -(B^\top P_{t+1} B + R)^{-1} B^\top P_{t+1} A
\end{aligned}$$

So the optimal policy for stochastic LQR is $u_t = L_t x_t$, where L_t is the same LQR feedback as in the deterministic LQR case. The total cost using the optimal policy is:

$$V_0(x_0) = x_0^\top P_0 x_0 + r_0 = x_0^\top P_0 x_0 + \sum_{t=1}^{T} \mathrm{tr}(P_t \Sigma_w)$$

if x_0 is deterministic. We see that the Riccati recursion and the feedback do not change.

Note that if x_0 is also random, say $x_0 \sim \mathcal{N}(\mu_x, \Sigma_w)$ then we would haveinstead obtain

$$V_0(x_0) = \mathbb{E}(x_0^\top P_0 x_0) + r_0 = \mu_x^\top P_0 \mu_x + \mathrm{tr}(P_0 \Sigma_w) + \sum_{t=1}^{T} \mathrm{tr}(P_t \Sigma_w).$$

In stochastic LQR, we do not have $x_t \to 0$. The process noise added at every time step causes the state to meander about zero, and it never quite settles down. This is why the cost has this ever-accumulating term that will go to infinity as T grows large. For this reason, it does not make sense to talk about the *steady-state* or *infinite-horizon* cost. As the horizon tends to infinity, so does the cost. However, we can talk about the average cost. This is found by taking the average and then the limit:

$$J_{\text{avg}} = \lim_{T \to \infty} \frac{1}{T} \left(x_0^\top P_0 x_0 + \sum_{t=1}^{T} \mathrm{tr}(P_t \Sigma_w) \right) = \mathrm{tr}(P_\infty \Sigma_w)$$

where $P_\infty = \lim_{T \to \infty} P_t$ is the solution of the discrete time algebraic Riccati equation. This can be shown in a manner similar to the problem of evaluating policy in next subsection.

Remark. Based on what we just derived, we conclude that there is a fundamental equivalence between the deterministic and stochastic versions of the LQR problem. Specifically, the two following quantities are the same:

1. The expected infinite-horizon cost of a deterministic LQR problem (no process noise), where the initial state is $x_0 \sim \mathcal{N}(0, \Sigma)$.

2. The average cost of a stochastic LQR problem where the process noise is $w_t \sim \mathcal{N}(0, \Sigma)$.

3.7 Evaluating a suboptimal policy

If instead of using the optimal infinite-horizon LQR policy $u_t = Kx_t$, we use some other policy $u_t = \hat{K}x_t$ such that $A + B\hat{K}$ is Schur. We can calculate the cost by substituting directly into the formula for the standard cost

$$J = \sum_{t=0}^{\infty}(x_t^\top Q x_t + u_t^\top R u_t) = \sum_{t=0}^{\infty} x_t^\top (Q + \hat{K}^\top R\hat{K})x_t$$

Now use the fact that $x_{t+1} = (A + B\hat{K})x_t$, which yields $x_t = (A + B\hat{K})^t x_0$ and we get

$$J = \sum_{t=0}^{\infty} x_0^\top (A^\top + \hat{K}^\top + B^\top)^t (Q + \hat{K}^\top R\hat{K})(A + B\hat{K})x_0 = x_o^\top \hat{P} x_0$$

The matrix $\hat{P}$ satisfies the Lyapunov equation:

$$(A + B\hat{K})^\top \hat{P}(A + B\hat{K}) - \hat{P} + (Q + \hat{K}^\top R\hat{K}) = 0$$

So to find the cost for this suboptima $\hat{K}$, we solve the Lyapunov equation above for $\hat{P}$, and then our cost is $x_0^\top \hat{P} x_0$. Note that $A + B\hat{K}$ is stable, otherwise the cost will be infinite.

Next we turn to discussion on policy evaluation in the stochastic linear regulator problem. We consider the standard discrete-time linear system with additive Gaussian noise:

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad w_t \sim \mathcal{N}(0, \sigma_w^2 I)$$

with a quadratic cost function:

$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t$$

Given a fixed linear feedback controller $u_t(x) = Kx$, the control input becomes:

$$u_t = Kx_t$$

leading to the closed-loop dynamics:

$$x_{t+1} = (A + BK)x_t + w_t$$

In infinite-horizon average-cost LQR, for a fixed linear state-feedback policy K,

the expected per-step cost is given by:

$$\lambda^\star = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\left(x_t^\top Q x_t + u_t^\top R u_t\right)\right].$$

This defines the *long-run average stage cost* under the policy K.

We can prove that the average cost admits a closed-form.

Proposition 3.13. *Let the closed-loop system evolve as $x_{t+1} = Lx_t + w_t$, with $L = A + BK$ and $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$. Let P_K be the unique positive semidefinite solution to the Lyapunov equation*

$$P_K = L^\top P_K L + (Q + K^\top R K).$$

Then the long-run average cost per time step under policy K is

$$\lambda^* = \sigma_w^2 \operatorname{tr}(P_K).$$

Proof. At stationarity, the state covariance Σ_x satisfies

$$\Sigma_x = L\Sigma_x L^\top + \sigma_w^2 I.$$

Premultiplying the Lyapunov equation by Σ_x and applying the cyclic trace identity yields

$$\operatorname{tr}((Q + K^\top R K)\Sigma_x) = \operatorname{tr}(P_K \sigma_w^2 I) = \sigma_w^2 \operatorname{tr}(P_K).$$

□

This tells us that the average cost admits a closed-form:

$$\lambda^\star = \sigma_w^2 \operatorname{tr}(P^\star),$$

where $P^\star \in \mathbb{R}^{n\times n}$ is the solution to the Lyapunov equation:

$$P^\star = L^{\star\top} P^\star L^\star - P^\star + Q + K^\top R K, \quad \text{with} \quad L^\star = A^\star + B^\star K.$$

Here, $P^\star$ is the value function matrix for policy K, and $\lambda^\star$ summarizes the expected steady-state cost per time step.

This compact expression shows that the per-step cost is simply the noise variance scaled by the trace of the value matrix.

3.8 Relationship to reinforcement learning

The standard components of an RL system are [15]:

- **State space** X: The set of all possible states the environment can be in. At time t, the environment is in state $x_t \in X$.
- **Action space** U: The set of all possible actions the agent can take. At time t, the agent selects an action $u_t \in U$.
- **Transition dynamics**: A (possibly unknown) rule that determines how the environment evolves. In many cases, this is modeled as a probabilistic function $x_{t+1} \sim p(\cdot \mid x_t, u_t)$.
- **Cost function**: A function$c(x_t, u_t)$ that provides feedback to the agent about the quality of its action in a given state.
- **Policy** π: A rule that tells the agent which action to take in each state.

To understand the relationship between the reinforcement learning and optimal control, in particular, LQR we provide brief introduction of reinforcement learning. First we define the Markov decision processes (MDP), which provides the formalism in which RL problems are usually posed.

A Markov decision process is a quintuple $(X, U, P_{sa}, \gamma, R)$, where $\gamma \in [0, 1)$ is called the discount factor, and $R : X \times U$ is the reward function and P_{xu} are the state transition probabilities..

The dynamics of an MDP proceeds as follows: We start in some state x_0, and get to choose some action $u_0 \in U$ to take in the MDP. As a result of this choice, the state of the MDP randomly transitions to some successor state x_1, drawn according to $x_1 \sim P_{x_0,u_0}$. Then, we get to pick another action u_1. Consequently the state trasitions again, now ,to some $x_2 \sim P_{x_1,u_1}$. We then pick up u_2 and so on....

$$x_0 \xrightarrow{u_0} x_1 \xrightarrow{u_1} x_2 \xrightarrow{u_2} x_3 \xrightarrow{u_3} \cdots$$

with actions $u_0, u_1, ...$ on the sequence of the states $x_0, , x_1, ...$ we get the payoff given by

$$R(x_0) + \gamma R(x_1) + \gamma^2 R(x_2) + \cdots .$$

In this thesis we will use the simpler state-rewards $R(s)$, though the generalization to state-action rewards $R(s, a)$ offers no special difficulties.

The goal in reinforcement learning is to choose actions over time so as to maximize the expected value of the total payoff:

$$\mathbb{E}[R(x_0) + \gamma R(x_1) + \gamma^2 R(x_2) + \cdots]$$

Note that the reward at timestep t is discounted by a factor of γ^t. Thus, to make this expectation large, we would like to accrue positive rewards as soon as possible (and postpone negative rewards as long as possible).

A policy is any function $\pi : X \to U$ mapping from the states to the actions. We say that we are executing some policy π if, whenever we are in state x, we take action $u = \pi(x)$. We also define the value function for a policy π according to

$$V^\pi(x) = \mathbb{E}[R(x_0) + \gamma R(x_1) + \gamma^2 R(x_2) + \cdots \ | x_0 = x, \pi]$$

In other words, $V^\pi(x)$ is just the expected sum of discounted rewards upon starting in state x, and taking actions according to π.

Given a fixed policy π, its value function V^π satisfies the Bellman equations:

$$V^\pi(x) = R(x) + \gamma \sum_{x' \in X} P_{x\pi(x)}(x') V^\pi(x')$$

This says that the expected sum of discounted rewards $V^\pi(x)$ for starting in x consists of two terms: First, the immediate reward $R(s)$ that we get right away simply for starting in state x, and second, the expected sum of future discounted rewards. Examining the second term in more detail, we see that the summation term above can be rewritten

$$\mathbb{E}_{x' \sim P_{x\pi(x)}}[V^\pi(x')].$$

This is the expected sum of discounted rewards for starting in state x') where x' is distributed according $P_{x\pi(x)}$, which is the distribution over where we will end up after taking the first action $u(x)$ in the MDP from state x. Thus, the second term above gives the expected sum of discounted rewards obtained after the first step in the MDP.

Next we define the optimal value function according to

$$V^*(x) = \max_\pi V^\pi(x).$$

In other words, this is the best possible expected sum of discounted rewards that can be attained using any policy. There is also a version of Bellman's equations for the optimal value function:

$$V^*(x) = R(x) + \max_{u \in U} \gamma \sum_{x' \in X} P_{xu}(x') V^*(x').$$

Note that the first term above is the immediate reward as before. The second term is the maximum over all actions a of the expected future sum of discounted rewards we will get upon after action u.

More generally we can formulate

$$V^*(x) = \max_{u \in U} [R(x,u) + \gamma \mathbb{E}_{x' \sim P_{xu}} [V^{\pi^*}(x')].$$

Finally we define a policy $\pi^* : X \to U$ as follows:

$$\pi^* = \arg\max_{u \in U} \sum_{x' \in X} P_{xu}(x') V^*(x').$$

As above it can be formulated in a moire general situation

$$\pi^*(x) = \arg\max_{u \in U} [R(x,u) + \gamma \mathbb{E}_{x' \sim P_{xu}} [V^{\pi^*}(x')].$$

Note that $\pi^*(x)$ gives the action u that attains the maximum. Clearly

$$V^*(x) = V^{\pi^*}(x) \geq V^{\pi}(x).$$

Note that in infinite horizon case, we have

$$\sum_{t=0}^{\infty} R(x_t, u_t) \gamma^t.$$

If the rewards are bounded by a constant $\bar{R}$, the payoff is bounded by

$$\left| \sum_{t=0}^{\infty} R(x_t, u_t) \gamma^t \right| \leq \bar{R} \sum_{t=0}^{\infty} \gamma^t < \frac{\bar{R}}{1-\gamma}.$$

Since the payoff is a finite sum the discount factor γ is not necessary anymore.

Now we turn to the LQR. We know that the state space is $X = \mathbb{R}^n$ and $U \subset \mathbb{R}^m$.

We assume linear transitions (with noise)

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad w_t \sim \mathcal{N}(0, \Sigma_w)$$

In the LQR setting the rewards

$$R^t(x_t, u_t) = -(x_t^\top Q x_t + u_t^\top R u_t)$$

which is negative because we minimize the quadratic cost. As we have shown in the Subsection 3.6 the noise, as long as it has zero mean, does not impact the optimal policy. But the value function does depend on the noise since the term r-term does.

A central distinction in reinforcement learning is between *model-based* and *model-free* approaches. Model-based methods explicitly estimate or use the system dynamics to compute an optimal control law, while model-free methods learn the optimal policy or value function directly from data.[16]

Model-Based vs. Model-Free Methods

One of the central distinctions in reinforcement learning is between **model-based** and **model-free** methods:

- **Model-Based Methods**: These methods attempt to learn or use a model of the environment?s dynamics (i.e., how states transition in response to actions). Once a model is available, control theory or planning techniques can be used to compute an optimal policy.
- **Model-Free Methods**: These methods directly learn the optimal policy or value function from data, without explicitly modeling the environment. They often rely on trial-and-error learning and can be simpler to implement but may require more data (i.e., they are less sample-efficient).

Both approaches have pros and cons. Model-based methods tend to be more sample-efficient, but can suffer if the learned model is inaccurate. Model-free methods avoid model bias, but often suffer from high variance and slower convergence.

4 Asymptotic analysis of policy evaluation

This section presents the core theoretical contributions discussed in the paper by Tu and Recht (2019) [6]. Their results compare the sample efficiency of model-based and model-free reinforcement learning methods for the Linear Quadratic Regulator (LQR) problem. The purpose of this thesis is to understand their theory and try to make the original theorems more accessible for the students at a level equivalent to our master program. To this end we also provide some commentary to clarify the intuition and implications of the underlying mathematics.

First we describe model-based and model free algorithms to make the text self-contained. Then we prove, according to the author's understanding, the three main theorems on policy evaluation in [6]. The main task is evaluating policy, described in Section 3.7, that is, for a given controller:

$$u_t = -Kx_t, K \in \mathbb{R}^{d\times n}$$

an unknown dynamical system:

$$x_{t+1} = A^\star x_t + B^\star u_t + w_t,$$

that produces a stable closed-loop system:

$$x_{t+1} = (A^* - B^*K)x_t + w_t,$$

We want to compute the (relative) value function $V^K(x)$:

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

We know from Section 3.7 that the problem can be solved by Lyapunov for given (A^*, B^*). The main difference is that we will study algorithms which only have input/output access to (A^*, B^*). 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. We are interested in the asymptotic behavior of algorithms as $T \to \infty$.

