LQ Optimal Tracking with Unbounded Cost for Unknown Dynamics: An Adaptive Dynamic Programming Approach

Sebastian Bernhard and Jürgen Adamy

***Abstract*— In case of unknown system dynamics, we consider linear quadratic optimal tracking on infinite horizons with generally unbounded cost. For the first time, we deal with this problem in the framework of adaptive dynamic programming. So far, existing methods require bounded costs which essentially limits the applicability and achievable performance. Thus, we develop a new algorithm that yields a strongly overtaking optimal control which is an adequate solution. After collecting measurement data in an exploration phase, the algorithm implicitly solves the necessary and sufficient algebraic equations in [3], but without knowledge of the dynamics. Then, implementing the control results in an optimal transition to an optimal stationary trajectory. A simulation example of almost exact tracking for an over-actuated system demonstrates a highly efficient saving of input-energy in contrast to state-of-the-art approaches.**

I. Introduction

In the recent past, methods for online learning of optimal controls for linear systems with unknown dynamics have become a matter of great interest [15], [16]. Concentrating on continuous-time systems, first, the essential linear quadratic regulator (LQR) problem was solved in [12], [21] by a technique which is known as *adaptive dynamic programming* [15]. From there on, several approaches achieved exact tracking [8], [9] and infinite-horizon linear quadratic optimal tracking (LQT) [17], [18], [20]. So far, the latter only handle LQT problems (LQTP) for the very limited case of *bounded* cost. Before we present a literature overview and discuss the advantages of an LQTP that allows for *unbounded* cost, let us state our **main problem** and our key contributions.

In this paper, we regard a disturbed LTI system

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u} + \boldsymbol{E}_\mathrm{d}\overline{\boldsymbol{x}}, \tag{1a}$$

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x} + \boldsymbol{D}_\mathrm{d}\overline{\boldsymbol{x}} \tag{1b}$$

with states $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}(0) = \boldsymbol{x}_0$, inputs $\boldsymbol{u} \in \mathbb{R}^m$, outputs $\boldsymbol{y} \in \mathbb{R}^p$ and state/output disturbances $\boldsymbol{E}_\mathrm{d}\overline{\boldsymbol{x}}$, $\boldsymbol{D}_\mathrm{d}\overline{\boldsymbol{x}}$ given by a *not stable* exosystem with states $\overline{\boldsymbol{x}} \in \mathbb{R}^{\overline{n}}$, $\overline{\boldsymbol{x}}(0) = \overline{\boldsymbol{x}}_0$:

$$\dot{\overline{\boldsymbol{x}}} = \overline{\boldsymbol{A}}\overline{\boldsymbol{x}}, \tag{2a}$$

$$\overline{\boldsymbol{y}} = \overline{\boldsymbol{C}}\overline{\boldsymbol{x}} \tag{2b}$$

which also generates the reference output $\overline{\boldsymbol{y}} \in \mathbb{R}^p$. We pose

Linear Quadratic Tracking Problem 1: For systems (1) and (2), find the optimal control $\boldsymbol{u}^*(\cdot)$ concerning the cost

$$J_{t_\mathrm{E},T}(\boldsymbol{u}(\cdot))\big|_{\boldsymbol{x}(t_\mathrm{E}),\,\overline{\boldsymbol{x}}(t_\mathrm{E})} = \tfrac{1}{2}\int_{t_\mathrm{E}}^{T} (\boldsymbol{y}-\overline{\boldsymbol{y}})^\mathsf{T}\boldsymbol{Q}(\boldsymbol{y}-\overline{\boldsymbol{y}}) + \boldsymbol{u}^\mathsf{T}\boldsymbol{R}\boldsymbol{u}\ \mathrm{d}t, \tag{3}$$

The authors are with the Institute of Automatic Control and Mechatronics, Control Methods and Robotics Lab; Technische Universität Darmstadt, Landgraf-Georg Str. 4, 64283 Darmstadt, Germany, {bernhard, adamy}@rmr.tu-darmstadt.de

with $t_\mathrm{E} \geq 0$, constant $\boldsymbol{Q} \succeq \boldsymbol{0}$ and $\boldsymbol{R} \succ \boldsymbol{0}$, on an infinite horizon $T \to \infty$ when the *data* of the complete dynamics:

$$\boldsymbol{A},\ \boldsymbol{B},\ \boldsymbol{E}_\mathrm{d} \text{ and } \overline{\boldsymbol{A}} \tag{4}$$

is **unknown**.

Even for known data, this is in effect a non-trivial problem for two reasons. First, the cost is unbounded in general, i.e. $\lim_{T\to\infty} J_{t_\mathrm{E},T}(\boldsymbol{u}(\cdot))|_{\boldsymbol{x}(t_\mathrm{E}),\,\overline{\boldsymbol{x}}(t_\mathrm{E})} = +\infty$ for any $\boldsymbol{u}(\cdot)$ (see [1]), which requires an adequate concept of optimality [6]. Second, there are no suitable sufficient conditions to prove optimality for the given LQTP, cf. [7].

To our advantage, the two mentioned problems have been rigorously solved [2], [3]. In contrast to [2], a time-invariant control of simple structure $\widehat{\boldsymbol{u}} = -\widehat{\boldsymbol{K}}(\widehat{\boldsymbol{x}} - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}) + \boldsymbol{F}^*\overline{\boldsymbol{x}}$ is derived in [3] based on necessary optimality criteria for $T \to \infty$ [10]. Under certain conditions, it is *strongly overtaking optimal* with respect to cost (3), see [2] for a definition. As a result, the transition to the stationary state: $\lim_{t\to\infty} \widehat{\boldsymbol{x}}(t) - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}(t) = \boldsymbol{0}$ is optimal with LQR feedback $\widehat{\boldsymbol{K}}$. Moreover, it is proven in [3] that $\boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}$ is an *optimal stationary state*. Thus, any other $\boldsymbol{u}(\cdot) \not\equiv \widehat{\boldsymbol{u}}(\cdot)$ leading to $\boldsymbol{x}(t) - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}(t) \not\to \boldsymbol{0}$ as $t \to \infty$ yields: $\lim_{T\to\infty} J_{t_\mathrm{E},T}(\boldsymbol{u}(\cdot)) - J_{t_\mathrm{E},T}(\widehat{\boldsymbol{u}}(\cdot)) = +\infty$ and, therefore, it requires infinitely more cost. Following the concept of *agreeable plans* [6], $\widehat{\boldsymbol{u}}(\cdot)$ is also shown to be a favorable approximation of the optimal control $\boldsymbol{u}_T^*(\cdot)$ on finite horizons $[t_\mathrm{E}, T]$ under relaxed assumptions.

Considering the transition, it is beneficial to be able to specify a different cost than (3). It allows us to handle differing or even conflicting requirements of stationary behavior and transition, cf. [3]. Consequently, we have to implement

$$\boldsymbol{u}^* = -\boldsymbol{K}^*(\boldsymbol{x}^* - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}) + \boldsymbol{F}^*\overline{\boldsymbol{x}} = -\boldsymbol{K}^*\widetilde{\boldsymbol{x}}^* + \boldsymbol{F}^*\overline{\boldsymbol{x}} \tag{5}$$

where control $\widetilde{\boldsymbol{u}}^* = -\boldsymbol{K}^*\widetilde{\boldsymbol{x}}^*$ solves a separately formulated LQR problem (LQRP) whereas $\lim_{t\to\infty} \boldsymbol{x}^*(t) - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}(t) = \boldsymbol{0}$ is still guaranteed. However, to obtain $\boldsymbol{K}^*$, $\boldsymbol{\Pi}_x^*$ and $\boldsymbol{F}^*$, it is required to solve necessary and sufficient algebraic equations in [3] which depend on the unknown data (4). In this light, we formulate our **main contributions**:

For the first time, we develop an adaptive dynamic programming (ADP) approach to determine control (5) in three steps: First, measurement data is collected during an exploration phase $[0, t_\mathrm{E}]$. Second, $\boldsymbol{u}^*(\cdot)$ is obtained by calculating

C1) a feedback matrix $\boldsymbol{K}_k$, $k \in \mathbb{N}_0$, iteratively such that $\lim_{k\to\infty} \|\boldsymbol{K}_k - \boldsymbol{K}^*\|_2 = 0$ as well as $\boldsymbol{\Pi}_x^*$ and $\boldsymbol{F}^*$ exactly such that $\lim_{t\to\infty} \boldsymbol{x}^*(t) - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}(t) = \boldsymbol{0}$ *without* requiring the unknown data (4).

C2) Third, implementing $\boldsymbol{u}^*(\cdot)$ on $[t_\mathrm{E}, \infty)$ guarantees an optimal transition to stationary state $\boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}$, which is

optimal on infinite horizons and approximates optimal solutions on finite horizons, under certain assumptions.