To keep this thesis self-contained, we provide a detailed descriptions of model-

based and model-free methods, respectively.

4.1 Model-Based Plugin Estimator

In model-based methods we have a controller an unknown dynamic system :

$$u_t = -Kx_t.$$

$$x_{t+1} = A^\star x_t + B^\star u_t + w_t,$$

that results in a closed-loop system:

$$x_{t+1} = (A^* - B^*K)x_t + w_t.$$

Define:

$$L^* = A^* - B^*K.$$

In the current situation, we do not know the feedback gain L^*, but we can run experiments and observe state transitions under the controller K.

Algorithm 1 Model-based Algorithm for Policy Evaluation: We run the system and observe:

$$x_t, \quad x_{t+1}.$$

we know:

$$u_t = -Kx_t.$$

Since in closed loop:

$$x_{t+1} = L^*x_t + w_t,$$

starting at $x_0 = 0$ (for simplicity) and driven by Gaussian white noise $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$. we have data pairs:

$$(x_t, x_{t+1}).$$

This is just a linear regression problem:

$$x_{t+1} = L^*x_t + w_t.$$

By collecting all samples column-wise, Let $X := \begin{pmatrix} x_0 & \cdots & x_{T-1} \end{pmatrix}$ and $X^+ :=$

$\begin{pmatrix} x_1 & \cdots & x_T \end{pmatrix}$. we can write the corresponding least-squares estimator as

$$X^+ = LX$$

whose solution is,

$$\widehat{L}(T) = (X^+X)^\top (XX^\top)^{-1}$$

by solve the normal equation

$$X^+X^\top = L(XX^\top)$$

or in terms of x_t,

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

But the matrix

$$\sum_{t=0}^{T-1} x_t x_t^\top$$

might be:

- singular, or
- very close to singular (i.e., has very small eigenvalues),

Although in the first mentioned case we can use any pseudo-inverse, it can still suffer numerical instability as in the second case. To overcome the ill-conditioning we make use of regularization, here we use the Tikhnov-type of regularization in order to take care of both cases. Introduce a parameter $\lambda > 0$. Solve the following optimization problem:

$$\min \|X^+ - LX\|_F^2 + \lambda^2 \|L\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm. Using the property of the Frobenius norm, we can re-write this problem as a new least-square problem in the form,

$$\min \left\| \begin{pmatrix} X^+ & 0 \end{pmatrix} - L \begin{pmatrix} X & \lambda I \end{pmatrix} \right\|_F^2.$$

As before it has a solution

$$\widehat{L}(T) = (X^+X^\top)(XX^\top + \lambda I)^{-1},$$

or equivalently

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

The matrix

$$\sum_{t=0}^{T-1} x_t x_t^\top + \lambda I$$

- is better conditioned (moving eigenvalues away from 0)
- stabilizes the numerical inversion.
- recovers the original regression problem as $\lambda \to 0$.

To see these points we perform an SVD on the matrix X: $X = U\begin{pmatrix}\Sigma_1 & 0\end{pmatrix}V^\top$, where Σ_1 is diagonal with the singular values of X, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_T \geq 0$, on the diagonal, and the matrices U and V are orthogonal with appropriate dimensions. A straightforward computation gives

$$X^\top(XX^\top + \lambda^2)^{-1} = V\begin{pmatrix}\Sigma_1 \\ 0\end{pmatrix}\begin{pmatrix}\frac{1}{\sigma_1^2+\lambda^2} & & & \\ & \frac{1}{\sigma_1^2+\lambda^2} & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_T^2+\lambda^2}\end{pmatrix}U^\top$$

$$= V\begin{pmatrix}\frac{\sigma_1}{\sigma_1^2+\lambda^2} & & & \\ & \frac{\sigma_2}{\sigma_1^2+\lambda^2} & & \\ & & \ddots & \\ & & & \frac{\sigma_T}{\sigma_T^2+\lambda^2} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0\end{pmatrix}U^\top$$

As $\lambda \to 0$ the diagonal block matrix in right hand side above tends to a diagonal matrix with $\frac{1}{\sigma_1}$, ..., $\frac{1}{\sigma_\ell}, 0, ..., 0$ where $\sigma_\ell > 0$ and $\ell \leq T$. Note also that this gives the computation of the pseudo-inverse.

It is worth pointing out that typical values for λ, in practical computations, are small, e.g., $\lambda = 0.001$ or 0.01.

Moreover, we introduce threshold criteria during model-based estimation to guarantee stability and numerical robustness. They are the spectral radius and the operator norm of the estimated matrix $\hat{L}(T)$, respectively:

$$\rho(\widehat{L}(T)) \leq \zeta, \qquad \|\widehat{L}(T)\| \leq \psi,$$

where $\zeta \in (0,1)$ and $\psi > 0$ are fixed constants.

The spectral radius condition $\rho(\widehat{L}(T)) \leq \zeta$ guarantees that the estimated closed-loop dynamics remain stable in discrete time, that is, all eigenvalues of $\widehat{L}(T)$ lie strictly inside the unit circle. Consider the discrete-time linear system

$$x_{t+1} = Lx_t, \qquad x_0 \in \mathbb{R}^n. \tag{10}$$

Stability means that the state x_t tends to zero as $t \to \infty$, i.e.,

$$\lim_{t\to\infty} x_t = 0 \quad \text{for all initial states } x_0. \tag{11}$$

For the initial state x_0, we have

$$x_t = L^t x_0. \tag{12}$$

This happens exactly when all eigenvalues of L lie strictly inside the unit circle in the complex plane, that is,

$$\rho(L) = \max_i |\lambda_i(L)| < 1. \tag{13}$$

The reason is that the matrix power L^t behaves like its eigenvalues raised to the power t: if all eigenvalues have magnitude smaller than one, then $L^t \to 0$ as $t \to \infty$, which implies $\lim_{t\to\infty} x_t = 0$ for any initial condition x_0.

The norm bound $\|\widehat{L}(T)\| \leq \psi$ prevents the occurrence of excessively large transient responses, thereby improving numerical stability in subsequent computations such as the Lyapunov equation solution. Although the spectral radius condition $\rho(L) < 1$ is sufficient to guarantee the *asymptotic stability* of the discrete-time system $x_{t+1} = Lx_t$, it does not necessarily ensure desirable numerical or transient properties. Asymptotic stability means that all eigenvalues of L lie strictly inside the unit circle in the complex plane, which implies that $\lim_{t\to\infty} L^t = 0$ and consequently $\lim_{t\to\infty} x_t = 0$ for any initial condition x_0. However, this criterion provides

information only about the long-term behavior of the system.

In practice, the matrix L may be *non-normal*, that is, it may not commute with its transpose. In such cases, even when $\rho(L) < 1$, the intermediate powers L^t can exhibit large transient amplifications before eventually decaying to zero. This phenomenon can lead to numerical instability or ill-conditioning in subsequent computations, such as solving the discrete-time Lyapunov equation used for policy evaluation.

If either of these threshold conditions is violated, the corresponding estimate is considered unreliable and is replaced by a null estimate, i.e., $\widehat{P}_{\text{plug}}(T) = 0$. This procedure ensures that only stable and well-conditioned estimates are used in the policy evaluation stage.This regularization improves the conditioning of the matrix and ensures numerical stability, avoiding extremely large or undefined values in the computation.

For the true closed-loop system:

$$P = Q + L^{*\top} P L^*.$$

Once we have $\hat{L}$, plug it in and solve:

$$\hat{P} = Q + \hat{L}^\top \hat{P} \hat{L}.$$

This is a standard discrete-time Lyapunov equation. We can solve it:

- Analytically for small systems.
- Numerically using built-in solvers.

Now we have:

$$\hat{P} \approx P.$$

The infinite-horizon cost for an initial state x_0 is:

$$J = x_0^\top P x_0.$$

So we get the estimate:

$$\hat{J} = x_0^\top \hat{P} x_0.$$

Now, we have estimated the expected cost of using K.
So, we use our data to fit L directly (closed-loop behavior). Then plug L into the

exact Lyapunov formula to get the cost.

We summarize the above discussion in the following pseudocode:

Algorithm 1 Model-based algorithm for policy evaluation.

Require: 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:

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

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

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

5: **else**

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

7: **end if**

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

4.2 Model-Free algorithm

Least-Squares Temporal Difference Learning is a popular family of algorithms for approximate policy evaluation in large Markov decision processes.

Algorithm 2 Model-Free (TD, LSTD) for Policy Evaluation At each time step, the current state is denoted by x_t, and the next state by x_{t+1}. The corresponding stage cost is given by

$$c_t = x_t^\top Q x_t + u_t^\top R u_t.$$

We want to estimate the cost-to-go (value function) for a given policy:

$$u_t = -Kx_t$$

Specifically, our objective is to estimate the matrix P^* in the value function:

$$V^K(x) = x^\top P^* x$$

This matrix P^* tells us how "costly" a state x is under policy K. The model-free

approach does not estimate how the system evolves (i.e., it does not learn A^*, B^*). It leverages the observation that the value function can be written as a linear function of $\text{svec}(xx^\top)$. This makes it possible to apply LSTD, a temporal difference method, to directly estimate the value function without building a model of the environment.

We provide the idea behind this method in this subsection.

We observe that:

$$V^K(x) = \sigma_w^2 \cdot x^\top P^* x = \sigma_w^2 \cdot \langle \text{svec}(P^*), \text{svec}(xx^\top) \rangle$$

This comes from solving the discrete-time Lyapunov equation. Here:

- $P^\star$ is a matrix that captures how cost accumulates over time under the closed-loop dynamics.
- σ_w^2 is the variance of the process noise in the system.

The quadratic form $x^\top P^\star x$ gives the cost associated with starting at state x. The equation is rewritten in a linear form using an inner product:

$$x^\top P^\star x = \langle \text{svec}(P^\star), \text{svec}(xx^\top) \rangle$$

Here:

- $\text{svec}(A)$: A function that vectorizes a symmetric matrix A by taking the upper triangle and flattening it into a vector (including proper weighting of off-diagonal terms).
- So, $\text{svec}(P^\star)$ is a vector representation of the matrix $P^\star$.
- $\text{svec}(xx^\top)$: similarly, vectorizes the outer product of the state with itself.

Thus, we now express the value function linearly in terms of a feature vector $\phi(x) := \text{svec}(xx^\top)$:

$$V^K(x) = \sigma_w^2 \langle w^\star, \phi(x) \rangle$$

where $w^\star = \text{svec}(P^\star)$.

This means that the value function is linear in the feature vector $\phi(x)$ a perfect setup for applying Temporal Difference Learning. LSTD is a model-free method:

- It does not try to learn the dynamics matrices A^*, B^*.

- Instead, it tries to directly learn the value function by solving a regression problem, using the idea that:

$$V(x_t) \approx c_t + V(x_{t+1}) - \lambda$$

where c_t is the immediate cost and λ is the average cost.

LSTD collects samples of states x_t, next states x_{t+1}, and costs, and uses them to solve a linear system to estimate $w^\star$, i.e., the value function weights. This means that the quadratic form $x^\top P^* x$ can be written as an inner product between the vectorized matrix P^* and the vectorized outer product $xx^\top$. This trick converts the nonlinear quadratic value function into a **linear regression** problem in terms of feature vectors:

$$\phi(x) := \text{svec}(xx^\top)$$

- It allows us to apply linear regression techniques (e.g., LSTD) to estimate P^*.
- It is commonly used in model-free reinforcement learning.

Uses the Bellman equation: $V(x_t) \approx c_t + V(x_{t+1})$. So each data point says: "My current value estimate should equal the cost now plus the value of the next state." TD methods adjust the value estimate (here parameterized as a quadratic form $V(x) = x^\top Px$) to make these approximate equalities hold on average. **So the data feeds a value estimator:** It never tries to figure out the system matrix L it just tries to get the value function to be consistent with the observed costs. **It's flexible:** Because, it doesn't need the model, only data. But it needs more data to achieve similar performance.

4.3 Asymptotic analysis

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, defined as:

$$\mathbb{E}[\|\hat{P} - P^\star\|_F^2]$$

where $\hat{P}$ is the estimated value function matrix, $P^\star$ is the true value function matrix, and $\|\cdot\|_F$ denotes the Frobenius norm.

Algorithm 2 Model-free algorithm for policy evaluation (LSTD) [17].

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 $\lambda_t \approx \sigma_w^2 \operatorname{tr}(P_*)$ from $\{x_t\}_{t=0}^{T}$.

3: Compute (recall that $\phi(x) = \operatorname{svec}(xx^\top)$):

$$\hat{w}_{\text{lstd}}(T) = \left(\sum_{t=0}^{T-1} \phi(x_t) \big(\phi(x_t) - \phi(x_{t+1}) \big)^\top \right)^{-1} \left(\sum_{t=0}^{T-1} (c_t - \lambda_t) \phi(x_t) \right),$$

4: Set $\hat{P}_{\text{lstd}}(T) = \operatorname{smat}(\hat{w}_{\text{lstd}}(T))$.

5: **return** $\hat{P}_{\text{lstd}}(T)$.

This quantity measures how close the estimated value function is to the true value function in terms of matrix distance. The expectation is taken over the randomness in the observed data, such as the process noise.
Our goal is to compare this risk for two algorithms:

- **Algorithm 1:** Model-based plug-in estimator.
- **Algorithm 2:** Model-free Least-Squares Temporal Difference (LSTD) estimator, which is based on the Temporal Difference (TD) learning principle.

In particular, we study the **asymptotic risk**:

$$\lim_{T \to \infty} T \cdot \mathbb{E}[\|\hat{P}(T) - P^\star\|_F^2]$$

This form reveals how the error scales with the number of samples T in the long run. A smaller asymptotic risk indicates better sample efficiency.
Our first theoretical result provides an **upper bound** on the asymptotic risk of the model-based Algorithm 1.

4.3.1 Asymptotic Risk of Model-Based Estimator

The first theorem provides an upper bound on the **asymptotic risk** of the model-based plugin estimator (Algorithm 1) for policy evaluation in Linear Quadratic Regulation (LQR). Specifically, it analyzes how accurately the estimator $\hat{P}_{\text{plug}}(T)$ approximates the true value matrix $P^\star$ as the number of samples (trajectory length T) grows.

- K: Fixed controller (feedback gain), where $u_t = Kx_t$
- $A^\star, B^\star$: Unknown system matrices
- $L^\star = A^\star + B^\star K$: Closed-loop system matrix
- $P^\star$: True value function matrix satisfying the Lyapunov equation
- $\hat{P}_{\text{plug}}(T)$: Estimated value function matrix from the model-based algorithm
- P_∞: Stationary covariance matrix of the state under policy K
- σ_w^2: Variance of the process noise
- $\|\cdot\|_F$: Frobenius norm
- $\otimes_s$: Symmetric Kronecker product

The risk is defined as:

$$\lim_{T\to\infty} T \cdot \mathbb{E}\left[\|\hat{P}_{\text{plug}}(T) - P^\star\|_F^2\right]$$

Next we want to prove the theorem that guarantees that under mild conditions (on stability and regularization parameters), the above limit is upper bounded by:

$$4 \cdot \operatorname{tr}\left[(I - L^{\star\top} \otimes_s L^{\star\top})^{-1} \left(L^{\star\top}(P^\star)^2 L^\star \otimes_s \sigma_w^2 (P_\infty)^{-1}\right) (I - L^{\star\top} \otimes_s L^{\star\top})^{-\top}\right]$$

- This expression quantifies how the accuracy of the value function estimate improves with more data.
- It shows that the estimation error decreases like $\mathcal{O}(1/T)$, and the constant depends on system matrices and noise level.
- The more stable the system (i.e., the smaller the spectral radius of $L^\star$), the smaller the risk.

For this purpose, we need some tools in the first step:

Lemma 4.1 ([6]). *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 $\lambda \geq 0$:*

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

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

$$L_\star P_\infty L_\star^\top - P_\infty + \sigma_w^2 I_n = 0.$$

We have that $\widehat{L}(T) \xrightarrow{a.s.} L_\star$, and furthermore:

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

Proof. Let $X = [x_0, x_1, \ldots, x_{T-1}]$ be data matrix and $W = [w_0, w_1, \ldots, w_{T-1}]$ be noise matrix. We know that the regularized least square problem at hand has the solution

$$\hat{L}(T) = (L_\star X + W)X^\top (XX^\top + \lambda I)^{-1},$$

from the derivation in Section 4.1. Then

$$\begin{aligned}
\hat{L}(T) - L_\star &= L_* XX^\top (XX^\top + \lambda I)^{-1} - L_* + WX^\top \left(XX^\top + \lambda I\right)^{-1} \\
&= \left(L_* XX^\top - L_*(XX^\top + \lambda I)\right)(XX^\top + \lambda I)^{-1} + WX^\top \left(XX^\top + \lambda I\right)^{-1} \\
&= -\lambda L_*(XX^\top + \lambda I)^{-1} + WX^\top \left(XX^\top + \lambda I\right)^{-1}
\end{aligned}$$

By Proposition 2.17,

$$\mathrm{vec}(\hat{L}(T) - L_\star) = \mathrm{vec}(-\lambda L_*(XX^\top + \lambda I)^{-1}) + \left(\left(XX^\top + \lambda I\right)^{-1} \otimes I\right) \cdot \mathrm{vec}(WX^\top)$$

Multipliying both sides by $\sqrt{T}$ yields

$$\begin{aligned}
&\sqrt{T}\mathrm{vec}(\hat{L}(T) - L_\star) \\
=& -\sqrt{T}\mathrm{vec}(\lambda L_*(XX^\top + \lambda I)^{-1}) + \left(\left(T^{-1}\left(XX^\top + \lambda I\right)\right)^{-1} \otimes I\right) \cdot \mathrm{vec}(T^{-1/2}WX^\top)
\end{aligned}$$

It is well-known that the process $\{x_t\}$ is geometrically ergodic because the system

$$x_{t+1} = L_\star x_t + w_t,$$

is driven by i.i.d. Gaussian noise $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$ and the matrix $L_\star$ is stable (i.e., all eigenvalues lie strictly inside the unit circle)[18]. Geometric ergodicity implies that the Markov chain $\{x_t\}$ converges to its unique stationary distribution at an exponential rate. This property ensures that the process mixes rapidly and satisfies the conditions of Central Limit Theorem for Markov chains 2.13 . Hence Theorem

2.13, together with the Cramér-Wold Theorem gives

$$\operatorname{vec}\left(T^{-1/2}WX^\top\right) = T^{-1/2}\sum_{t=1}^{T}\operatorname{vec}(w_t x_t^\top) \xrightarrow{D} \mathcal{N}\left(0,\Sigma\right), \tag{14}$$

where

$$\Sigma = \mathbb{E}_{x\sim\nu_\infty,w}\left[\operatorname{vec}(wx^\top)\operatorname{vec}(wx^\top)^\top\right].$$

We now analyze the scaled sum:

$$T^{-1/2}\sum_{t=1}^{T}\operatorname{vec}(w_t x_t^\top),$$

which is a vector-valued sum of outer products of w_t and x_t. Here:

- x_t evolves as a geometrically ergodic Markov chain,
- $w_t \sim \mathcal{N}(0,\sigma_w^2 I)$ are i.i.d. Gaussian noise, independent of x_t.

Define a function:

$$f(x_t, w_t) := \operatorname{vec}(w_t x_t^\top).$$

Then the sum becomes:

$$T^{-1/2}\sum_{t=1}^{T} f(x_t, w_t),$$

which is a functional of the augmented process $\{(x_t, w_t)\}$. Since $\{x_t\}$ is geometrically ergodic and $\{w_t\}$ is i.i.d. and independent of x_t, the joint process is also geometrically ergodic.

Theorem 2.9 states that for a geometrically ergodic Markov chain $\{x_t\}$ and a Borel function f with finite second moment:

$$T\left(\frac{1}{T}\sum_{t=1}^{T} f(x_t) - \mathbb{E}_\pi[f(x)]\right) \xrightarrow{D} \mathcal{N}(0,\Sigma_f).$$

In our case:

$$\mathbb{E}[w_t] = 0 \quad \Rightarrow \quad \mathbb{E}[w_t x_t^\top] = 0,$$

so the summands are already zero-mean, and we directly get:

$$T^{-1/2}\sum_{t=1}^{T}\operatorname{vec}(w_t x_t^\top) \xrightarrow{D} \mathcal{N}(0,\Sigma).$$

The Cramér-Wold theorem [1] that convergence of all one-dimensional projections $a^\top Z_T \to \mathcal{N}(0, a^\top \Sigma a)$ implies:

$$Z_T \xrightarrow{D} \mathcal{N}(0, \Sigma).$$

Therefore, we conclude:

$$\operatorname{vec}\left(T^{-1/2} W X^\top\right) = T^{-1/2} \sum_{t=1}^{T} \operatorname{vec}(w_t x_t^\top) \xrightarrow{D} \mathcal{N}\left(0, \Sigma\right),$$

with

$$\Sigma = \mathbb{E}_{x \sim \nu_\infty, w}\left[\operatorname{vec}(w x^\top)\operatorname{vec}(w x^\top)^\top\right].$$

We now compute the asymptotic covariance of the term $\operatorname{vec}(w x^\top)$ using standard identities:

$$\operatorname{vec}(w x^\top) = (x \otimes I_n) w,$$

where $x \in \mathbb{R}^n$, $w \sim \mathcal{N}(0, \sigma_w^2 I_n)$. We compute the expected outer product:

$$\mathbb{E}_{x \sim \nu_\infty, w}\left[\operatorname{vec}(w x^\top)\operatorname{vec}(w x^\top)^\top\right] = \mathbb{E}_{x,w}\left[(x \otimes I_n) w w^\top (x^\top \otimes I_n)\right].$$

Since $w \sim \mathcal{N}(0, \sigma_w^2 I_n)$, we have:

$$\mathbb{E}[w w^\top] = \sigma_w^2 I_n,$$

so the expression simplifies to:

$$\sigma_w^2 \cdot \mathbb{E}_{x \sim \nu_\infty}\left[(x \otimes I_n)(x^\top \otimes I_n)\right].$$

$$(x \otimes I_n)(x^\top \otimes I_n) = (x x^\top) \otimes I_n,$$

which gives:

$$\sigma_w^2 \cdot \mathbb{E}_{x \sim \nu_\infty}\left[(x x^\top) \otimes I_n\right] = \sigma_w^2 (P_\infty \otimes I_n),$$

where $P_\infty := \mathbb{E}[x x^\top]$ is the stationary covariance matrix of the process $\{x_t\}$. Thus,

[1] **Cramér–Wold Theorem.**[19] Let $X_n = (X_{n1}, \ldots, X_{nk})$ and $X = (X_1, \ldots, X_k)$ be random vectors in $\mathbb{R}^k$. Then $X_n \xrightarrow{d} X$ if and only if

$$\sum_{i=1}^{k} t_i X_{ni} \xrightarrow{d} \sum_{i=1}^{k} t_i X_i$$

for every $(t_1, \ldots, t_k) \in \mathbb{R}^k$. Thus, convergence of all one-dimensional linear projections characterizes joint convergence in distribution.

the asymptotic covariance of the term is:

$$\mathbb{E}_{x,w}\left[\mathrm{vec}(wx^\top)\mathrm{vec}(wx^\top)^\top\right] = \sigma_w^2(P_\infty \otimes I_n).$$

Since X is the matrix formed by the state trajectory $\{x_t\}$ and $\{x_t\}$ is ergodic, its time average converges almost surely to the expected value under the stationary distribution:

$$\frac{1}{T}\sum_{t=0}^{T-1} x_t x_t^\top = \frac{1}{T}XX^\top \xrightarrow{\text{a.s.}} \mathbb{E}[xx^\top] = P_\infty.$$

From earlier in the proof, we have:

$$\mathrm{vec}\left(T^{-1/2}W^\top X\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_w^2\left(P_\infty \otimes I_n\right)\right).$$

Next we analyze the asymptotics of $\left(\left(T^{-1}\left(XX^\top + \lambda I\right)\right)^{-1} \otimes I\right) \cdot \mathrm{vec}(T^{-1/2}WX^\top)$. Since

$$\left(T^{-1}\left(XX^\top + \lambda I\right)\right)^{-1} \otimes I = \left(T^{-1}XX^\top + T^{-1}\lambda I\right)^{-1} \otimes I$$

we see that the second term in the inverse tends to 0 as T is tends to ∞ while the first term

$$\frac{1}{T}XX^\top \xrightarrow{\text{a.e.}} P_\infty$$

by the ergodic theorem. Hence

$$\left(T^{-1}\left(XX^\top + \lambda I\right)\right)^{-1} \otimes I \xrightarrow{\text{a.e.}} P_\infty^{-1} \otimes I_n.$$

Then by **Proposition** 2.7, if $A_T \to A$ (a.s. or in probability), and $Y_T \xrightarrow{D} Y$, then:

$$A_T Y_T \xrightarrow{D} AY.$$

Thus, we conclude:

$$\left(\left(T^{-1}\left(XX^\top + \lambda I\right)\right)^{-1} \otimes I\right) \cdot \mathrm{vec}(T^{-1/2}WX^\top) \xrightarrow{D} \mathcal{N}\left(0, \sigma_w^2\left(P_\infty^{-1} \otimes I_n\right)\right).$$

□

Remark 4.2. Note that the goal of the lemma is to oet the system evolve as

$$x_{t+1} = L^\star x_t + w_t, \tag{15}$$

where $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$, and define the regularized least squares estimator:

$$\hat{L}(T) = \arg\min_{L} \left\{ \frac{1}{2} \sum_{t=0}^{T-1} \|x_{t+1} - Lx_t\|^2 + \frac{\lambda}{2} \|L\|_F^2 \right\}. \tag{16}$$

The lemma shows that

$$\hat{L}(T) \xrightarrow{\text{a.s.}} L^\star, \tag{17}$$

$$\sqrt{T} \cdot \text{vec}(\hat{L}(T) - L^\star) \xrightarrow{D} \mathcal{N}\left(0, \sigma_w^2 \left(P_\infty^{-1} \otimes I_n\right)\right), \tag{18}$$

with P_∞ is the stationary covariance matrix of x_t.