In addition,

C3) $\boldsymbol{K}^*$ may solve a separated LQR problem in order to consider a transition cost different from (3) and to obtain an optimal transition $\|\widetilde{\boldsymbol{x}}^*(t)\|_2 \leq M\mathrm{e}^{-\alpha t}$, $M \in \mathbb{R}^{>0}$ with a specified degree of stability $\alpha > 0$.

C4) The approach is well suited in case of *under-actuation* $m < p$, where $\lim_{t\to\infty} \boldsymbol{y}(t) - \overline{\boldsymbol{y}}(t) = \boldsymbol{0}$ is usually infeasible, and *over-actuation* $\operatorname{rank}(\boldsymbol{B}) > p$ without any need of solving additional parametric optimization problems in contrast to, e.g., [8], [9].

In the sequel, we give a short **literature review** and highlight some advantages of our approach.

In the domain of interest, the first results considered the online identification of the LQR feedback $\boldsymbol{K}^*$ of a plant (1a), $\overline{\boldsymbol{x}}(\cdot) \equiv \boldsymbol{0}$, with unknown system matrix $\boldsymbol{A}$ [21] and additionally unknown input matrix $\boldsymbol{B}$ [12]. The proposed methods exploit the iterative computation of the *Riccati matrix* given by [13] and involve two separated steps. First, collecting measurement data in a phase where the system is excited by exploration noise and stabilized by an initial feedback $\boldsymbol{K}_0$. Second, computing iteratively a $\boldsymbol{K}_k$ by a procedure equivalent to [13] but independent of unknown data (4) such that convergence occurs: $\lim_{k\to\infty} \boldsymbol{K}_k = \boldsymbol{K}^*$. This technique known as adaptive dynamic programming (ADP) or, at times, (measurement-) *data-driven adaptive optimal control* is used to derive the first part of C1). Moreover, we extend the results of [12] to account for C3), i.e. guaranteeing a specified degree of stability. This is useful for preserving the degree of stability given by $\boldsymbol{K}_0$.

Regarding the problem of tracking references, all results [8], [9] and [17], [18], [20] consider the same setup as ours, i.e. LTI-systems (1a) with references (and disturbances) generated by an exosystem (2a). We separate two groups.

Based on output regulation theory, [8] and [9] introduce exact tracking for unknown dynamics without steady-state error: $\lim_{t\to\infty} \boldsymbol{y}(t) - \overline{\boldsymbol{y}}(t) = \boldsymbol{0}$. This is achieved by an LQR-transition, obtained via ADP methods as above, to a stationary state that is the solution of the necessary and sufficient *regulator equations* [19]. Before solving these, the parts depending on unknown matrices have to be identified. In [8], [9], this requires to solve repeatedly a set of equations for matrix variables. In addition, if (1a) is over-actuated, i.e. $\operatorname{rank}(\boldsymbol{B}) > \operatorname{rank}(\boldsymbol{C})$, the sought solution is not unique. Then, an optimization problem by [14] is considered in [8], [9] which has to be solved online for $n\overline{n} + m\overline{n}$ variables. In any case, we require to solve a single set of equations only once and, thus, our implementation is less complicated. Furthermore, we show by simulation that our approach is more efficient in using additional actuators.

The second group [17], [18], [20] considers LQT problems on infinite horizons which are essentially limited to bounded cost. All of them merge (1a), (2a) into an augmented system with state $\boldsymbol{x}_{\text{aug}}^{\mathsf{T}} = \begin{bmatrix}\boldsymbol{x}^{\mathsf{T}} & \overline{\boldsymbol{x}}^{\mathsf{T}}\end{bmatrix}$. Then, they solve LQTP 1 as an LQRP of the augmented system. Indeed, this is a common approach for the so-called servo-problem [1, Sec. 4.2]. In [18], LQTP 1 is solved by the methods in [12] in order to obtain the desired LQR $\boldsymbol{u}^*_{\text{aug}}(\cdot)$. This requires, however, that a $\boldsymbol{P}_{\text{aug}} \in \mathbb{R}^{(n+\overline{n})\times(n+\overline{n})}$ exists such that $J_{t_{\text{E}},\infty}(\boldsymbol{u}^*_{\text{aug}}(\cdot)) = \frac{1}{2}\boldsymbol{x}_{\text{aug}}(t_{\text{E}})^{\mathsf{T}}\boldsymbol{P}_{\text{aug}}\boldsymbol{x}_{\text{aug}}(t_{\text{E}})$, i.e. the cost is *bounded*. Following [1], this is only true when the augmented system is stabilizable. Necessarily, [18] assumes that $\overline{\boldsymbol{A}}$ is *Hurwitz*, i.e. $\overline{\boldsymbol{y}}(\cdot)$ must be exponentially decreasing which is a crucial restriction. The same assumption is obviously needed to apply the results of [20]. There, an actor-critic structure (cf. [16]) uses a critic to improve the actor. The critic precisely mimics the same quadratic form $\frac{1}{2}\boldsymbol{x}_{\text{aug}}^{\mathsf{T}}\boldsymbol{P}_{\text{aug}}\boldsymbol{x}_{\text{aug}}$ via a Q-function (cf. [20, Lemma 2]) which necessarily needs to exist. In contrast we assume, without loss of generality, that all eigenvalues of $\overline{\boldsymbol{A}}$ lie in the closed right half-plane.

To handle the problem above, [17] uses a discount-factor $\gamma > 0$ by multiplying the integrand of (3) by $\mathrm{e}^{-2\gamma t}$. Similar to [1, Sec. 3.5], the discounted LQRP for the augmented system $\boldsymbol{x}_{\text{aug}}$ can be reformulated as a standard LQRP. Then again, the solution in [17] can be obtained by using methods in [12], [21] if $\gamma > 0$ is chosen such that $\overline{\boldsymbol{A}} - \gamma\boldsymbol{I}$ is *Hurwitz*. That is a milder assumption as above. Using a discount factor, however, leads to several shortcomings which are absent in our approach. First, the discount factor lowers the degree of stability of the closed-loop dynamics in an opposite manner as in [1, Sec. 3.5]. One can observer that it may cause unstable closed loops when the open loop is unstable. Second, it is rather not-transparent how the choice of $\gamma > 0$ affects the stationary tracking performance.

Altogether, [17], [18], [20] have to use the same cost for transition and stationary behavior. Then for "large" weights $\boldsymbol{Q}$, input constraints may be violated during transition. This is critical, if "large" weights are desired for *almost exact tracking* (AET), i.e. implementing arbitrarily small stationary tracking errors on any finite horizon of interest. This motivates C4) that allows us to implement AET more efficiently for over-actuated systems than [8], [9], cf. Section V.

Summarizing, there is strong motivation to solve LQTP 1 for generally unbounded cost in case of unknown dynamics, i.e. without restrictive assumptions [18] or modifications of (3) as [17]. To achieve C1-4), we proceed as follows. After some preliminaries, we solve our problem for (4) assumed known in Section III. In Section IV, we modify ADP in [9], [12] to account for C3). Then, we present a new set of equations which is solved for $\boldsymbol{\Pi}_x^*$, $\boldsymbol{F}^*$ in view of C1-2). An online algorithm and a simulation demonstrating C4) are provided in Section V. Finally, we give concluding remarks.

II. MATHEMATICAL PRELIMINARIES

We start by defining operations which are used in ADP, e.g. [8], [12], [21]. For a matrix $\boldsymbol{P} = \boldsymbol{P}^{\mathsf{T}} \in \mathbb{R}^{n\times n}$, the bijective mapping $\widehat{\boldsymbol{P}} : \mathbb{R}^{n\times n} \to \mathbb{R}^{0.5n(n+1)}$ is defined as

$$\widehat{\boldsymbol{P}} := \begin{bmatrix} p_{1,1} & 2p_{1,2} & \cdots & 2p_{1,n} & p_{2,2} & 2p_{2,3} & \cdots & 2p_{n-1,n} & p_{n,n}\end{bmatrix}^{\mathsf{T}} \quad (6)$$

where $p_{i,j}$ is the element in the i-th row and j-th column of $\boldsymbol{P}$. For vector $\boldsymbol{x} \in \mathbb{R}^n$, we define $\underline{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^{0.5n(n+1)}$ as

$$\underline{\boldsymbol{x}} := \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_n & x_2^2 & x_2x_3 & \cdots & x_{n-1}x_n & x_n^2\end{bmatrix}^{\mathsf{T}}. \quad (7)$$