Lemma 4.3 ([6]). *Consider the system $x_{t+1} = L_\star x_t + w_t$ with $w_t \sim \mathcal{N}(0, \sigma_w^2 I_n)$ and let $\widehat{L}(T)$ be the regularized least-squares estimator:*

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

Then, for any $\delta \in (0, 1)$, there exists a constant C depending only on $L_\star, \lambda, n$, such that with probability at least $1 - \delta$:

$$\|\widehat{L}(T) - L_\star\| \leq C \sqrt{\frac{\log(1/\delta)}{T}}.$$

Furthermore, for any $p \geq 1$, there exists C_p depending on $L_\star, \lambda, n, p$ such that:

$$\mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] \leq \frac{C_p}{T^{p/2}}.$$

Proof. Recall in the notation of the proof of Lemma 4.1,

$$\widehat{L}(T) - L_\star = \lambda L_*(XX^\top + \lambda I_n)^{-1} + WX^\top \left(XX^\top + \lambda I_n\right)^{-1}$$

Suppose we are on an event where $XX^\top$ is invertible. Let $X = U\Sigma V^\top$ denote the compact SVD of X. Then

$$\|\widehat{L}(T) - L_\star\| \leq \frac{\lambda \|L_\star\|}{\lambda_{\min}(XX^\top + \lambda I_n)} + \|WX^\top (XX^\top + \lambda I_n)^{-1}\|.$$

$$\leq \frac{\lambda \|L_\star\|}{\lambda_{\min}(XX^\top + \lambda I_n)} + \|WX^\top (XX^\top)^{-1}\|.$$

The inequality holds since $(XX^\top + \lambda I_n)^{-2} \preceq (XX^\top)^{-2}$. Thus for $M = WX^\top$, conjugating both sides gives

$$M(XX^\top + \lambda I_n)^{-2} M^\top \preceq M(XX^\top)^{-2} M^\top,$$

so

$$\|M(XX^\top + \lambda I_n)^{-1}\| \le \sqrt{\lambda_{\max}(M(XX^\top + \lambda I_n)^{-2} M^\top)} \le \|M(XX^\top)^{-1}\|.$$

According to the theorem[2] stated in the footnote, for $T \ge C_{L_\star,n} \log(1/\delta)$, there exists an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$ such that on $\mathcal{E}$,

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

Hence on $\mathcal{E}$ we have

$$\|\widehat{L}(T) - L_\star\| \le C_{L_\star,n,\lambda} \sqrt{\tfrac{\log(1/\delta)}{T}}.$$

For the remainder of the proof, $O(\cdot)$ hides constants depending only on $L_\star, n, p, \lambda$. We bound the p-th moment as follows. Decompose:

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

[2] **Theorem** [20] Fix $\varepsilon, \delta \in (0,1)$, $T \in \mathbb{N}$, and $0 \prec \Gamma_{\mathrm{sb}} \preceq \overline{\Gamma}$. Let $(X_t, Y_t)_{t \ge 1} \in (\mathbb{R}^d \times \mathbb{R}^n)^T$ be a random sequence such that:

(a) $Y_t = A_\star X_t + \eta_t$, where $\eta_t \mid \mathcal{F}_t$ is σ^2-sub-Gaussian and mean zero,

(b) $X_1, \dots, X_T$ satisfies the $(k, \Gamma_{\mathrm{sb}}, p)$-small ball condition, and

(c) $\mathbb{P}\left[\sum_{t=1}^T X_t X_t^\top \not\preceq T\overline{\Gamma}\right] \le \delta$.

Then if

$$T \ge \frac{10k}{p^2}\left(\log\left(\frac{1}{\delta}\right) + 2d\log(10/p) + \log\det(\overline{\Gamma}\Gamma_{\mathrm{sb}}^{-1})\right),$$

we have

$$\mathbb{P}\left[\left\|\widehat{A}(T) - A_\star\right\|_{\mathrm{op}} > \frac{90\sigma}{p}\sqrt{\frac{n + d\log(10/p) + \log\det(\overline{\Gamma}\Gamma_{\mathrm{sb}}^{-1}) + \log(1/\delta)}{T\lambda_{\min}(\Gamma_{\mathrm{sb}})}}\right] \le 3\delta.$$

On $\mathcal{E}$, using $(a+b)^p \le 2^{p-1}(a^p + b^p)$, we have

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

On the other hand, always

$$\|\widehat{L}(T) - L_\star\|^p \le 2^{p-1}(\|L_\star\|^p + \|WX^\top\|^p/\lambda^p).$$

Hence

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

Now, bound $\mathbb{E}[\|WX^\top\|^{2p}]$. By Hölder's inequality,

$$\mathbb{E}[\|WX^\top\|^{2p}] = \mathbb{E}\left[\left\|\sum_{t=0}^{T-1} w_t x_t^\top\right\|^{2p}\right] \le T^{2p-1}\sum_{t=1}^{T}\mathbb{E}[\|w_t\|^{2p}\|x_t\|^{2p}].$$

Since $x_t \sim \mathcal{N}(0, P_\infty)$ in stationarity,

$$\mathbb{E}[\|WX^\top\|^{2p}] \le O(T^{2p}),$$

where P_∞ is the stationary covariance.

Thus

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

Finally, choose $\delta = O(1/T^{3p})$, which gives $O(1/T^{p/2})$.

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

□

Next we prove:

Lemma 4.4 ([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 $\hat{L}(T)$ denote the least-squares estimator:*

$$\hat{L}(T) = \arg\min_{L\in\mathbb{R}^{n\times n}}\left\{\frac{1}{2}\sum_{t=0}^{T-1}\|x_{t+1} - Lx_t\|_2^2 + \frac{\lambda}{2}\|L\|_F^2\right\}. \tag{19}$$

Fix a finite $p \ge 1$. Let $C_{L_\star,\lambda,n}$ and $C'_{L_\star,\lambda,n,p}$ denote constants that depend only on $L_\star, \lambda, n$ (respectively $L_\star, \lambda, n, p$) and not on T or δ. Fix $\delta \in (0,1)$. With probability

at least $1-\delta$*, as long as* $T \geq C_{L_\star,\lambda,n} \log(1/\delta)$*, we have*

$$\|\hat{L}(T) - L_\star\| \leq C'_{L_\star,\lambda,n} \sqrt{\frac{\log(1/\delta)}{T}}. \tag{20}$$

Furthermore, as long as $T \geq C_{L_\star,\lambda,n,p}$*, it holds that*

$$\mathbb{E}\Big[\|\hat{L}(T) - L_\star\|^p\Big] \leq C'_{L_\star,\lambda,n,p}\, T^{-p/2}. \tag{21}$$

Proof. The least-squares solution can be written in closed form:

$$\hat{L}(T) - L_\star = -\lambda L_\star (XX^\top + \lambda I_n)^{-1} + WX^\top (XX^\top + \lambda I_n)^{-1}.$$

Assume that the matrix $XX^\top$ is invertible, and let $X = U\Sigma V^\top$ denote its compact singular value decomposition (SVD). Then,

$$\|\hat{L}(T) - L_\star\| \leq \frac{\lambda \|L_\star\|}{\lambda_{\min}(XX^\top + \lambda I_n)} + \|WX^\top (XX^\top + \lambda I_n)^{-1}\|.$$

Using the fact that $\lambda_{\min}(XX^\top + \lambda I_n) \geq \lambda_{\min}(XX^\top)$, we have

$$\|WX^\top (XX^\top + \lambda I_n)^{-1}\| \leq \|WX^\top (XX^\top)^{-1}\|.$$

Letting $M = WX^\top$, we obtain

$$|M(XX^\top + \lambda I_n)^{-1}| \leq \sqrt{\lambda_{\max}(M(XX^\top + \lambda I_n)^{-2} M^\top)} \leq \|M(XX^\top)^{-1}\|.$$

Hence,

$$\|\hat{L}(T) - L_\star\| \leq \frac{\lambda \|L_\star\|}{\lambda_{\min}(XX^\top)} + \|WX^\top (XX^\top)^{-1}\|.$$

As stated in The Theorem in footnote 2, for $T \geq C_{L_\star,\lambda,n} \log(1/\delta)$, there exists an event $\mathcal{E}$ with probability $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ such that

$$\|WX^\top (XX^\top)^{-1}\| \leq C_{L_\star,\lambda,n} \sqrt{\frac{\log(1/\delta)}{T}}, \quad X^\top X \succeq C_{L_\star,\lambda,n}\, T\, I_n.$$

Substituting this bound yields

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

For the moment bound, we decompose

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

On $\mathcal{E}$, using $(a+b)^p \le 2^{p-1}(a^p + b^p)$ for $a, b \ge 0$, we have

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

We next bound $\mathbb{E}[\|WX^\top\|^p]$ using Hölder's inequality:

$$\mathbb{E}\Big[\|WX^\top\|^p\Big] = \mathbb{E}\left[\left\|\sum_{t=0}^{T-1} w_t x_t^\top\right\|^p\right] \le T^{2p-1}\sum_{t=0}^{T-1}\mathbb{E}[\|w_t\|^p\|x_t\|^p].$$

Since w_t and x_t are Gaussian and bounded by the stationary covariance P_∞, this term is $\mathcal{O}(T^{2p})$.

Combining these estimates and choosing $\delta = \mathcal{O}(T^{-3p})$ ensures that the remainder term is negligible. Hence, for T sufficiently large,

$$\mathbb{E}\Big[\|\hat{L}(T) - L_\star\|^p\Big] = \mathcal{O}(T^{-p/2}),$$

which completes the proof of Lemma 4.4. □

Theorem 4.5 ([6]). *[Upper bound on the asymptotic risk] Let K stabilize $(A^\star, B^\star)$. Define $L^\star = A^\star + B^\star K$ to be the closed-loop matrix and let $\rho(L^\star) \in (0,1)$ denote its spectral radius. Recall that $P^\star$ is the solution to the discrete-time Lyapunov equation:*

$$(L^\star)^\top P^\star L^\star - P^\star + Q + K^\top R K = 0.$$

Then, Algorithm 1 with thresholds (ζ, ψ) satisfying $\zeta \in (\rho(L^\star), 1)$, $\psi \in (\|L^\star\|, \infty)$, and any fixed regularization parameter $\lambda > 0$, has the asymptotic risk upper bound:

$$\begin{aligned}&\lim_{T\to\infty} T \cdot \mathbb{E}\left[\|\hat{P}_{plug}(T) - P^\star\|_F^2\right]\\ \le& 4 \cdot \mathrm{tr}\left[(I - L^{\star\top} \otimes_s L^{\star\top})^{-1}\left(L^{\star\top}(P^\star)^2 L^\star \otimes_s \sigma_w^2 (P_\infty)^{-1}\right)(I - L^{\star\top} \otimes_s L^{\star\top})^{-\top}\right],\end{aligned}$$

where $P_\infty = dlyap(L^\star, \sigma_w^2 I)$ 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.

This Theorem provides an upper bound on the asymptotic risk of the model-based plugin estimator for policy evaluation. The goal is to analyze the following

quantity:

$$\lim_{T\to\infty} T \cdot \mathbb{E}\left[\|\hat{P}_{\text{plug}}(T) - P^\star\|_F^2\right]$$

Proof of Theorem 4.5 . We divide the proof to the following steps:

Step 1. Analyze the Estimator for the Closed-Loop Matrix he plugin estimator uses the trajectory data to estimate the closed-loop matrix:

$$\hat{L}(T) = \hat{A}(T) + \hat{B}(T)K \approx L^\star = A^\star + B^\star K$$

Using results from Markov chain theory, specifically a Central Limit Theorem (CLT), one can show:

$$\sqrt{T} \cdot \text{vec}(\hat{L}(T) - L^\star) \xrightarrow{D} \mathcal{N}(0, \Sigma)$$

Step 2. Apply the Delta Method The value function matrix $P^\star$ satisfies the discrete Lyapunov equation:

$$P^\star = \text{dlyap}(L^\star, Q + K^\top RK)$$

We view the map $L \mapsto P(L) := \text{dlyap}(L, Q + K^\top RK)$ as a smooth function, and apply the *delta method* to derive the asymptotic distribution of:

$$\sqrt{T} \cdot \text{svec}(\hat{P}_{\text{plug}}(T) - P^\star)$$

This involves computing the Jacobian of the Lyapunov operator and applying it to the limiting distribution of $\hat{L}(T)$.

Step 3. Compute the Asymptotic Covariance Using the linear approximation from the delta method, we compute the asymptotic covariance matrix of the estimation error:

$$\text{Cov}\left[\sqrt{T} \cdot \text{svec}(\hat{P}_{\text{plug}}(T) - P^\star)\right]$$

The expected squared Frobenius norm is then given by the trace of this covariance:

$$\lim_{T\to\infty} T \cdot \mathbb{E}\left[\|\hat{P}_{\text{plug}}(T) - P^\star\|_F^2\right] = \text{tr}(\text{Cov}\left[\sqrt{T} \cdot \text{svec}(\hat{P}_{\text{plug}}(T) - P^\star)\right]$$

Step 4. Prove Uniform Integrability To rigorously justify taking the expectation limit,

the proof shows that the sequence

$$T \cdot \|\hat{P}_{\text{plug}}(T) - P^\star\|_F^2$$

is *uniformly integrable*. This ensures that convergence in distribution implies convergence in expectation.

Now we do it step by step.

Step 1 (Estimator for the Closed-Loop Matrix). As pointed out earlier, by Theorem 2.13 together with Lemma 4.1,

$$\sqrt{T}\,\mathrm{vec}\Big(\widehat{L}(T) - L_\star\Big) \overset{d}{\Rightarrow} \mathcal{N}(0, \Sigma_L),$$

for some covariance matrix Σ_L determined by the stationary covariance of the closed-loop process and the noise level. (The scaling is $\sqrt{T}$, as usual for CLTs.)

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 a perturbation X. By a straightforward application of the implicit function theorem, we have:

$$[DP(L_\star)](X) = \mathrm{dlyap}(L_\star, X^\top P_\star L_\star + L_\star^\top P_\star X),$$

where $P_\star$ is the solution to the discrete-time Lyapunov equation at $L_\star$. Let Γ denote the matrix such that

$$\Gamma \mathrm{vec}(S) = \mathrm{svec}(S)$$

for any symmetric matrix S. Let Π be the orthonormal matrix such that

$$\Pi \mathrm{vec}(X) = \mathrm{vec}(X^\top)$$

for all square matrices X. It is not hard to verify that:

$$\Pi^\top (A \otimes B)\Pi = B \otimes A.$$

With this notation, we proceed as follows:

$$\begin{aligned}\operatorname{svec}([DP(L_\star)](X)) &= (I - L_\star^\top \otimes_s L_\star^\top)^{-1}\operatorname{svec}(X^\top P_\star L_\star + L_\star^\top P_\star X) \\ &= (I - L_\star^\top \otimes_s L_\star^\top)^{-1}\Gamma\operatorname{vec}(X^\top P_\star L_\star + L_\star^\top P_\star X) \\ &= (I - L_\star^\top \otimes_s L_\star^\top)^{-1}\Gamma\left((I_n \otimes L_\star^\top)P_\star + (L_\star^\top \otimes I_n)P_\star\right)\operatorname{vec}(X).\end{aligned}$$

This vectorized form of the derivative is necessary for applying the Delta Method to the function $P(\cdot)$ evaluated at the estimator $\hat{L}(T)$. It allows us to compute the asymptotic distribution of:

$$\sqrt{T} \cdot \operatorname{svec}(P(\hat{L}(T)) - P_\star),$$

as a linear transformation of the asymptotic distribution of $\hat{L}(T)$ around $L_\star$.

Step 2 (Application of the Delta Method). Now, we apply the **delta method** to transfer the asymptotic normality of the closed-loop matrix estimate $\widehat{L}(T)$ to the value function estimate $\widehat{P}_{\text{plug}}(T)$. Recall from Lemma that we have

$$\sqrt{T} \cdot \operatorname{vec}(\widehat{L}(T) - L_\star) \xrightarrow{D} \mathcal{N}(0, \Sigma_L),$$

where $\widehat{L}(T) := \widehat{A}(T) + \widehat{B}(T)K$ is the estimated closed-loop matrix and $L_\star := A_\star + B_\star K$ is the true one.
Define the function

$$f(L) := \operatorname{dlyap}(L, Q + K^\top RK),$$

which maps the closed-loop matrix L to the solution P of the discrete-time Lyapunov equation:

$$P = L^\top PL + Q + K^\top RK.$$

Then, we can express the plugin estimator as

$$\widehat{P}_{\text{plug}}(T) = f(\widehat{L}(T)), \quad \text{and} \quad P_\star = f(L_\star).$$

The delta method asserts that if f is differentiable at $L_\star$, then

$$\sqrt{T} \cdot \operatorname{vec}\left(\widehat{P}_{\text{plug}}(T) - P_\star\right) \xrightarrow{D} \mathcal{N}\left(0, J_f(L_\star)\Sigma_L J_f(L_\star)^\top\right),$$

where $J_f(L_\star)$ denotes the Jacobian of the function f evaluated at $L_\star$.
We treat the map $f(L) = \text{dlyap}(L, Q+K^\top RK)$ as a smooth operator. The derivative of f at $L_\star$ is obtained using standard results from matrix calculus and control theory. Letting $P_\star := f(L_\star)$, the Jacobian is:

$$J_f(L_\star) = (I - L_\star \otimes_s L_\star)^{-1} \left(L_\star^\top \otimes_s P_\star + P_\star \otimes_s L_\star \right),$$

where $\otimes_s$ denotes the *symmetric Kronecker product* and $\text{svec}(\cdot)$ vectorizes symmetric matrices.
Applying the delta method yields:

$$\sqrt{T} \cdot \text{svec}\left(\widehat{P}_{\text{plug}}(T) - P_\star \right) \xrightarrow{D} \mathcal{N}(0, \Sigma_P),$$

where the asymptotic covariance matrix Σ_P is:

$$\Sigma_P = (I - L_\star \otimes_s L_\star)^{-1} \left(L_\star^\top \otimes_s P_\star + P_\star \otimes_s L_\star \right) \cdot \Sigma_L \cdot \left[\left(L_\star^\top \otimes_s P_\star + P_\star \otimes_s L_\star \right)^\top (I - L_\star \otimes_s L_\star)^{-1} \right].$$

This completes Step 2 and prepares us to compute the asymptotic risk bound in Step 3.

Step 3 (Asymptotic Covariance Computation). We are studying the convergence in distribution of:

$$T \cdot \text{svec}(P(\hat{L}(T)) - P^\star),$$

where:

- $\hat{L}(T)$ is the estimated (regularized) closed-loop matrix from least squares,
- $P(\cdot)$ maps a stable matrix $L \in \mathbb{R}^{n \times n}$ to the unique solution P of the discrete-time Lyapunov equation:

$$(A + BK)^\top P (A + BK) - P + Q + K^\top RK = 0.$$

Since $P(L)$ is smooth in L, we can apply the *Delta Method* to the CLT for $\hat{L}(T)$ (established in Lemma 4.1) to derive the asymptotic distribution of $P(\hat{L}(T))$. Using the Delta Method, we approximate:

$$T \cdot \text{svec}(P(\hat{L}(T)) - P^\star) \approx [DP(L^\star)] \left(T \cdot \text{vec}(\hat{L}(T) - L^\star) \right),$$

and from Lemma 4.1:

$$T \cdot \mathrm{vec}(\hat{L}(T) - L^\star) \xrightarrow{D} \mathcal{N}\left(0, \sigma_w^2 (P_\infty^{-1} \otimes I_n)\right).$$

Hence, we need to compute and apply the linear map $[DP(L^\star)](\cdot)$.
The Fréchet derivative of the Lyapunov solution map is:

$$[DP(L^\star)](X) = \mathrm{dlyap}\left(L^\star, X^\top P^\star L^\star + L^{\star\top} P^\star X\right),$$

and its vectorized form is:

$$\mathrm{svec}([DP(L^\star)](X)) = \left(I - L^{\star\top} \otimes_s L^{\star\top}\right)^{-1} \Gamma \left((I_n \otimes L^{\star\top})P^\star + (L^{\star\top} \otimes I_n)P^\star\right) \mathrm{vec}(X),$$

as established earlier in the proof. Here: - $\otimes_s$ denotes the symmetric Kronecker product, - Γ maps full vectorization to symmetric vectorization (i.e., vec to svec). The final asymptotic covariance becomes:

$$\sigma_w^2 \left(I - L^{\star\top} \otimes_s L^{\star\top}\right)^{-1} V \left(I - L^{\star\top} \otimes_s L^{\star\top}\right)^{-\top},$$

where the intermediate matrix V is:

$$V := \Gamma \left[\left((L^{\star\top} P^\star \otimes I_n)\Pi + (I_n \otimes L^{\star\top} P^\star)\right)(P_\infty^{-1} \otimes I_n)\left((L^{\star\top} P^\star \otimes I_n)\Pi + (I_n \otimes L^{\star\top} P^\star)\right)^\top\right] \Gamma^\top.$$

We aim to bound the asymptotic covariance matrix V, which appears in the central limit theorem derived via the Delta method. Recall that:

$$\sqrt{T} \cdot \mathrm{svec}(P(\hat{L}(T)) - P_\star) \xrightarrow{D} \mathcal{N}\left(0, \sigma_w^2 (I - L_\star^\top \otimes L_\star^\top)^{-1} V (I - L_\star^\top \otimes L_\star^\top)^{-\top}\right)$$

where V is defined as:

$$V := \Gamma \left[\left((L_\star^\top P_\star \otimes I_n)\Pi + (I_n \otimes L_\star^\top P_\star)\right)(P_\infty^{-1} \otimes I_n)\left((L_\star^\top P_\star \otimes I_n)\Pi + (I_n \otimes L_\star^\top P_\star)\right)^\top\right] \Gamma^\top$$

We apply the inequality from Zhang (2005, Chapter 3, page 94)[21], which states that for any matrices X, Y and positive definite matrices F, G:

$$(X+Y)(F+G)^{-1}(X+Y)^\top \preceq XF^{-1}X^\top + YG^{-1}Y^\top$$

In our case:

$$X = (L_\star^\top P_\star \otimes I_n)\Pi$$
$$Y = (I_n \otimes L_\star^\top P_\star)$$
$$F = G = P_\infty \otimes I_n$$

Applying this inequality gives:

$$V \preceq 2\Gamma \left[(L_\star^\top P_\star \otimes I_n)\Pi(P_\infty^{-1} \otimes I_n)\Pi^\top (P_\star L_\star \otimes I_n) + (I_n \otimes L_\star^\top P_\star)(P_\infty^{-1} \otimes I_n)(I_n \otimes P_\star L_\star)\right] \Gamma^\top$$

Using properties of Kronecker products and the fact that $\Pi\Pi^\top = I$, we simplify each term:

$$\Gamma \left[(L_\star^\top P_\star \otimes I_n)\Pi(P_\infty^{-1} \otimes I_n)\Pi^\top (P_\star L_\star \otimes I_n)\right] = \Gamma \left[L_\star^\top P_\star^2 L_\star \otimes P_\infty^{-1}\right]$$
$$\Gamma \left[(I_n \otimes L_\star^\top P_\star)(P_\infty^{-1} \otimes I_n)(I_n \otimes P_\star L_\star)\right] = \Gamma \left[L_\star^\top P_\star^2 L_\star \otimes P_\infty^{-1}\right]$$

Adding these together and applying linearity of Γ, We use the inequality:

$$(A+B)(A+B)^\top \preceq 2AA^\top + 2BB^\top,$$

we obtain:

$$V \preceq 4\left(L_\star^\top P_\star^2 L_\star \otimes P_\infty^{-1}\right)$$

We want to evaluate the asymptotic risk:

$$\lim_{T\to\infty} T \cdot \mathbb{E}\left[\|P(\hat{L}(T)) - P^\star\|_F^2\right].$$

Let:

$$Z_T := T \cdot \mathrm{svec}(P(\hat{L}(T)) - P^\star).$$

Then:

$$\|Z_T\|_2^2 = T \cdot \|P(\hat{L}(T)) - P^\star\|_F^2.$$

Hence, the asymptotic risk becomes:

$$\lim_{T\to\infty} \mathbb{E}\left[\|Z_T\|_2^2\right].$$

Step 4 (To justify interchanging the limit and expectation) We require that $\|Z_T\|_2^2$ is *uniformly integrable*. If uniform integrability holds, then convergence in distribution implies convergence in expectation. Since:

$$Z_T \xrightarrow{D} \mathcal{N}\left(0, \sigma_w^2 (I - L^{\star\top} \otimes L^{\star\top})^{-1} V (I - L^{\star\top} \otimes L^{\star\top})^{-\top}\right),$$

we conclude:

$$\lim_{T\to\infty} T \cdot \mathbb{E}\left[\|P(\hat{L}(T)) - P^\star\|_F^2\right] = \mathbb{E}\left[\|Z_\infty\|_2^2\right] = \text{tr(Covariance)}.$$

Thus, the final asymptotic risk bound becomes:

$$\begin{aligned}&\lim_{T\to\infty} T \cdot \mathbb{E}\left[\|P(\hat{L}(T)) - P^\star\|_F^2\right] \\ \leq& 4 \cdot \text{tr}\left((I - L^{\star\top} \otimes L^{\star\top})^{-1} \left(L^{\star\top} P^{\star 2} L^\star \otimes \sigma_w^2 P_\infty^{-1}\right) (I - L^{\star\top} \otimes L^{\star\top})^{-\top}\right).\end{aligned}$$

We now show that the sequence $\{\|Z_T\|_F^2\}$ is uniformly integrable.

According to Lemma 4.3, we define $\delta_T := O(1/T^{p/2})$, and let T be sufficiently large so that the high-probability event $\mathcal{E}_{\text{bdd}}$ satisfies

$$\mathbb{P}(\mathcal{E}_{\text{bdd}}) \geq 1 - \delta_T,$$

and on this event, we have the operator norm bound

$$\|\widehat{L}(T) - L_\star\| \leq O\left(\sqrt{\frac{\log(1/\delta_T)}{T}}\right).$$

Additionally, we choose T large enough so that this deviation is controlled as

$$\|\widehat{L}(T) - L_\star\| \leq \min\left(\frac{\gamma - \rho_\star}{C_\star}, \psi - \|L_\star\|\right),$$

ensuring that $\widehat{L}(T)$ remains within a stable neighborhood of $L_\star$. We define the following compact set of admissible matrices:

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

We now introduce a convex combination:

$$\widetilde{L}(\alpha) := \alpha \widehat{L}(T) + (1 - \alpha) L_\star, \quad \text{for any } \alpha \in (0, 1).$$

By convexity of $\mathcal{G}$, it follows that $\widetilde{L}(\alpha) \in \mathcal{G}$ for all $\alpha \in (0,1)$, provided that T is large enough.

Since the Lyapunov operator $P(L)$ is Fréchet differentiable on the stable set, we apply the mean value theorem for operator-valued functions to obtain:

$$\|P(\widehat{L}(T)) - P_\star\| = \|[DP(\widetilde{L}(\alpha))](\widehat{L}(T) - L_\star)\|,$$

for some $\alpha \in (0,1)$.

Because the map $L \mapsto P(L)$ is continuously differentiable and $\widetilde{L}(\alpha) \in \mathcal{G}$, we have:

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

This establishes a Lipschitz-type bound on the value function estimation error in terms of the matrix estimation error. The constant S depends only on the compact set $\mathcal{G}$ and not on T, ensuring that the bound holds uniformly with high probability.

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

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

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

$$\begin{aligned}
\mathbb{E}\Big[\|\widehat{P}_{\text{plug}}(T) - P_\star\|^p\Big] &= \mathbb{E}\Big[\|\widehat{P}_{\text{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\text{bdd}}}\Big] + \mathbb{E}\Big[\|\widehat{P}_{\text{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\text{bdd}}^c}\Big] \\
&\le \mathbb{E}\Big[\|\widehat{P}_{\text{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\text{bdd}}}\Big] + \mathbb{E}\Big[\|\widehat{P}_{\text{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\text{bdd}}^c \cap \mathcal{E}_{\text{Alg}}}\Big] \\
&\quad + \mathbb{E}\Big[\|\widehat{P}_{\text{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\text{bdd}}^c \cap \mathcal{E}_{\text{Alg}}^c}\Big].
\end{aligned}$$

We assume T is sufficiently large.