Furthermore, we will make use of the *Kronecker product* denoted by $\boldsymbol{A} \otimes \boldsymbol{B} : \mathbb{R}^{m\times n} \times \mathbb{R}^{o\times p} \to \mathbb{R}^{mo\times np}$ for which the following relations hold:

$$\big(\boldsymbol{A} \otimes \boldsymbol{B}\big)^\mathsf{T} = \boldsymbol{A}^\mathsf{T} \otimes \boldsymbol{B}^\mathsf{T}, \tag{8a}$$

$$\big(\boldsymbol{A} \otimes \boldsymbol{B}\big)\big(\boldsymbol{C} \otimes \boldsymbol{D}\big) = \boldsymbol{AC} \otimes \boldsymbol{BD}, \tag{8b}$$

$$\boldsymbol{L}^{(o,m)}\big(\boldsymbol{A} \otimes \boldsymbol{B}\big)\boldsymbol{L}^{(n,p)} = \boldsymbol{B} \otimes \boldsymbol{A}, \tag{8c}$$

$$\mathrm{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\mathsf{T} \otimes \boldsymbol{A})\mathrm{vec}(\boldsymbol{B}), \tag{8d}$$

$$\boldsymbol{x}^\mathsf{T}\boldsymbol{P}\boldsymbol{x} = \underline{\boldsymbol{x}}^\mathsf{T}\widehat{\boldsymbol{P}} = \big(\boldsymbol{x}^\mathsf{T} \otimes \boldsymbol{x}^\mathsf{T}\big)\,\boldsymbol{T}^\mathsf{T}\widehat{\boldsymbol{P}}, \tag{8e}$$

where we used $\mathrm{vec}(\boldsymbol{A}) : \mathbb{R}^{n\times m} \to \mathbb{R}^{nm}$ that stacks the columns of $\boldsymbol{A}$ and the commutation matrix $\boldsymbol{L}^{(q,r)} = \sum_{i=1}^{q}\sum_{j=1}^{r}(\boldsymbol{e}_{i,q}\boldsymbol{e}_{j,r}^\mathsf{T}) \otimes (\boldsymbol{e}_{j,r}\boldsymbol{e}_{i,q}^\mathsf{T})$ with the unit vector $\boldsymbol{e}_{k,l}$ of length l with k-th element equal one, zero else. Notice that in (8e) the mappings (6), (7) and $\underline{\boldsymbol{x}} = \boldsymbol{T}\,(\boldsymbol{x} \otimes \boldsymbol{x})$ were utilized, where each row of $\boldsymbol{T}$ is given by a specific $\boldsymbol{e}_{i,n^2}^\mathsf{T}$ according to (7) and $\boldsymbol{T}$ has full row rank.

We also use the *notations*: The zero and identity matrix of specified dimensions by $\boldsymbol{0}_{a\times b}$ and $\boldsymbol{I}_a$, a positive (semi-) definite matrix $\boldsymbol{M}$ by $\boldsymbol{M} \succ (\succeq)\boldsymbol{0}$, the spectrum of a matrix $\boldsymbol{M}$ by $\sigma(\boldsymbol{M})$, 2-norm: $\|\cdot\|_2$ and Frobenius norm: $\|\cdot\|_\mathrm{F}$.

III. Time-Invariant Control in Linear Quadratic Tracking

Based on [3], we derive control $\boldsymbol{u}^*(\cdot)$ accounting for C1-3) if the data (4) is known. Besides, Kleinman's algorithm [13] to solve *algebraic Riccati equations* (ARE) is stated.

In view of C3), we want to consider different costs for transition and stationary behavior as suggested by [3]. Exploiting the structure of $\boldsymbol{u}^*(\cdot)$ given by (5), we rewrite the corresponding solution of (1a) as $\boldsymbol{x} = \widetilde{\boldsymbol{x}} + \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}$ dividing the stationary behavior $\boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}$ that is determined by

$$\boldsymbol{\Pi}_x^*\overline{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{\Pi}_x^* + \boldsymbol{B}\boldsymbol{F}^* + \boldsymbol{E}_\mathrm{d}$$

from the transition $\widetilde{\boldsymbol{x}}$ defined by the transient dynamics

$$\dot{\widetilde{\boldsymbol{x}}} = \boldsymbol{A}\widetilde{\boldsymbol{x}} + \boldsymbol{B}\widetilde{\boldsymbol{u}} \tag{9}$$

with $\widetilde{\boldsymbol{x}}_0 = \boldsymbol{x}_0 - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}_0$ and $\widetilde{\boldsymbol{u}} = \boldsymbol{u} - \boldsymbol{F}^*\overline{\boldsymbol{x}}$. This allows us to formulate an LQRP which accounts for an optimal transition:

***Linear Quadratic Regulator Problem** 2:* For given $\alpha \geq 0$ and dynamics (9), find the control $\widetilde{\boldsymbol{u}}^*(\cdot)$ which minimizes

$$\widetilde{J}(\widetilde{\boldsymbol{u}}(\cdot))\big|_{\widetilde{\boldsymbol{x}}(t_\mathrm{E})} = \tfrac{1}{2}\int_{t_\mathrm{E}}^{\infty} \mathrm{e}^{2\alpha t}\left(\widetilde{\boldsymbol{x}}^\mathsf{T}\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{x}} + \widetilde{\boldsymbol{u}}^\mathsf{T}\widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{u}}\right)\,\mathrm{d}t, \tag{10}$$

where $\widetilde{\boldsymbol{Q}} \succeq \boldsymbol{0}$, $\widetilde{\boldsymbol{R}} \succ \boldsymbol{0}$ are constant and $t_\mathrm{E} \geq 0$.

At all times, we consider the following assumptions

Assumption 1: $(\boldsymbol{A} + \alpha\boldsymbol{I}, \boldsymbol{B})$ is stabilizable.

Assumption 2: For some $\boldsymbol{V}^\mathsf{T}\boldsymbol{V} = \boldsymbol{C}^\mathsf{T}\boldsymbol{Q}\boldsymbol{C}$, $(\boldsymbol{V}, \boldsymbol{A})$ is detectable and for some $\boldsymbol{W}^\mathsf{T}\boldsymbol{W} = \widetilde{\boldsymbol{Q}}$, $(\boldsymbol{W}, \boldsymbol{A})$ is observable.

Assumption 3: It holds $\sigma(\overline{\boldsymbol{A}}) \subset \{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) \geq 0\}$.

Assumption 4: The characteristic polynomial of $\overline{\boldsymbol{A}}$ is a minimal polynomial (for a definition see, e.g., [19, p. 68]).

Asmp. 1 and 2 are standard in LQTP. The first is true $\forall\alpha \geq 0$ if $(\boldsymbol{A}, \boldsymbol{B})$ is controllable. The second implies that $\forall\alpha \geq 0$, $(\boldsymbol{W}, \boldsymbol{A} + \alpha\boldsymbol{I})$ is observable. In view of Asmp. 3, we simply include any asymptotically stable part of (2a) into (1a). Asmp. 4 specifies that the geometric multiplicity of each distinct eigenvalue of $\overline{\boldsymbol{A}}$ equals one. It is also needed in, e.g., [8], [9], [17], [18], and we will remark where it is required.

We give a result extending [3, Coroll. 10], where we firstly consider *bounded* references in view of stationary tracking:

Lemma 1: For any given $\alpha \geq 0$ and if $\forall\overline{\boldsymbol{x}}_0\ \exists M \in \mathbb{R}^{>0}$ such that $\|\overline{\boldsymbol{x}}(t)\|_2 \leq M$, $\forall t \in [0, \infty)$, then for the control

$$\boldsymbol{u}^* = \underbrace{-\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{P}(\boldsymbol{x}^* - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}})}_{-\boldsymbol{K}^*\widetilde{\boldsymbol{x}}^*}\underbrace{-\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{\Pi}_\phi^*\overline{\boldsymbol{x}}}_{+\boldsymbol{F}^*\overline{\boldsymbol{x}}}, \tag{11}$$

where *Riccati* matrix $\boldsymbol{P} = \boldsymbol{P}^\mathsf{T} \succ \boldsymbol{0}$ uniquely solves ARE

$$\boldsymbol{P}(\boldsymbol{A} + \alpha\boldsymbol{I}) + (\boldsymbol{A} + \alpha\boldsymbol{I})^\mathsf{T}\boldsymbol{P} - \boldsymbol{P}\boldsymbol{B}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{P} = -\widetilde{\boldsymbol{Q}} \tag{12}$$

and $\boldsymbol{\Pi}_x^*$, $\boldsymbol{\Pi}_\phi^*$ with $\boldsymbol{S} = \boldsymbol{C}^\mathsf{T}\boldsymbol{Q}(\overline{\boldsymbol{C}} - \boldsymbol{D}_\mathrm{d})$ uniquely solve

$$\begin{bmatrix}\boldsymbol{\Pi}_x\\ \boldsymbol{\Pi}_\phi\end{bmatrix}\overline{\boldsymbol{A}} = \underbrace{\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\\ -\boldsymbol{C}^\mathsf{T}\boldsymbol{Q}\boldsymbol{C} & -\boldsymbol{A}^\mathsf{T}\end{bmatrix}}_{=\boldsymbol{H}}\begin{bmatrix}\boldsymbol{\Pi}_x\\ \boldsymbol{\Pi}_\phi\end{bmatrix} + \begin{bmatrix}\boldsymbol{E}_\mathrm{d}\\ \boldsymbol{S}\end{bmatrix}, \tag{13}$$

it is satisfied for any given $\boldsymbol{x}(t_\mathrm{E}) \in \mathbb{R}^n$ and $\overline{\boldsymbol{x}}(t_\mathrm{E}) \in \mathbb{R}^{\overline{n}}$:

1) The pair $(\widetilde{\boldsymbol{x}}^*, \widetilde{\boldsymbol{u}}^*)$ with $\widetilde{\boldsymbol{u}}^* = -\boldsymbol{K}^*\widetilde{\boldsymbol{x}}^*$ is the optimal solution of LQRP 2. That is, $\boldsymbol{u}^*$ leads to an **optimal transition** $\lim_{t\to\infty}\widetilde{\boldsymbol{x}}^* = \boldsymbol{0}$ with respect to cost (10).
2) The closed-loop matrix $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}^*$ is *Hurwitz* and the degree of stability α is guaranteed, i.e. $\max_{\lambda\in\sigma(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}^*)}\mathrm{Re}\,(\lambda) < -\alpha$.
3) For any admissible control $\boldsymbol{u}(\cdot) \not\equiv \boldsymbol{u}^*(\cdot)$ on $[t_\mathrm{E}, \infty)$ such that $\boldsymbol{x}(t) - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}(t) \not\to \boldsymbol{0}$ for $t \to \infty$ it holds
$$\lim_{T\to\infty} J_{t_\mathrm{E},T}(\boldsymbol{u}(\cdot)) - J_{t_\mathrm{E},T}(\boldsymbol{u}^*(\cdot)) = +\infty.$$
That is, $\boldsymbol{u}^*(\cdot)$ leads to an **optimal stationary solution** in the sense that any deviation from $\boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}$ results in infinite additional cost.

Proof: 1) and 2) are standard results, c.f. [1, Sec. 3.5]. In view of 3), we extend [3, Coroll. 10]. Note that (13) parametrizes the particular solution of the *Hamiltonian* system which emerges from the necessary optimality conditions for LQTP 1 by [10]. Due to the uniqueness of solutions of (13), e.g. see [4, Thm. 3], $\boldsymbol{u}^*(\cdot)$ and the unique optimal solution $\widehat{\boldsymbol{u}}(\cdot)$ of LQTP 1 derived in [3, Thm. 9] share the same stationary state $\boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}$ and pre-filter $\boldsymbol{F}^*\overline{\boldsymbol{x}}$, i.e. $\widehat{\boldsymbol{u}}(\cdot) = -\widehat{\boldsymbol{K}}(\widehat{\boldsymbol{x}} - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}) + \boldsymbol{F}^*\overline{\boldsymbol{x}}$. Actually, if $\widetilde{\boldsymbol{Q}} = \boldsymbol{C}^\mathsf{T}\boldsymbol{Q}\boldsymbol{C}$, $\widetilde{\boldsymbol{R}} = \boldsymbol{R}$ and $\alpha = 0$, then $\boldsymbol{u}^*(\cdot) \equiv \widehat{\boldsymbol{u}}(\cdot)$.

As a consequence, we have $\lim_{t\to\infty}\boldsymbol{x}^*(t) - \widehat{\boldsymbol{x}}(t) = \boldsymbol{0}$ and $\lim_{t\to\infty}\boldsymbol{u}^*(t) - \widehat{\boldsymbol{u}}(t) = \boldsymbol{0}$ both exponentially in t. By means of the *calculus of variations*, omitting the technicalities, it is rather evident to conclude that $\lim_{T\to\infty} J_{t_\mathrm{E},T}(\boldsymbol{u}^*(\cdot)) - J_{t_\mathrm{E},T}(\widehat{\boldsymbol{u}}(\cdot)) \leq M(\boldsymbol{x}(t_\mathrm{E}), \overline{\boldsymbol{x}}(t_\mathrm{E}))$, $M(\boldsymbol{x}(t_\mathrm{E}), \overline{\boldsymbol{x}}(t_\mathrm{E})) > 0$. For showing this, one could study the first case in the given case study of [3, proof of Thm. 9]. Furthermore, we recall that it indeed holds $\lim_{T\to\infty} J_{t_\mathrm{E},T}(\boldsymbol{u}(\cdot)) - J_{t_\mathrm{E},T}(\widehat{\boldsymbol{u}}(\cdot)) = +\infty$, see [3, Coroll. 10].

Using both, 3) readily follows after minor calculations which we leave to the reader. ∎

In case of *unbounded* references and disturbances, 3) cannot be shown. But then, one is interested in a finite horizon rather

than $T \to \infty$. To this end, we adopt a result proven in [3, Thm. 12] which focuses on the stationary behavior here:

Lemma 2: Consider $T_1 \geq 0$: $\boldsymbol{x}(T_1) = \boldsymbol{\Pi}_x^* \overline{\boldsymbol{x}}(T_1)$. If

$$\mathrm{Re}(\lambda_H) + \mathrm{Re}(\overline{\lambda}) < 0 \tag{14}$$

for all $\lambda_H \in \sigma(\boldsymbol{H}) \cap \{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) < 0\}$ and $\overline{\lambda} \in \sigma(\overline{\boldsymbol{A}})$, the pair $(\boldsymbol{\Pi}_x^* \overline{\boldsymbol{x}}, \boldsymbol{F}^* \overline{\boldsymbol{x}})$ which solves (1a) on $[T_1, \infty)$ satisfies

$$J_{T_1,\theta}(\boldsymbol{F}^* \overline{\boldsymbol{x}})\big|_{\overline{\boldsymbol{x}}(T_1)} - \lim_{T \to \infty} J_{T_1,\theta}(\boldsymbol{u}_T^*(\cdot))\big|_{\overline{\boldsymbol{x}}(T_1)} = 0 \tag{15}$$

for any $\theta \geq T_1$ and any given $\overline{\boldsymbol{x}}(T_1) \in \mathbb{R}^{\overline{n}}$, where $\boldsymbol{u}_T^*(\cdot)$ is the optimal control of LQTP 1 on a finite horizon $[T_1, T]$. ■

Clearly, (15) is a very desirable property which we put into words. Assume the transition phase ends after some T_1, i.e. $\boldsymbol{x}^*(T_1) \approx \boldsymbol{\Pi}_x^* \overline{\boldsymbol{x}}(T_1)$, and T is large, then the stationary behavior $\boldsymbol{\Pi}_x^* \overline{\boldsymbol{x}}$ induced by $\boldsymbol{u}^*(\cdot)$ closely approximates the optimal solution $\boldsymbol{x}_T^*(\cdot)$ induced by $\boldsymbol{u}_T^*(\cdot)$ on the interval $[T_1, \theta]$. That is, both produce approximatively the same cost as implied by (15). In this way, $\boldsymbol{u}^*(\cdot)$ qualifies as an *agreeable plan* on $[T_1, T]$, cf. [6]. Meaning that for a large horizon T, which may not be exactly known, it is agreeable to apply $\boldsymbol{u}^*(\cdot)$ instead of $\boldsymbol{u}_T^*(\cdot)$ over a large interval.

Remark 1: Condition (14) needs only to be considered if $\exists \overline{\lambda} \in \sigma(\overline{\boldsymbol{A}})$: $\mathrm{Re}(\overline{\lambda}) > 0$ and should be satisfied for a "large" $\boldsymbol{Q}$ in most cases.

Our main focus lies on deriving (11) without the need of the unknown data (4). The associated cost of $\boldsymbol{u}^*(\cdot)$ will serve as a starting point later on. It is concisely written as

$$\begin{aligned} J_{t,T}(\boldsymbol{u}^*(\cdot)) = \tfrac{1}{2}\boldsymbol{x}(t)^\mathsf{T} \boldsymbol{P} \boldsymbol{x}(t) + \boldsymbol{x}(t)^\mathsf{T}(-\boldsymbol{P}\boldsymbol{\Pi}_x^* + \boldsymbol{\Pi}_\phi^*)\overline{\boldsymbol{x}}(t) \\ + z(\overline{\boldsymbol{x}}(t), \boldsymbol{x}(T), \overline{\boldsymbol{x}}(T), t, T) \end{aligned} \tag{16}$$

if $\widetilde{\boldsymbol{Q}} = \boldsymbol{C}^\mathsf{T} \boldsymbol{Q} \boldsymbol{C}$, $\widetilde{\boldsymbol{R}} = \boldsymbol{R}$, $\alpha = 0$, some terms substituted by z. Note that (16) is generally unbounded as $\lim_{T \to \infty} z = \infty$.