Case 1: On $\mathcal{E}_{\text{bdd}}$. Since $\mathcal{E}_{\text{bdd}} \subseteq \mathcal{E}_{\text{Alg}}$, by Lemma 4.3:

$$\begin{aligned}
&\mathbb{E}\Big[\|\widehat{P}_{\text{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\text{bdd}}}\Big] \\
=&\mathbb{E}\Big[\|\widehat{P}_{\text{plug}}(T) - P_\star\|^p \mathbf{1}_{\mathcal{E}_{\text{bdd}} \cap \mathcal{E}_{\text{Alg}}}\Big] \le S^p\, \mathbb{E}[\|\widehat{L}(T) - L_\star\|^p] \le O(1/T^{p/2}).
\end{aligned}$$

Case 2: On $\mathcal{E}_{\mathbf{bdd}}^c \cap \mathcal{E}_{\mathbf{Alg}}$. Here, using compactness of $\mathcal{G}_{\mathrm{Alg}}$:

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

Case 3: On $\mathcal{E}_{\mathbf{bdd}}^c \cap \mathcal{E}_{\mathbf{Alg}}^c$. In this event,

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

Putting it together. Combining all cases:

$$\mathbb{E}\Big[\|\widehat{P}_{\mathrm{plug}}(T) - P_\star\|^p\Big] \leq O(1/T^{p/2}).$$

Recall that $Z_T = \mathrm{svec}(\widehat{P}_{\mathrm{plug}}(T) - P_\star)$. For any finite $\gamma > 0$ and $T \geq \Omega(1)$:

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

Thus $\sup_T \mathbb{E}[\|Z_T\|^{2+\gamma}] < \infty$, showing $\{Z_T\}$ is uniformly integrable. This completes the proof of Theorem 4.5. □

4.3.2 Asymptotic analysis of model-free algorithm for policy evaluation (LSTD)

We turn to asymptotic analysis of model-free method.

We consider a model-free method for policy evaluation, which does *not* attempt to estimate the system dynamics $A^\star, B^\star$. Instead, we use the Least-Squares Temporal Difference (LSTD) learning algorithm. **Task:** Evaluate a fixed policy $\pi(x) = Kx$ in the Linear Quadratic Regulator (LQR) setting. In LQR, the value function for policy K takes the form:

$$V_K(x) = x^\top P^\star x,$$

where $P^\star$ satisfies the Lyapunov equation under policy K. The goal is to estimate $P^\star$, or an approximation $\hat{P}$, using only data.

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

We are given a fixed linear state-feedback policy:

$$u_t = Kx_t,$$

and a rollout horizon of T time steps, determining the amount of data to be collected.

Execute the policy $u_t = \pi(x_t) = Kx_t$ on the unknown system starting from $x_0 = 0$. The system evolves according to:

$$x_{t+1} = A^\star x_t + B^\star u_t + w_t,$$

where $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$ is system noise.

Collect the sequence of states $\{x_t\}_{t=0}^{T}$ resulting from this rollout. This state trajectory is the only information available to the algorithm — there is no access to the true system matrices $A^\star$ or $B^\star$.

We know that $\lambda^* = \sigma_w^2 P^\star$ by Proposition 3.13. However the matrix $P^\star$ is unknown in practice. So we approximate $\lambda^\star$ empirically as:

$$\hat{\lambda} = \frac{1}{T} \sum_{t=0}^{T-1} c_t.$$

However, in theoretical analyses (such as the LSTD algorithm), one may assume access to the true value $\lambda^\star$ to isolate and analyze algorithmic performance without estimation noise.

The value function has the form:

$$V_K(x) = x^\top P^\star x.$$

Define:

$$w^\star = \mathrm{svec}(P^\star), \quad \phi(x) = \mathrm{svec}(xx^\top),$$

then the value function can be written as:

$$V_K(x) = w^{\star\top} \phi(x).$$

So $w^\star$ is the “true" weight vector corresponding to $P^\star$.

What $\widehat{w}_{\mathbf{lstd}}(T)$ means: Since $P^\star$ is unknown, we estimate $w^\star$ from data.

After collecting a trajectory of length T, the LSTD regression computes:

$$\widehat{w}_{\text{lstd}}(T) = A^{-1}b,$$

where

$$A = \sum_{t=0}^{T-1} \phi(x_t) \left(\phi(x_t) - \phi(x_{t+1})\right)^\top,$$

$$b = \sum_{t=0}^{T-1} (c_t - \lambda_t)\phi(x_t).$$

This $\widehat{w}_{\text{lstd}}(T)$ is the empirical least-squares estimate of the true vector $w^\star$. In the algorithm, we first estimated

$$\widehat{w}_{\text{lstd}}(T),$$

which is a vector.

But the true value function is quadratic:

$$V_K(x) = x^\top P^\star x,$$

where $P^\star$ is a matrix.

So, to interpret the result, we need to "reshape" the vector $\widehat{w}_{\text{lstd}}(T)$ back into a symmetric matrix.

Because quadratic forms can be written as a linear model:

$$x^\top P x = \langle P, xx^\top \rangle = w^\top \phi(x),$$

with:

$$w = \text{svec}(P) \quad \text{(vector form of matrix } P\text{)},$$

$$\phi(x) = \text{svec}(xx^\top).$$

This representation makes it possible to use linear regression (specifically, LSTD) to estimate w.

So

$$\widehat{P}_{\text{lstd}}(T) = \text{smat}\left(\widehat{w}_{\text{lstd}}(T)\right)$$

It means:

> "Take the estimated vector of coefficients and reshape it into the matrix form that defines the quadratic value function."

In simple words:

- $\widehat{w}_{\text{lstd}}(T)$ is just the vector of parameters learned by LSTD.
- $\text{smat}(\cdot)$ converts that vector back into a matrix.
- $\widehat{P}_{\text{lstd}}(T)$ is the final estimate of the value function matrix you actually care about.

The system evolves as

$$x_{t+1} = L^\star x_t + w_t, \quad w_t \sim \mathcal{N}(0, \sigma_w^2 I).$$

Since w_t is Gaussian with mean zero, the state x_t also has mean zero in steady state. x_t is random (because of the noise w_t).

To analyze algorithms like LSTD, we need to understand the distribution of the states in steady state.

$$\text{Cov}(x_t) = \mathbb{E}[x_t x_t^\top] - \mathbb{E}[x_t]\mathbb{E}[x_t]^\top.$$

Since the process is mean zero, $\mathbb{E}[x_t] = 0$, the covariance simplifies to:

$$\text{Cov}(x_t) = \Sigma_t = \mathbb{E}[x_t x_t^\top].$$

That's exactly what we call P_∞.

Now, we compute Σ_{t+1}:

$$\begin{aligned}
\Sigma_{t+1} &= \mathbb{E}[x_{t+1} x_{t+1}^\top] \\
&= \mathbb{E}[(L^\star x_t + w_t)(L^\star x_t + w_t)^\top] \\
&= \mathbb{E}[L^\star x_t x_t^\top L^{\star\top}] + \mathbb{E}[L^\star x_t w_t^\top] + \mathbb{E}[w_t x_t^\top L^{\star\top}] + \mathbb{E}[w_t w_t^\top].
\end{aligned}$$

The cross terms vanish since w_t is independent of x_t and $\mathbb{E}[w_t] = 0$. Thus,

$$\Sigma_{t+1} = L^\star \Sigma_t L^{\star\top} + \sigma_w^2 I_n.$$

The covariance satisfies, for all t,

$$\Sigma_{t+1} = L^\star \Sigma_t L^{\star\top} + \sigma_w^2 I_n.$$

Define the linear–affine map:

$$T(\Sigma) := L^\star \Sigma L^{\star\top} + \sigma_w^2 I_n.$$

This map is continuous in Σ.

Suppose $\Sigma_t \to P_\infty$. Then, by continuity of T,

$$P_\infty = \lim_{t\to\infty} \Sigma_{t+1} = \lim_{t\to\infty} T(\Sigma_t) = T\left(\lim_{t\to\infty} \Sigma_t\right) = T(P_\infty).$$

Therefore,

$$P_\infty = L^\star P_\infty L^{\star\top} + \sigma_w^2 I_n,$$

which is the discrete Lyapunov equation. So P_∞ is a Lyapunov solution.

Suppose P and $\tilde{P}$ both satisfy:

$$P = L^\star P L^{\star\top} + \sigma_w^2 I, \quad \tilde{P} = L^\star \tilde{P} L^{\star\top} + \sigma_w^2 I.$$

Subtracting the equations gives:

$$\Delta := P - \tilde{P} \quad \Rightarrow \quad \Delta = L^\star \Delta L^{\star\top}.$$

Vectorizing both sides:

$$\mathrm{vec}(\Delta) = (L^\star \otimes L^\star)\mathrm{vec}(\Delta).$$

If $\rho(L^\star) < 1$, then:

$$\rho(L^\star \otimes L^\star) = \rho(L^\star)^2 < 1.$$

Thus, the only fixed point is $\mathrm{vec}(\Delta) = 0$, hence $\Delta = 0$, so $P = \tilde{P}$.

(Equivalently, one can write:

$$(I - L^\star \otimes L^\star)\mathrm{vec}(P) = \mathrm{vec}(\sigma_w^2 I),$$

and the matrix $I - L^\star \otimes L^\star$ is invertible because $\rho(L^\star \otimes L^\star) < 1$.)

Iterating the recursion from any $\Sigma_0 \succeq 0$ gives:

$$\Sigma_t = L^{\star t} \Sigma_0 (L^{\star\top})^t + \sum_{k=0}^{t-1} L^{\star k} (\sigma_w^2 I)(L^{\star\top})^k \succeq 0.$$

Taking $t \to \infty$ and using $\rho(L^\star) < 1$ so that the first term vanishes, yields:

$$P_\infty = \sum_{k=0}^{\infty} L^{\star k}(\sigma_w^2 I)(L^{\star\top})^k \succeq 0,$$

So P_∞ is positive semidefinite (PSD).

Lemma 4.6 ([6]). *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 $\widehat{\Theta}(N)$ 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 $\left\{\zeta_t^{(i)} := (x_t^{(i)}, u_t^{(i)})\right\}_{t=0}^{T}{}_{i=1}^{N}$:*

$$\widehat{\Theta}(N) = \arg\min_{(A,B)\in\mathbb{R}^{n\times(n+d)}} \frac{1}{2N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\left\|x_{t+1}^{(i)} - Ax_t^{(i)} - Bu_t^{(i)}\right\|_2^2 + \frac{\lambda}{2}\left\|[A \quad B]\right\|_F^2.$$

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

$$A_\star P_\infty A_\star^\top - P_\infty + \sigma_u^2 B_\star B_\star^\top + \sigma_w^2 I_n = 0.$$

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

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

Proof. See [6] Appendix D.2 □

Lemma 4.6 states two important properties of the least-squares estimator $\widehat{\Theta}(N)$:

1. **Consistency:**
$$\widehat{\Theta}(N) \xrightarrow{\text{a.s.}} \Theta_\star = (A_\star, B_\star).$$
That is, as the number of independent trajectories N grows, the least-squares estimator converges almost surely to the true system parameters.

2. **Asymptotic Distribution:**
The scaled estimation error converges in distribution to a multivariate Gaussian:
$$\sqrt{N}\cdot\mathrm{vec}(\widehat{\Theta}(N) - \Theta_\star) \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma_u^2}{T}\begin{bmatrix} P_\infty^{-1} & 0 \\ 0 & \frac{1}{\sigma_u^2} I_d \end{bmatrix} \otimes I_n + o\left(\frac{1}{T}\right)\right).$$

This expression gives the asymptotic covariance of the estimator.

Lemma 4.7 ([6]). *Let $x_{t+1} = A_\star x_t + B_\star u_t + w_t$ be a linear system driven by $u_t = Kx_t$ 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*

$$\begin{aligned}
A_\infty &:= \mathbb{E}_{\substack{x\sim\nu_\infty,\\ x'\sim\mathcal{P}(\cdot|x,\pi(x))}} \left[\phi(x)\left(\phi(x) - \phi(x')\right)^\top\right], \\
B_\infty &:= \mathbb{E}_{\substack{x\sim\nu_\infty,\\ x'\sim\mathcal{P}(\cdot|x,\pi(x))}} \left[\left(\left(\phi(x') - \psi(x)\right)^\top w_\star\right)^2 \phi(x)\phi(x)^\top\right], \\
\psi(x) &:= \mathbb{E}_{x'\sim\mathcal{P}(\cdot|x,\pi(x))} \left[\phi(x')\right], \\
w_\star &:= \operatorname{svec}(P_\star).
\end{aligned}$$

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

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

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 $\widehat{w}_{\mathrm{lstd}}(T) \xrightarrow{a.s.} w_\star$ and furthermore:

$$\sqrt{T}\left(\widehat{w}_{\mathrm{lstd}}(T) - w_\star\right) \overset{d}{\Rightarrow} \mathcal{N}\left(0, A_\infty^{-1} B_\infty A_\infty^{-\top}\right).$$

This lemma considers the asymptotic distribution of the least squares temporal learning for LQR The Lemma statements:

1. **Consistency:**

   $$\widehat{w}_{\mathrm{lstd}}(T) \xrightarrow{\text{a.s.}} w^\star.$$

   That is, the LSTD estimator converges almost surely to the true value-function parameter.

2. **Asymptotic Normality:**

   $$T \cdot \left(\widehat{w}_{\mathrm{lstd}}(T) - w^\star\right) \xRightarrow{T\to\infty} \mathcal{N}\left(0,\ A_\infty^{-1} B_\infty A_\infty^{-\top}\right).$$

This lemma provides a central limit theorem for the LSTD estimator in the LQR setting:

- As more data is collected ($T \to \infty$), the LSTD estimate of the value-function matrix $P^\star$ (in its vectorized form) becomes increasingly accurate.

- The convergence rate is $\mathcal{O}(1/T)$.

- The asymptotic distribution is Gaussian, with covariance determined by matrices A_∞ and B_∞, which depend on the data distribution and the LQR parameters.

Proof. Let the per-stage cost be

$$c_t = x_t^\top (Q + K^\top R K) x_t.$$

From the Bellman equation for the average-cost setting, we have

$$c_t - \lambda_\star = (\phi(x_t) - \psi(x_t))^\top w_\star,$$

where $\phi(x) = \mathrm{svec}(xx^\top)$, $\psi(x) = \mathbb{E}[\phi(x') \mid x]$, and $w_\star = \mathrm{svec}(P_\star)$.