Finally, let us introduce Kleinman's algorithm [13] which lays the foundation for adaptive dynamic programming in, e.g., [9], [12], [21]. We adopt the results of [13] by minor modifications to our case of a specified degree of stability:

Lemma 3: For a given $\alpha \geq 0$, suppose $\boldsymbol{K}_0$ such that $(\boldsymbol{A} + \alpha \boldsymbol{I} - \boldsymbol{B}\boldsymbol{K}_0)$ is *Hurwitz*. Consider

$$\boldsymbol{P}_k \boldsymbol{A}_k + \boldsymbol{A}_k^\mathsf{T} \boldsymbol{P}_k + 2\alpha \boldsymbol{P}_k = -\widetilde{\boldsymbol{Q}}_k, \tag{17a}$$

$$\text{For } k \geq 1\text{: } \boldsymbol{K}_k = \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B}^\mathsf{T} \boldsymbol{P}_{k-1}, \tag{17b}$$

with $\boldsymbol{A}_k = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}_k$ and $\widetilde{\boldsymbol{Q}}_k = \widetilde{\boldsymbol{Q}} + \boldsymbol{K}_k^\mathsf{T} \widetilde{\boldsymbol{R}} \boldsymbol{K}_k$ then for any $k \in \mathbb{N}_0$ the following holds:

1) $(\boldsymbol{A} + \alpha \boldsymbol{I} - \boldsymbol{B}\boldsymbol{K}_k)$ is *Hurwitz*. As a consequence, $\max_{\lambda \in \sigma(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}_k)} \mathrm{Re}\,(\lambda) < -\alpha$.
2) $\boldsymbol{P}_k$ is symmetric and $\boldsymbol{P}_k \succeq \boldsymbol{P}_{k+1} \succeq \boldsymbol{P} \succ \boldsymbol{0}$ where $\boldsymbol{P}$ uniquely solves (12).
3) $\lim_{k \to \infty} \boldsymbol{P}_k = \boldsymbol{P}$. ■

IV. Application to Unknown Dynamics: An ADP Approach

For the first time, we achieve C2) despite the unknown data (4). That is, implementing $\boldsymbol{u}^*(\cdot)$ as in (11) with desired properties as stated in Lemma 1 and 2. Apparently, we have to find $\boldsymbol{K}^*$, $\boldsymbol{\Pi}_x^*$ and $\boldsymbol{F}^*$.

To this end, we present C1). Firstly, we show how to obtain $\boldsymbol{K}_k \approx \boldsymbol{K}^*$ iteratively. In the process, unknown $\boldsymbol{B}$ and $\boldsymbol{E}_\mathrm{d}$ are also provided. It is based on adaptive dynamic programming (ADP) [9], [12]. We extend these methods to achieve a desired degree of stability C3). Secondly, we introduce a new approach which allows us to determine $\boldsymbol{\Pi}_x^*$ and $\boldsymbol{F}^*$ *exactly* in spite of the unknown $\boldsymbol{A}$, $\overline{\boldsymbol{A}}$ and approximation $\boldsymbol{P}_k \approx \boldsymbol{P}$.

A. Optimal Transition of Unknown Dynamics

Our goal is to solve LQRP 2 despite the unknown data (4). LQRPs for $\alpha = 0$ were approached in [9], [12] and [21] by the method of ADP. Following [12], the idea is to collect *measurement data* containing sufficient information on the unknown dynamics (4). Thus, we collect measurement data depending on $\boldsymbol{x}(t)$, $\boldsymbol{u}(t)$ and $\overline{\boldsymbol{x}}(t)$ during an exploration phase $[0, t_\mathrm{E}]$ where the system (1a) is excited by

$$\boldsymbol{u} = -\boldsymbol{K}_0 \boldsymbol{x} + \boldsymbol{e} \tag{18}$$

with $\boldsymbol{e}$ being a *known* exploration noise. Afterwards, a set of equations must be provided which is equivalent to (17) but depends only on the collected measurement data instead of unknown structural data (4). Then, iteratively solving these equations gives $\boldsymbol{K}_k \to \boldsymbol{K}^*$ for $k \to \infty$ as desired.

Remark 2: When the open loop is not stable, requiring $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}_0$ being *Hurwitz* is a crucial step in Lemma 3 and, thus, in [9], [12], etc.. As suggested by [8], some information on the system parameters is usually available. Suppose we know convex polytopes $\mathcal{A}$ and $\mathcal{B}$, then an optimization-based design [5] can be used to obtain a stabilizing $\boldsymbol{K}_0$ for any $\boldsymbol{A} \in \mathcal{A}$, $\boldsymbol{B} \in \mathcal{B}$ and to quantify a minimal degree of stability α_0. This motivates our extension of the results in [9], [12], i.e. we can specify $0 \leq \alpha < \alpha_0$ in LQRP 2. By choosing α close to α_0, advantageously, any $\boldsymbol{K}_k$ will preserve the degree of stability $\alpha \approx \alpha_0$ independent of our choice of $\widetilde{\boldsymbol{Q}}$, $\widetilde{\boldsymbol{R}}$.

Subsequently, we derive the desired set of equations and give a unique solution. We follow the derivations in [9], [12] and extend these for the case $\alpha \geq 0$. Let us denote $\boldsymbol{x}(t)$ given by (1a) for control law (18). We integrate the derivative of $\boldsymbol{x}(t)^\mathsf{T} \boldsymbol{P}_k \boldsymbol{x}(t)$, $k \in \mathbb{N}_0$, over some period $[t_i, t_{i+1}] \subset [0, t_\mathrm{E}]$ with $t_{i+1} \geq t_i$, $i \in \mathbb{N}$ and $t_1 = 0$. It holds

$$\begin{aligned} &\boldsymbol{x}^\mathsf{T}(t_{i+1}) \boldsymbol{P}_k \boldsymbol{x}(t_{i+1}) - \boldsymbol{x}^\mathsf{T}(t_i) \boldsymbol{P}_k \boldsymbol{x}(t_i) + 2\alpha \int_{t_i}^{t_{i+1}} \boldsymbol{x}^\mathsf{T} \boldsymbol{P}_k \boldsymbol{x} \, \mathrm{d}\tau \\ &= \int_{t_i}^{t_{i+1}} \boldsymbol{x}^\mathsf{T} \Big(\boldsymbol{A}_k^\mathsf{T} \boldsymbol{P}_k + \boldsymbol{P}_k \boldsymbol{A}_k + 2\alpha \boldsymbol{P}_k + 2\left(\boldsymbol{K}_k \boldsymbol{x}\right)^\mathsf{T} \boldsymbol{B}^\mathsf{T} \boldsymbol{P}_k \Big) \boldsymbol{x} \\ &\qquad + 2\boldsymbol{u}^\mathsf{T} \boldsymbol{B}^\mathsf{T} \boldsymbol{P}_k \boldsymbol{x} + 2\left(\boldsymbol{E}_\mathrm{d} \overline{\boldsymbol{x}}\right)^\mathsf{T} \boldsymbol{P}_k \boldsymbol{x} \, \mathrm{d}\tau \\ &= \int_{t_i}^{t_{i+1}} -\boldsymbol{x}^\mathsf{T} \widetilde{\boldsymbol{Q}}_k \boldsymbol{x} + 2\left(\boldsymbol{K}_k \boldsymbol{x} + \boldsymbol{u}\right)^\mathsf{T} \widetilde{\boldsymbol{R}} \boldsymbol{K}_{k+1} \boldsymbol{x} + 2\overline{\boldsymbol{x}}^\mathsf{T} \boldsymbol{E}_\mathrm{d}^\mathsf{T} \boldsymbol{P}_k \boldsymbol{x} \, \mathrm{d}\tau \end{aligned} \tag{19}$$

where we used (17a) and $\boldsymbol{B}^\mathsf{T} \boldsymbol{P}_k = \widetilde{\boldsymbol{R}} \boldsymbol{K}_{k+1}$ based on (17b).

At step k, (19) obviously depends on three unknown variables: $\boldsymbol{K}_{k+1}$, symmetric $\boldsymbol{P}_k$ and the product $\boldsymbol{E}_\mathrm{d}^\mathsf{T} \boldsymbol{P}_k$. Consequently, we follow the suggestion of [12] and regard $i = 1, \ldots, h$ time intervals such that $[0, t_\mathrm{E}] = \cup_{i=1}^h [t_i, t_{i+1}]$. Then, stacking the h equations (19) leads to a set of equations which can be solved for the unknowns similarly as in [9].

With this in mind, (19) is reformulated by means of the mappings (6), (7) and relations (8), see Section II:

$$\begin{aligned}\Big(\underline{\boldsymbol{x}}(t_{i+1}) - \underline{\boldsymbol{x}}(t_i) + 2\alpha \boldsymbol{T}\int_{t_i}^{t_{i+1}} (\boldsymbol{x}\otimes\boldsymbol{x})\ \mathrm{d}\tau\Big)^{\mathsf{T}} \widehat{\boldsymbol{P}}_k \\ = \int_{t_i}^{t_{i+1}} -\big(\boldsymbol{x}^{\mathsf{T}}\otimes\boldsymbol{x}^{\mathsf{T}}\big)\mathrm{vec}(\widetilde{\boldsymbol{Q}}_k) + 2\left(\boldsymbol{x}^{\mathsf{T}}\otimes\overline{\boldsymbol{x}}^{\mathsf{T}}\right)\mathrm{vec}(\boldsymbol{E}_{\mathrm{d}}^{\mathsf{T}}\boldsymbol{P}_k) \\ +2\Big(\big(\boldsymbol{x}^{\mathsf{T}}\otimes\boldsymbol{x}^{\mathsf{T}}\big)\big(\boldsymbol{I}_n\otimes\boldsymbol{K}_k^{\mathsf{T}}\widetilde{\boldsymbol{R}}\big) + \big(\boldsymbol{x}^{\mathsf{T}}\otimes\boldsymbol{u}^{\mathsf{T}}\big)\big(\boldsymbol{I}_n\otimes\widetilde{\boldsymbol{R}}\big)\Big) \\ \cdot\,\mathrm{vec}(\boldsymbol{K}_{k+1})\ \mathrm{d}\tau. \quad (20)\end{aligned}$$

Before we stack h equations (20), let us introduce

$$\boldsymbol{\Delta}_{xx} = \begin{bmatrix}\underline{\boldsymbol{x}}(t_2)-\underline{\boldsymbol{x}}(t_1) & \cdots & \underline{\boldsymbol{x}}(t_{h+1})-\underline{\boldsymbol{x}}(t_h)\end{bmatrix}^{\mathsf{T}}, \quad (21a)$$

$$\boldsymbol{\Lambda}_{xx} = \begin{bmatrix}\int_{t_1}^{t_2}\boldsymbol{x}\otimes\boldsymbol{x}\ \mathrm{d}\tau & \cdots & \int_{t_h}^{t_{h+1}}\boldsymbol{x}\otimes\boldsymbol{x}\ \mathrm{d}\tau\end{bmatrix}^{\mathsf{T}}, \quad (21b)$$

$$\boldsymbol{\Lambda}_{xu} = \begin{bmatrix}\int_{t_1}^{t_2}\boldsymbol{x}\otimes\boldsymbol{u}\ \mathrm{d}\tau & \cdots & \int_{t_h}^{t_{h+1}}\boldsymbol{x}\otimes\boldsymbol{u}\ \mathrm{d}\tau\end{bmatrix}^{\mathsf{T}}, \quad (21c)$$

$$\boldsymbol{\Lambda}_{x\overline{x}} = \begin{bmatrix}\int_{t_1}^{t_2}\boldsymbol{x}\otimes\overline{\boldsymbol{x}}\ \mathrm{d}\tau & \cdots & \int_{t_h}^{t_{h+1}}\boldsymbol{x}\otimes\overline{\boldsymbol{x}}\ \mathrm{d}\tau\end{bmatrix}^{\mathsf{T}} \quad (21d)$$

with $\boldsymbol{\Delta}_{xx}\in\mathbb{R}^{h\times 0.5n(n+1)}$, $\boldsymbol{\Lambda}_{xx}\in\mathbb{R}^{h\times n^2}$, $\boldsymbol{\Lambda}_{xu}\in\mathbb{R}^{h\times nm}$, $\boldsymbol{\Lambda}_{x\overline{x}}\in\mathbb{R}^{h\times n\overline{n}}$. Finally, one can obtain the set of equations

$$\boldsymbol{\Theta}_k\begin{bmatrix}\widehat{\boldsymbol{P}}_k\\ \mathrm{vec}(\boldsymbol{K}_{k+1})\\ \mathrm{vec}(\boldsymbol{E}_{\mathrm{d}}^{\mathsf{T}}\boldsymbol{P}_k)\end{bmatrix} = -\boldsymbol{\Lambda}_{xx}\mathrm{vec}(\widetilde{\boldsymbol{Q}}_k) \quad (22)$$

with $\boldsymbol{\Theta}_k\in\mathbb{R}^{h\times(0.5n(n+1)+nm+n\overline{n})}$ given by

$$\boldsymbol{\Theta}_k = \begin{bmatrix}\widetilde{\boldsymbol{\Delta}}_{xx} & -2\Big(\boldsymbol{\Lambda}_{xx}\big(\boldsymbol{I}_n\otimes\boldsymbol{K}_k^{\mathsf{T}}\widetilde{\boldsymbol{R}}\big)+\boldsymbol{\Lambda}_{xu}\big(\boldsymbol{I}_n\otimes\widetilde{\boldsymbol{R}}\big)\Big) & -2\boldsymbol{\Lambda}_{x\overline{x}}\end{bmatrix}$$

where $\widetilde{\boldsymbol{\Delta}}_{xx} = \boldsymbol{\Delta}_{xx} + 2\alpha\boldsymbol{\Lambda}_{xx}\boldsymbol{T}^{\mathsf{T}}$. When $\boldsymbol{\Theta}_k$ has full column rank, clearly

$$\begin{bmatrix}\widehat{\boldsymbol{P}}_k\\ \mathrm{vec}(\boldsymbol{K}_{k+1})\\ \mathrm{vec}(\boldsymbol{E}_{\mathrm{d}}^{\mathsf{T}}\boldsymbol{P}_k)\end{bmatrix} = -\left(\boldsymbol{\Theta}_k^{\mathsf{T}}\boldsymbol{\Theta}_k\right)^{-1}\boldsymbol{\Theta}_k^{\mathsf{T}}\boldsymbol{\Lambda}_{xx}\mathrm{vec}(\boldsymbol{Q}_k) \quad (23)$$

is the least square solution of (22). Existence can be guaranteed by a condition stated in the following lemma which is similar to [12, Lem. 6]:

Lemma 4: Suppose it holds

$$\mathrm{rank}\begin{pmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_{xx} & \boldsymbol{\Lambda}_{xu} & \boldsymbol{\Lambda}_{x\overline{x}}\end{bmatrix}\end{pmatrix} = \tfrac{1}{2}n(n+1)+nm+n\overline{n} \quad (24)$$

for the given $[0,t_{\mathrm{E}}] = \cup_{i=1}^{h}[t_i,t_{i+1}]$, then, for any $k\in\mathbb{N}_0$, $\boldsymbol{\Theta}_k$ has full column rank.

Proof:... goes along the same lines as [9, proof of Lem. 3 for $i=0$] with some modifications to cover $\alpha\geq 0$. ■

Remark 3: Obviously, the exploration noise $\boldsymbol{e}$ needs to be *sufficiently rich* in the sense that (24) holds. A typical choice is random noise or sums of sinusoidal signals [12]. From a practical perspective, the latter is preferable as it does not violate actuator rate constraints. Due to the structure of (21), we may need less frequencies than the suggested $\frac{1}{2}\big(\frac{1}{2}n(n+1)+nm+n\overline{n}\big)$ by [11, Sec. 5.2.1], cf. Section V.

Based on the derivation of (19), we know that, for any $k\in\mathbb{N}_0$, the solution $\boldsymbol{P}_k$, $\boldsymbol{K}_{k+1}$ of (17) also satisfies (22). Hence, an exact solution of (22) exists which, of course, is given by the least square solution (23) when (24) holds. As a consequence, iteratively solving (17) is equivalent to iteratively calculating (23). Following this discussion, the next theorem is self-evident. It extends [9, Thm. 2].

Theorem 1: For a given $\alpha\geq 0$, suppose $\boldsymbol{K}_0$ such that $\boldsymbol{A}+\alpha\boldsymbol{I}-\boldsymbol{B}\boldsymbol{K}_0$ is *Hurwitz* and condition (24) holds. For any $k\in\mathbb{N}_0$, $\boldsymbol{P}_k\succ\boldsymbol{0}$ and $\boldsymbol{K}_{k+1}$ given by the policy iteration (23) satisfy (17) and it holds $\lim_{k\to\infty}\boldsymbol{P}_k=\boldsymbol{P}$, $\lim_{k\to\infty}\boldsymbol{K}_k=\boldsymbol{K}^*$ where $\boldsymbol{P}$, $\boldsymbol{K}^*$ as in Lemma 1. ■

With $\boldsymbol{P}_k$ invertible for any $k\in\mathbb{N}_0$, we can *exactly* calculate $\boldsymbol{B}=\boldsymbol{P}_k^{-1}\boldsymbol{K}_{k+1}^{\mathsf{T}}\widetilde{\boldsymbol{R}}$ and $\boldsymbol{E}_{\mathrm{d}}$ from the third argument of (23) for some arbitrary $k\in\mathbb{N}_0$.