The LSTD estimator is defined by

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

Subtracting $w_\star$, we write

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

Dividing by T, this can be expressed as

$$\hat{w}_{\mathrm{lstd}}(T) - w_\star = \left(\frac{1}{T} \sum_{t=0}^{T-1} \phi(x_t)(\phi(x_t) - \phi(x_{t+1}))^\top \right)^{-1} \left(\frac{1}{T} \sum_{t=0}^{T-1} \phi(x_t)(\phi(x_{t+1}) - \psi(x_t))^\top w_\star \right).$$

We now analyze the two terms. So the analysis splits into two empirical averages:

$$\frac{1}{T}\sum_{t=0}^{T-1}\phi(x_t)\left(\phi(x_t)-\phi(x_{t+1})\right)^\top,$$

which converges to A_∞.

$$\frac{1}{T}\sum_{t=0}^{T-1}\phi(x_t)\left(\phi(x_{t+1})-\psi(x_t)\right)^\top w_\star,$$

which (after CLT scaling) converges in distribution to a Gaussian with covariance B_∞.

Consider the process $z_t = (x_t, w_t)$. Since x_{t+1} depends only on z_t, the sequence $\{z_t\}$ forms a Markov chain.

Moreover, the stationary distribution of this chain is

$$\nu_\infty \times \mathcal{N}(0, \sigma_w^2 I_n).$$

By the ergodic theorem, we have almost sure convergence

$$\frac{1}{T}\sum_{t=0}^{T-1}\phi(x_t)(\phi(x_t)-\phi(x_{t+1}))^\top \xrightarrow{\text{a.s.}} A_\infty,$$

so that its inverse converges to A_∞^{-1} by the continuous mapping theorem.

Next, Theorem CTL for Markov chain 2.13 together with the Cramér–Wold theorem as stated in footnote 1 implies

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

Combining the two pieces, we conclude

$$\sqrt{T}(\hat{w}_{\text{lstd}}(T) - w_\star) \overset{D}{\Longrightarrow} \mathcal{N}(0,\ A_\infty^{-1} B_\infty A_\infty^{-T}),$$

as required. □

Theorem 4.8 ([6])**.** *Let K stabilize $(A_\star, B_\star)$. Define $L_\star$ to be the closed-loop matrix $L_\star = A_\star + B_\star K$. Recall that $P_\star$ is the solution to the discrete-time Lyapunov equation that parameterizes the value function $V^K(x)$. We have that Algorithm 2 with the cost*

estimates λ_t set to the true cost $\lambda_\star := \sigma_w^2 tr(P_\star)$ satisfies the asymptotic risk lower bound:

$$\begin{aligned}
&\liminf_{T\to\infty} T\cdot\mathbb{E}\Big[\big\|\hat{P}_{\mathrm{lstd}}(T)-P_\star\big\|_F^2\Big]\\
&\geq\ 4R_{\mathrm{plug}}8\sigma_w^2\, tr\Big((P_\infty\otimes_s L_\star^\top P_\star^2 L_\star)\,(I-L_\star^\top\otimes_s L_\star^\top)^{-1}\,(P_\infty^{-1}\otimes_s P_\infty^{-1})\,(I-L_\star\otimes_s L_\star)^{-\top}\Big)\\
&\quad+\ 16\sigma_w^2\, tr\Big((I-L_\star^\top\otimes_s L_\star^\top)^{-1}\,(L_\star^\top P_\star^2 L_\star\otimes_s P_\infty^{-1})\,(I-L_\star\otimes_s L_\star)^{-\top}\Big).
\end{aligned}$$

Here, $R_{\mathrm{plug}} := \lim_{T\to\infty} T\cdot\mathbb{E}\Big[\big\|\hat{P}_{\mathrm{plug}}(T)-P_\star\big\|_F^2\Big]$ is the asymptotic risk of the plugin estimator, and $P_\infty = \mathrm{dlyap}(L_\star^\top, \sigma_w^2 I_n)$ 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.

Proof. Let $\hat{w}_{\mathrm{lstd}}(T) := \mathrm{svec}\Big(\hat{P}_{\mathrm{lstd}}(T)\Big)$ and $w_\star := \mathrm{svec}(P_\star)$. By Lemma 4.7 we have the central limit theorem

$$\sqrt{T}\Big(\hat{w}_{\mathrm{lstd}}(T)-w_\star\Big)\ \xRightarrow{D}\ \mathcal{N}\Big(0,\ \Sigma_\infty\Big),\qquad \Sigma_\infty := A_\infty^{-1}B_\infty A_\infty^{-\top},$$

where

$$A_\infty :=\ E_{\nu_\infty}\Big[\phi(x)\Big(\phi(x)-\phi(x')\Big)^\top\Big],\qquad B_\infty :=\ E_{\nu_\infty}\Big[\Big((\phi(x))^\top w_\star-\psi(x)\Big)^2\,\phi(x)\phi(x)^\top\Big],$$

with $\phi(x) = \mathrm{svec}(xx^\top)$ and $\psi(x) = \mathbb{E}[\phi(x') \mid x]$. Since the svec map is an isometry for symmetric matrices, we have $\|\hat{P}_{\mathrm{lstd}}(T)-P_\star\|_F^2 = \|\hat{w}_{\mathrm{lstd}}(T)-w_\star\|_2^2$.

Apply Lemma 4.1 with $X_T := \sqrt{T}\Big(\hat{w}_{\mathrm{lstd}}(T)-w_\star\Big)$ and $f(u) = \|u\|_2^2$ and use the fact that f is nonnegative and continuous, Lemma 2.8 yields

$$\liminf_{T\to\infty} T\,\mathbb{E}\Big[\|\hat{P}_{\mathrm{lstd}}(T)-P_\star\|_F^2\Big]\ \geq\ \mathbb{E}\Big[\|Z\|_2^2\Big]\ =\ \mathrm{tr}(\Sigma_\infty),\quad Z\sim\mathcal{N}(0,\Sigma_\infty).$$

Hence

$$\liminf_{T\to\infty} T\,\mathbb{E}\Big[\|\hat{P}_{\mathrm{lstd}}(T)-P_\star\|_F^2\Big]\ \geq\ \mathrm{tr}\Big(A_\infty^{-1}B_\infty A_\infty^{-\top}\Big). \tag{22}$$

It remains to rewrite the right-hand side explicitly. Using the identities

$$\mathrm{svec}(LXL^\top) = (L\otimes_s L)\,\mathrm{svec}(X),\qquad E_{\nu_\infty}\Big[\phi(x)\phi(x)^\top\Big] = 2\,(P_\infty\otimes_s P_\infty),$$

and the conditional expectation $\psi(x) =\ E[\phi(x') \mid x] = (L_\star\otimes_s L_\star)\phi(x) + \sigma_w^2\,\mathrm{svec}(I_n)$,

together with Gaussian fourth-moment formulas, one obtains after routine algebra that

$$\operatorname{tr}\Big(A_\infty^{-1} B_\infty A_\infty^{-\top}\Big) \;\geq\; 8\sigma_w^2 \operatorname{tr}\Big((P_\infty \otimes_s L_\star^\top P_\star^2 L_\star)\,(I - L_\star^\top \otimes_s L_\star^\top)^{-1}(P_\infty^{-1} \otimes_s P_\infty^{-1})\,(I - L_\star \otimes_s L_\star)^{-\top}\Big)$$

$$+\, 16\sigma_w^2 \operatorname{tr}\Big((I - L_\star \otimes_s L_\star)^{-1}(L_\star^\top P_\star^2 L_\star \otimes_s P_\infty^{-1})\,(I - L_\star \otimes_s L_\star)^{-\top}\Big),$$

which is exactly the lower bound displayed in the proof of Theorem 4.8. Finally, compare this lower bound with the plugin estimator's asymptotic risk upper bound provided by Theorem 4.5 The comparison shows that the plugin estimator attains an asymptotic risk no larger than the LSTD lower bound above, proving the claim of Theorem 4.8. □

The proof is done under an idealized setting:

In Algorithm 2 (LSTD), we assume access to the true average cost:

$$\lambda^\star = \sigma_w^2 \operatorname{tr}(P^\star).$$

In practice, $\lambda^\star$ is not known, it must be estimated from data. [6]

So, the analysis is somewhat optimistic: it assumes access to a quantity that is unavailable in real-world settings. [6]

- Using the true cost $\lambda^\star$ instead of an estimate only helps LSTD.
- If $\lambda^\star$ had to be estimated from data, the risk of LSTD would actually be higher.

Therefore, the inequality in Theorem 4.8,

$$\text{plugin risk} \leq \text{LSTD risk},$$

holds even more strongly in practice.

The comparison in Theorem 4.5 is already generous toward LSTD.

Even with the unrealistic advantage of knowing $\lambda^\star$, LSTD still shows higher asymptotic risk.

In real-world use (when $\lambda^\star$ is unknown), the performance gap between model-based and model-free approaches would be even wider.

We proved that model-free LSTD is less sample-efficient than the plugin method *even if we give it extra information* (the true cost $\lambda^\star$) that it wouldn't have access

to in practice.[6]
Our goal is to compare two methods for policy evaluation in LQR: the *plugin* (model-based) estimator and the *LSTD* (model-free) estimator, in terms of their *risk*, i.e. the expected estimation error.
For the plugin (model-based) estimator, the error arises primarily from estimating the closed-loop matrix $L_\star = A_\star + B_\star K$. In contrast, for the LSTD (model-free) estimator, the error comes from solving the regression equations that involve the quadratic features $\phi(x) = \mathrm{svec}(xx^\top)$.
Analyzing the risk for *all possible* LQR systems is infeasible, as the algebra quickly becomes intractable for arbitrary closed-loop matrices.
A standard technique in statistics and learning theory [22, 23, 24, 25] is, therefore, to restrict attention to a carefully chosen *structured family* of problem instances. Such families are simple enough to allow explicit calculation of risks, yet rich enough to capture the essential difficulty of the estimation task. Following this approach, Tu and Recht [26] introduce the family $\mathcal{F}(\rho, d, K)$ of structured LQR systems, which we also adopt in our analysis.

In LQR, with system matrices $(A_\star, B_\star)$ and a fixed feedback policy $u_t = Kx_t$, the state evolves as

$$x_{t+1} = (A_\star + B_\star K)x_t + w_t.$$

The matrix

$$L_\star = A_\star + B_\star K$$

is the *closed-loop matrix*, which determines both stability ($\rho(L_\star) < 1$) and the evolution of the states. Thus, once K is fixed, performance analysis is reduced to studying $L_\star$.

Instead of allowing $L^\star$ to be arbitrary, we define a restricted structured family:

$$\begin{aligned}&\mathcal{F}(\rho, d, K)\\ &:= \left\{(A^\star, B^\star) : L^\star = A^\star + B^\star K = \tau P_E + \gamma I_n,\ (\tau, \gamma) \in (0, 1),\ \tau + \gamma \le \rho,\ \dim(E) \le d\right\}.\end{aligned}$$

Form of $L^\star$: It must be a combination of two parts:

$$L^\star = \tau P_E + \gamma I_n.$$

- I_n: the $n \times n$ identity matrix(cost matrix normalized).

$$Q + K^\top RK = I_n$$

- P_E: a projection matrix onto some subspace $E \subseteq \mathbb{R}^n$. A projection matrix satisfies $P_E^2 = P_E$, and its rank is $\dim(E)$.
- $\tau, \gamma \in (0, 1)$: scalars that determine the balance between the projection and identity parts.
- $\tau + \gamma \leq \rho$: ensures the spectral radius $\rho(L^\star) \leq \rho$, which implies stability.
- Smaller ρ means more stable systems.
- $\dim(E) \leq d$: the perturbation subspace has dimension at most d.

As a results of Theorems 4.5 and 4.8 and 2: The asymptotic risk bounds for the model-based (plugin) and model-free (LSTD) estimators are as follows:

Plugin Estimator

For the plugin estimator, the asymptotic risk satisfies:

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

LSTD Estimator

For the model-free LSTD estimator, the asymptotic risk satisfies:

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

This implies that, as the sample size $T \to \infty$, the asymptotic risk of LSTD decays at a rate of $\mathcal{O}\left(\frac{n^3}{T}\right)$, up to constants.

Both bounds show that the risk decreases at the rate of $\mathcal{O}(1/T)$ as the number of samples T increases. However, the dependence on the state dimension n differs significantly:

- **Plugin estimator:** risk scales as $\mathcal{O}(n^2)$

- **LSTD estimator:** risk scales as $\Omega(n^3)$

This result demonstrates that the model-free LSTD estimator incurs an additional asymptotic risk scaling factor of n relative to the model-based plug-in estimator. Consequently, the plug-in approach exhibits superior sample efficiency, particularly in high-dimension.

Table 1: Asymptotic risk scaling of plugin (model-based) vs. LSTD (model-free) estimators for the family $\mathcal{F}(\rho, d, K)$.

Estimator	**Asymptotic risk (per T samples)**	**Scaling in n**
Plugin (model-based)	$O\left(\frac{\rho^2 n^2}{(1-\rho^2)^3\, T}\right)$	$O\left(\frac{n^2}{T}\right)$
LSTD (model-free)	$\Omega\left(\frac{\rho^2 n^3}{(1-\rho^2)^3\, T}\right)$	$\Omega\left(\frac{n^3}{T}\right)$

4.3.3 A minimax lower bound on the risk

A concluding result for policy evaluation establishes a minimax lower bound on the estimation risk for any algorithm, taken over the family $\mathcal{F}(\rho, d, K)$.