B. Optimal Stationary Behavior of Unknown Dynamics

In the previous section, we obtained $\boldsymbol{P}_k$, $\boldsymbol{K}_{k+1}$ as approximations of $\boldsymbol{P}$, $\boldsymbol{K}^*$ and $\boldsymbol{B}$, $\boldsymbol{E}_{\mathrm{d}}$ exactly. In order to implement $\boldsymbol{u}^*(\cdot)$, it is left to determine $\boldsymbol{\Pi}_x^*$ and $\boldsymbol{F}^*=-\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{\Pi}_\phi^*$, i.e. we need the *exact* solution of (13) despite $\boldsymbol{A}$ and $\overline{\boldsymbol{A}}$ unknown. We present a new methodology. As a result, we solve LQTP 1 with unbounded cost in a stationarily optimal sense despite the unknown data for the first time.

As we see, for obtaining (19) a term similar to the quadratic, first addend in (16) was analyzed. In the sequel, we follow the idea of providing an analogous analysis for a term similar to the bilinear, second addend in (16).

First, we take $\boldsymbol{P}_k$ for some $k\in\mathbb{N}_0$ and consider

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}^{\mathsf{T}}(t)\boldsymbol{P}_k\boldsymbol{\Pi}_x\overline{\boldsymbol{x}}(t)\\ &= \underbrace{\boldsymbol{x}^{\mathsf{T}}(\boldsymbol{P}_k\boldsymbol{\Pi}_x\overline{\boldsymbol{A}}+\boldsymbol{A}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{\Pi}_x)\overline{\boldsymbol{x}}}_{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}_k\left(\boldsymbol{\Pi}_x\overline{\boldsymbol{A}}-\boldsymbol{A}\boldsymbol{\Pi}_x\right)\overline{\boldsymbol{x}}+\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Omega}_k\boldsymbol{\Pi}_x\overline{\boldsymbol{x}}} + (\boldsymbol{B}\boldsymbol{u}+\boldsymbol{E}_{\mathrm{d}}\overline{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{\Pi}_x\overline{\boldsymbol{x}} \quad (25)\end{aligned}$$

where we used (17a) to replace $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{P}_k=-\boldsymbol{P}_k\boldsymbol{A}+\boldsymbol{\Omega}_k$ with

$$\boldsymbol{\Omega}_k = -\widetilde{\boldsymbol{Q}}_k - 2\alpha\boldsymbol{P}_k + \boldsymbol{P}_k\boldsymbol{B}\boldsymbol{K}_k + \boldsymbol{K}_k^{\mathsf{T}}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{P}_k.$$

After replacing $\boldsymbol{\Pi}_x\overline{\boldsymbol{A}}-\boldsymbol{A}\boldsymbol{\Pi}_x$ in (25) with the help of the first line of (13), integration of (25) over an interval $[t_i,t_{i+1}]$ as defined in Section IV-A yields

$$\begin{aligned}&\boldsymbol{x}^{\mathsf{T}}(t_{i+1})\boldsymbol{P}_k\boldsymbol{\Pi}_x\overline{\boldsymbol{x}}(t_{i+1}) - \boldsymbol{x}^{\mathsf{T}}(t_i)\boldsymbol{P}_k\boldsymbol{\Pi}_x\overline{\boldsymbol{x}}(t_i)\\ &\quad= \int_{t_i}^{t_{i+1}} -\boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{\Pi}_\phi\overline{\boldsymbol{x}} + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{E}_{\mathrm{d}}\overline{\boldsymbol{x}}\\ &\qquad+\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Omega}_k\boldsymbol{\Pi}_x\overline{\boldsymbol{x}} + (\boldsymbol{B}\boldsymbol{u}+\boldsymbol{E}_{\mathrm{d}}\overline{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{\Pi}_x\overline{\boldsymbol{x}}\ \mathrm{d}\tau. \quad (26)\end{aligned}$$

Second, we integrate the derivative of $\boldsymbol{x}(t)\boldsymbol{\Pi}_\phi\overline{\boldsymbol{x}}(t)$:

$$\begin{aligned}&\boldsymbol{x}^{\mathsf{T}}(t_{i+1})\boldsymbol{\Pi}_\phi\overline{\boldsymbol{x}}(t_{i+1}) - \boldsymbol{x}^{\mathsf{T}}(t_i)\boldsymbol{\Pi}_\phi\overline{\boldsymbol{x}}(t_i)\\ &= \int_{t_i}^{t_{i+1}}\boldsymbol{x}^{\mathsf{T}}\left(\boldsymbol{\Pi}_\phi\overline{\boldsymbol{A}}+\boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Pi}_\phi\right)\overline{\boldsymbol{x}} + (\boldsymbol{B}\boldsymbol{u}+\boldsymbol{E}_d\overline{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{\Pi}_\phi\overline{\boldsymbol{x}}\ \mathrm{d}\tau\\ &= \int_{t_i}^{t_{i+1}} -\boldsymbol{x}^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{C}\boldsymbol{\Pi}_x\overline{\boldsymbol{x}} + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{S}\overline{\boldsymbol{x}} + (\boldsymbol{B}\boldsymbol{u}+\boldsymbol{E}_d\overline{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{\Pi}_\phi\overline{\boldsymbol{x}}\ \mathrm{d}\tau\end{aligned} \quad (27)$$

where we used the second line of (13).

Summarizing, we have found two equations (26) and (27) which are satisfied by the unique solution of (13) but do not depend on the unknown $\boldsymbol{A}$, $\overline{\boldsymbol{A}}$. Again as in Section IV-A, the idea is to obtain a set of equations equivalent to (13) by stacking (26), (27) for h time intervals. For this purpose, we

$$\overline{\boldsymbol{\Theta}} = \begin{bmatrix} \boldsymbol{\Delta}_{\overline{x}x}(\boldsymbol{I}_{\overline{n}} \otimes \boldsymbol{P}_k) - \boldsymbol{\Lambda}_{\overline{x}x}(\boldsymbol{I}_{\overline{n}} \otimes \boldsymbol{\Omega}_k) - \boldsymbol{\Lambda}_{\overline{x}u}(\boldsymbol{I}_{\overline{n}} \otimes \boldsymbol{B}^{\mathsf{T}}\boldsymbol{P}_k) - \boldsymbol{\Lambda}_{\overline{x}\overline{x}}(\boldsymbol{I}_{\overline{n}} \otimes \boldsymbol{E}_{\mathrm{d}}^{\mathsf{T}}\boldsymbol{P}_k) & \boldsymbol{\Lambda}_{\overline{x}x}(\boldsymbol{I}_{\overline{n}} \otimes \boldsymbol{P}_k\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}}) \\ \boldsymbol{\Lambda}_{\overline{x}x}(\boldsymbol{I}_{\overline{n}} \otimes \boldsymbol{C}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{C}) & \boldsymbol{\Delta}_{\overline{x}x} - \boldsymbol{\Lambda}_{\overline{x}u}(\boldsymbol{I}_{\overline{n}} \otimes \boldsymbol{B}^{\mathsf{T}}) - \boldsymbol{\Lambda}_{\overline{x}\overline{x}}(\boldsymbol{I}_{\overline{n}} \otimes \boldsymbol{E}_{\mathrm{d}}^{\mathsf{T}}) \end{bmatrix} \quad (29)$$

could rewrite (26), (27) in a similar manner as done in (20) by means of the relations (8). Due to the space restrictions, however, we omit these technical steps here and proceed directly to formulate the desired set of equations.

By stacking measurements for $i = 1, \ldots, h$, we have

$$\boldsymbol{\Delta}_{\overline{x}x} = \begin{bmatrix} \cdots & \overline{\boldsymbol{x}}(t_{i+1}) \otimes \boldsymbol{x}(t_{i+1}) - \overline{\boldsymbol{x}}(t_i) \otimes \boldsymbol{x}(t_i) & \cdots \end{bmatrix}^{\mathsf{T}}, \quad (28a)$$

$$\boldsymbol{\Lambda}_{\overline{x}x} = \boldsymbol{\Lambda}_{x\overline{x}}\boldsymbol{L}^{(n,\overline{n})}, \quad (28b)$$

$$\boldsymbol{\Lambda}_{\overline{x}u} = \begin{bmatrix} \int_{t_1}^{t_2} \overline{\boldsymbol{x}} \otimes \boldsymbol{u} \, \mathrm{d}\tau & \cdots & \int_{t_h}^{t_{h+1}} \overline{\boldsymbol{x}} \otimes \boldsymbol{u} \, \mathrm{d}\tau \end{bmatrix}^{\mathsf{T}}, \quad (28c)$$

$$\boldsymbol{\Lambda}_{\overline{x}\overline{x}} = \begin{bmatrix} \int_{t_1}^{t_2} \overline{\boldsymbol{x}} \otimes \overline{\boldsymbol{x}} \, \mathrm{d}\tau & \cdots & \int_{t_h}^{t_{h+1}} \overline{\boldsymbol{x}} \otimes \overline{\boldsymbol{x}} \, \mathrm{d}\tau \end{bmatrix}^{\mathsf{T}} \quad (28d)$$

where we used $\boldsymbol{L}^{(1,1)} = 1$ and with $\boldsymbol{\Delta}_{\overline{x}x} \in \mathbb{R}^{h \times \overline{n}n}$, $\boldsymbol{\Lambda}_{\overline{x}u} \in \mathbb{R}^{h \times \overline{n}m}$, $\boldsymbol{\Lambda}_{\overline{x}\overline{x}} \in \mathbb{R}^{h \times \overline{n}^2}$. With $\overline{\boldsymbol{\Theta}} \in \mathbb{R}^{2h \times 2n\overline{n}}$ given in (29) on top of this page, we obtain

$$\overline{\boldsymbol{\Theta}} \begin{bmatrix} \mathrm{vec}(\boldsymbol{\Pi}_x) \\ \mathrm{vec}(\boldsymbol{\Pi}_\phi) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_{\overline{x}x}\mathrm{vec}(\boldsymbol{P}_k\boldsymbol{E}_{\mathrm{d}}) \\ \boldsymbol{\Lambda}_{\overline{x}x}\mathrm{vec}(\boldsymbol{S}) \end{bmatrix}. \quad (30)$$