Theorem 4.9 ([6]). *Fix $\alpha, \rho \in (0,1)$ and suppose that K satisfies $Q + K^\top R K = I_n$. Suppose that n is greater than an absolute constant and $T \gtrsim n(1-\rho^2)/\rho^2$. We have that:*

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

where the infimum is taken over all estimators $\widehat{P}$ taking input $\{x_t\}_{t=0}^T$.[6]

Proof. To establish minimax lower bounds, we use Fano's method. The idea is to construct a set of alternative problem instances that are:

(i) Statistically indistinguishable (i.e., their distributions have small Kullback-Leibler divergence), and

(ii) Well-separated in the quantity of interest (i.e., the target matrices P_i and P_j satisfy $\|P_i - P_j\|_F$ is large).

Any estimator must incur significant error on at least one of the instances.

We choose N different d-dimensional subspaces $E_1, \ldots, E_N \subset \mathbb{R}^n$ with associated orthogonal projections P_{E_i}. A classical packing result on the Grassmannian (Pajor)[27] guarantees the existence of such a family with pairwise separation:

$$N \geq \exp(n(n-d))$$

$$\|P_{E_i} - P_{E_j}\|_F \gtrsim d.$$

Define matrices:

$$A_i := \tau P_{E_i} + \gamma I_n, \quad \text{with } \tau, \gamma \in (0,1), \text{ and } \tau + \gamma = \rho < 1.$$

Then each system defined by:

$$x_{t+1} = A_i x_t + w_t, \quad w_t \sim \mathcal{N}(0, \sigma^2 I)$$

is stable. Define the Lyapunov solution:

$$P_i = \text{dlyap}(A_i, I_n).$$

These P_i matrices are well-separated due to the distinct subspaces E_i.

Use the chain rule for KL divergence of Markov processes:

$$\text{KL}(\mathbb{P}_i, \mathbb{P}_j) = \sum_{t=1}^{T} \mathbb{E}_i \left[\text{KL}\left(\mathcal{N}(A_i x_{t-1}, \sigma^2 I) \| \mathcal{N}(A_j x_{t-1}, \sigma^2 I) \right) \right].$$

The Gaussian KL reduces to:

$$\frac{1}{2\sigma^2} \mathbb{E}_i \left[\|(A_i - A_j) x_{t-1}\|_2^2 \right].$$

Summing gives:

$$\text{KL}(\mathbb{P}_i, \mathbb{P}_j) \lesssim \frac{T}{1-\rho^2} \|A_i - A_j\|_F^2.$$

With $A_i - A_j = \tau(P_{E_i} - P_{E_j})$, we obtain:

$$\text{KL}(\mathbb{P}_i, \mathbb{P}_j) \lesssim \frac{\tau^2 T}{1-\rho^2}.$$

Recall that:

$$P = \sum_{k \geq 0} (A^k)^\top A^k.$$

Using the recursion and only keeping first-order terms in τ, we find:

$$P_i - P_j \approx \sum_{k \geq 0} \sum_{\ell=0}^{k-1} \gamma^{2k-2-\ell} \tau (P_{E_i} - P_{E_j}).$$

Evaluating the geometric sum gives:

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

With average KL divergence $\leq \frac{1}{2} \log N$ and separation Δ in P, Fano's inequality implies:

$$\inf_{\widehat{P}} \sup_i \mathbb{E} \|\widehat{P} - P_i\|_F^2 \gtrsim \Delta^2.$$

Substituting $\Delta \asymp \frac{\gamma\tau}{(1-\gamma^2)^2} d$ and optimizing over parameters subject to $\gamma + \tau = \rho$ and KL constraints (this is where $T \gtrsim \frac{n(1-\rho^2)}{\rho^2}$ comes in), and setting $d \asymp n$, we conclude:

$$\inf_{\widehat{P}} \sup \mathbb{E} \|\widehat{P} - P_\star\|_F^2 \gtrsim \frac{\rho^2 n^2}{(1-\rho^2)^3 T}.$$

□

5 Discussions

In this study, the Linear Quadratic Regulator (LQR) problem was adopted as an analytical framework to investigate and compare model-based and model-free control methodologies. The LQR formulation offers a mathematically tractable and well-established setting in which the optimal control law can be derived in closed form. As noted in [15], it represents one of the few problems where dynamic programming yields an exact analytical solution, allowing the separation of estimation and control aspects. This makes the LQR problem particularly suitable for evaluating learning-based control algorithms, since it enables a systematic and rigorous analysis of their convergence properties, sample efficiency, and robustness to disturbances. Furthermore, the LQR framework provides a neutral platform to compare model-based and model-free systems: model-based methods exploit knowledge or estimation of the system dynamics to compute optimal feedback gains, whereas model-free methods learn control policies directly from observed trajectories without relying on explicit models. Thus, the LQR setting isolates the fundamental trade-off between *model accuracy* and *data efficiency*, revealing the inherent strengths and limitations of both approaches under a unified theoretical framework.

5.1 Mathematical tractability

The Linear Quadratic Regulator (LQR) framework, provides a mathematically tractable and analytically solvable formulation for optimal control of linear dynamical systems. The system dynamics are given by

$$x_{t+1} = Ax_t + Bu_t, \quad x_0 \in \mathbb{R}^n, \tag{23}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ denote the system and input matrices, respectively. The control objective is to minimize the quadratic performance index

$$J = \sum_{t=0}^{\infty} \left(x_t^\top Q x_t + u_t^\top R u_t \right), \tag{24}$$

where $Q \succeq 0$ and $R \succ 0$ are symmetric weighting matrices that penalize deviations in the state and control effort. Dynamic programming yields that the value function is quadratic in the state, $V(x) = x^\top P x$, where the matrix P satisfies the discrete

algebraic Riccati equation

$$P = Q + A^\top PA - A^\top PB(B^\top PB + R)^{-1}B^\top PA. \tag{25}$$

The optimal control law is then expressed in state-feedback form as

$$u_t = -Kx_t, \quad K = (B^\top PB + R)^{-1}B^\top PA. \tag{26}$$

Under standard assumptions such as the stabilizability of the pair (A, B) and the detectability of $(A, Q^{1/2})$, the feedback matrix K ensures asymptotic stability of the closed-loop system $x_{t+1} = (A - BK)x_t$. Because the Linear Quadratic Regulator (LQR) problem is *linear in its dynamics* and *quadratic in its cost*, every component of the optimal control law can be expressed in closed form. This implies that there is no model approximation error, since the optimal policy

$$u_t = -Kx_t \tag{27}$$

can be computed exactly. As a result, any observed performance difference between algorithms cannot originate from nonlinearities or approximation artifacts; it must arise purely from the way the system model is obtained (in model-based methods) or from how the control policy is learned directly from data (in model-free methods). This property makes the LQR problem an ideal and transparent framework for comparing model-based and model-free reinforcement learning algorithms, as it isolates the effects of estimation and learning without the confounding influence of model or numerical inaccuracies.[11]

5.2 Stability guarantees

The LQR law ensures a stable closed-loop system, which is essential for analyzing convergence in both model-based and model-free algorithms. As established in Theorem 3.11 (Section 3.4), the infinite-horizon linear–quadratic regulator (LQR) problem

$$\min_{u_t} \sum_{t=0}^{\infty} \left(x_t^\top Q x_t + u_t^\top R u_t \right) \quad \text{subject to} \quad x_{t+1} = Ax_t + Bu_t,$$

with $Q \succeq 0$ and $R \succ 0$, admits a unique positive semidefinite solution P to the discrete algebraic Riccati equation (DARE)

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

and the optimal control policy

$$u_t = -Lx_t, \qquad L = (R + B^\top PB)^{-1}B^\top PA,$$

renders the closed-loop system asymptotically stable.

Due to its analytical structure, guaranteed optimality, and explicit stability properties, the LQR serves as an ideal benchmark for comparing control strategies. It allows for a clear and rigorous examination of the fundamental differences between model-based and model-free reinforcement learning approaches within a well-understood theoretical setting.[11]

5.3 Comparison between Model-Based and Model-Free Methods

Both the model-based and model-free approaches rely on observed trajectories under a **known LQR controller**. However, they differ fundamentally in how the observed data is used to estimate the system performance. Table 2 summarizes the key differences between the two. In both methods, we observe trajectories:

$$x_0,\ x_1,\ x_2,\ \ldots,\ x_T.$$

Since the controller K is known, we also know:

$$u_t = -Kx_t.$$

So, raw data is:

Time	State x_t	Next state x_{t+1}	Control u_t	Stage cost c_t
t	x_t	x_{t+1}	u_t	$c_t = x_t^\top Q x_t + u_t^\top R u_t$

How each method uses this data

Table 2: Comparison between model-based and model-free (TD, LSTD) policy evaluation methods.

	Model-Based	**Model-Free (TD, LSTD)**
What data is used?	(x_t, x_{t+1})	(x_t, x_{t+1}, c_t)
What is estimated?	The **closed-loop matrix** $L = A - BK$	The **value function** directly
How is it used?	Fit a model: $x_{t+1} \approx Lx_t \rightarrow$ solve Lyapunov: $P = Q + L^\top PL$	Use Bellman consistency: $V(x_t) = c_t + V(x_{t+1})$, then adjust V using all samples
End result	Plug-in P gives cost	Direct estimate for P gives cost

Minimax Optimality of the Plugin Estimator:[6] Theorem 4.9 shows that the plugin (model-based) estimator achieves the minimax optimal rate for policy evaluation in the Linear Quadratic Regulator (LQR) setting. In other words:

- The lower bound establishes that no estimator can achieve an error smaller than the order
$$\frac{\rho^2 n^2}{(1-\rho^2)^3 T}$$
when evaluated over the family $\mathcal{F}(\rho, d, K)$.
- The upper bound from Theorem 4.5 demonstrates that the plugin algorithm actually attains this rate.

Together, these results imply that the plugin method is **asymptotically minimax optimal**:

- Its risk decays at the best possible rate as the sample size T increases.
- Its dependence on the system dimension n and the spectral radius ρ is the best achievable up to constant factors.

The asymptotic risk bounds presented in Theorems 4.5 and 4.8 establish a clear efficiency gap between the two approaches. Specifically, Theorem 4.5 shows that the model-based (plug-in) estimator achieves an asymptotic risk of order $\mathcal{O}\left(\frac{n^2}{T}\right)$, while Theorem 4.8 proves that the model-free (LSTD) estimator exhibits a lower bound of

$\Omega\left(\frac{n^3}{T}\right)$. Hence, the model-free method requires approximately an order of magnitude more data (in the state dimension n) to attain comparable accuracy. Theorem 4.9 further demonstrates that the plug-in estimator attains the minimax optimal rate of convergence, $\frac{\rho^2 n^2}{(1-\rho^2)^3 T}$, implying that no estimator can asymptotically achieve a smaller risk over the considered family of LQR systems $\mathcal{F}(\rho, d, K)$. Together, these results confirm that the superiority of the model-based approach is not merely algorithmic but is instead a fundamental statistical property of the problem.

AI Statement

AI assistance was employed during the preparation of this thesis exclusively for editorial support, such as grammar refinement and restructuring of sentences. All scientific content, derivations, and results are produced solely by the author and the supervisor.

References

[1] The Editors of Encyclopaedia Britannica. Expected value, 2025. Accessed on September 1, 2025.

[2] Galin L Jones. Markov chain monte carlo methods. *Handbook of computational statistics*, pages 179–195, 2004.

[3] Marco Taboga. Variance. https://www.statlect.com/fundamentals-of-probability/variance, 2017. Accessed: September 23, 2025.

[4] Gregory Gundersen. Ergodic markov chains, 2019. Accessed: 2025-07-08.

[5] J. A. Wooldridge. Chapter 2: Lecture notes for stat 581–582–583. Technical report, Department of Statistics, University of Washington, 2006. Revised 12/1/2006.

[6] Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Proceedings of the 32nd Annual Conference on Learning Theory (COLT)*, volume 99, pages 1–48. PMLR, 2019.

[7] Don Lemons. *An Introduction to Stochastic Processes in Physics*. Johns Hopkins University Press, 2003.

[8] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Springer, 1998.

[9] Jiawei Han. Kullback–leibler divergence. https://hanj.cs.illinois.edu/cs412/bk3/KL-divergence.pdf, 2020. Lecture notes for CS412: Introduction to Data Mining, University of Illinois at Urbana–Champaign.

[10] Steven W. Nydick. A different(ial) way: Matrix derivatives again. Presentation, University of Minnesota, May 2012.

[11] Mikael Johansson. *Linear quadratic and model predictive control*. h, h edition, 2018.

[12] Hax Bradley and Magnanti. Dynamic programming, chapter 11. Technical report, MIT, 1997.

[13] C. I. Byrnes, A. Lindquist, and Y. Zhou. On the nonlinear dynamics of kalman filtering. *SIAM Journal on Control and Optimization*, 32:744–789, 1994.

[14] T. Kailath. *Linear System Theory*. Prentice Hall, 2nd edition, 1980. with contributions by Wilson J. Rugh.

[15] Dimitri P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019.

[16] Yishao Zhou. Private communication, 2025. Personal correspondence.

[17] Kembey Gbarayor Jr. Linear dynamical systems: A machine learning framework for financial time series analysis. Senior honors thesis, Brown University, Department of Computer Science, Providence, RI, USA, 2020.

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

[19] Juan Antonio Cuesta-Albertos, Ricardo Fraiman, and Thomas Ransford. A sharp form of the cramér–wold theorem. *Journal of Theoretical Probability*, 20(2):201–209, 2007.

[20] Max Simchowitz, Horia Mania, Stephen Tu, Michael Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference on Learning Theory (COLT)*, pages 439–473, 2018.

[21] Fuzhen Zhang. *The Schur Complement and Its Applications*, volume 4 of *Numerical Methods and Algorithms*. Springer, 2005.

[22] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[23] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

[24] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[25] Po-Ling Loh. On lower bounds for statistical learning theory. *Entropy*, 19(11):617, 2017.

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

[27] Alain Pajor. Metric entropy of the grassmann manifold. In *Convex Geometric Analysis*, volume 34, pages 181–188. American Mathematical Society, 1998.