With respect to solving (30), it is formulated:

Lemma 5: For the given $[0, t_{\mathrm{E}}] = \cup_{i=1}^{h}[t_i, t_{i+1}]$, suppose

$$\mathrm{rank}\,(\boldsymbol{\Lambda}_{\overline{x}x}) = n\overline{n} \quad (31)$$

and condition (14) hold, then $\overline{\boldsymbol{\Theta}}$ in (29) has full column rank.

Proof: We prove this by showing that

$$\overline{\boldsymbol{\Theta}} \begin{bmatrix} \mathrm{vec}(\boldsymbol{N}) \\ \mathrm{vec}(\boldsymbol{M}) \end{bmatrix} = \boldsymbol{0} \quad (32)$$

necessarily implies $\boldsymbol{N} = \boldsymbol{M} = \boldsymbol{0}$. To do so, we provide an analogous analysis of $\boldsymbol{x}(t)^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{N}\overline{\boldsymbol{x}}(t)$ and $\boldsymbol{x}(t)^{\mathsf{T}}\boldsymbol{M}\overline{\boldsymbol{x}}(t)$ as in (26) and (27), respectively. According to (25), integrating the derivative over $[t_i, t_{i+1}]$ and rearranging gives

$$\begin{aligned} &\boldsymbol{x}^{\mathsf{T}}(t_{i+1})\boldsymbol{P}_k\boldsymbol{N}\overline{\boldsymbol{x}}(t_{i+1}) - \boldsymbol{x}^{\mathsf{T}}(t_i)\boldsymbol{P}_k\boldsymbol{N}\overline{\boldsymbol{x}}(t_i) + \int_{t_i}^{t_{i+1}} -\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Omega}_k\boldsymbol{N}\overline{\boldsymbol{x}} \\ &\quad - (\boldsymbol{B}\boldsymbol{u} + \boldsymbol{E}_{\mathrm{d}}\overline{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{N}\overline{\boldsymbol{x}} + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{M}\overline{\boldsymbol{x}} \, \mathrm{d}\tau. \\ &\quad = \int_{t_i}^{t_{i+1}} \boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}_k\left(-\boldsymbol{A}\boldsymbol{N} + \boldsymbol{N}\overline{\boldsymbol{A}} + \boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{M}\right)\overline{\boldsymbol{x}} \, \mathrm{d}\tau. \end{aligned}$$

Corresponding to the first equation in (27), rearranging yields

$$\begin{aligned} &\boldsymbol{x}^{\mathsf{T}}(t_{i+1})\boldsymbol{M}\overline{\boldsymbol{x}}(t_{i+1}) - \boldsymbol{x}^{\mathsf{T}}(t_i)\boldsymbol{M}\overline{\boldsymbol{x}}(t_i) \\ &\quad + \int_{t_i}^{t_{i+1}} \boldsymbol{x}^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{C}\boldsymbol{N}\overline{\boldsymbol{x}} - (\boldsymbol{B}\boldsymbol{u} + \boldsymbol{E}_{\mathrm{d}}\overline{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{M}\overline{\boldsymbol{x}} \, \mathrm{d}\tau \\ &\quad = \int_{t_i}^{t_{i+1}} \boldsymbol{x}^{\mathsf{T}}\left(\boldsymbol{M}\overline{\boldsymbol{A}} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{M} + \boldsymbol{C}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{C}\boldsymbol{N}\right)\overline{\boldsymbol{x}} \, \mathrm{d}\tau. \end{aligned}$$

Choosing $i = 1, \ldots, h$, one can observe that the (i)-th row and $(h+i)$-th row on the left of (32) are precisely given by the left hand side of the just stated first and second equation, respectively. Consequently, (32) is equivalent to:

$$\begin{bmatrix} \boldsymbol{\Lambda}_{\overline{x}x}(\boldsymbol{I}_{\overline{n}} \otimes \boldsymbol{P}_k)\mathrm{vec}\Big(-\boldsymbol{A}\boldsymbol{N} + \boldsymbol{N}\overline{\boldsymbol{A}} + \boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{M}\Big) \\ \boldsymbol{\Lambda}_{\overline{x}x}\mathrm{vec}\Big(\boldsymbol{M}\overline{\boldsymbol{A}} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{M} + \boldsymbol{C}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{C}\boldsymbol{N}\Big) \end{bmatrix} = \boldsymbol{0}.$$

Due to (31) and $\boldsymbol{P}_k$ invertible, it results

$$\begin{bmatrix} \boldsymbol{N} \\ \boldsymbol{M} \end{bmatrix} \overline{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}} \\ -\boldsymbol{C}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{C} & -\boldsymbol{A}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{N} \\ \boldsymbol{M} \end{bmatrix} \quad (33)$$

which is the homogeneous part of (13). But, condition (14) is sufficient for $\sigma(\overline{\boldsymbol{A}}) \cap \sigma(\boldsymbol{H}) = \emptyset$ since $\boldsymbol{H}$ is *Hamiltonian*. Thus, cf. [19], $\boldsymbol{N} = \boldsymbol{M} = \boldsymbol{0}$ uniquely solve (33) and (32). ■

Taking the structure of $\boldsymbol{\Lambda}_{xx}$, $\boldsymbol{\Lambda}_{xu}$ and orthogonality of commutation matrix $\boldsymbol{L}^{(n,\overline{n})}$ in (28b) into account, we conclude

Corollary 1: Condition (24) readily implies (31). ■

Under these conditions, the least square solution of (30) reads

$$\begin{bmatrix} \mathrm{vec}(\boldsymbol{\Pi}_x^*) \\ \mathrm{vec}(\boldsymbol{\Pi}_\phi^*) \end{bmatrix} = \left(\overline{\boldsymbol{\Theta}}^{\mathsf{T}}\overline{\boldsymbol{\Theta}}\right)^{-1}\overline{\boldsymbol{\Theta}}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{\Lambda}_{\overline{x}x}\mathrm{vec}(\boldsymbol{P}_k\boldsymbol{E}_{\mathrm{d}}) \\ \boldsymbol{\Lambda}_{\overline{x}x}\mathrm{vec}(\boldsymbol{S}) \end{bmatrix}. \quad (34)$$

Following the derivations of (26) and (27), an exact solution of (30) is certainly given by the unique solution of (13). When (31) is true, this exact solution is unique and coincides with the least square solution (34); hence, (30) and (13) are equivalent sets of equations. Based on this discussion, the next theorem is self-evident again.

Theorem 2: For an arbitrary $k \in \mathbb{N}_0$, let $\boldsymbol{P}_k$ satisfy (17) where $\boldsymbol{K}_0$ such that $\boldsymbol{A} + \alpha\boldsymbol{I} - \boldsymbol{B}\boldsymbol{K}_0$ is *Hurwitz* for a given $\alpha \geq 0$. If conditions (14) and (31) are satisfied, then $\boldsymbol{\Pi}_x^*$ and $\boldsymbol{\Pi}_\phi^*$ obtained from (34) uniquely solve (13). ■

Remark 4: Asmp. 4 ensures that $\mathrm{rank}\,(\boldsymbol{\Lambda}_{x\overline{x}}) = n\overline{n}$ can be satisfied. Hence, it must also hold in [8], [9], [17] and [18]. It is dispensable if exosystem (2a) can be excited by exploration noise: $\overline{\boldsymbol{B}}\overline{\boldsymbol{e}}$. This is usually possible for the part of (2a) generating $\overline{\boldsymbol{y}}$ and requires minor modifications of (29).

V. Online Algorithm and Simulation

We briefly summarize our new results by providing an online algorithm for finding $\boldsymbol{u}^*(\cdot)$. We assume that Asmp. 1-4 and condition (14) hold.

***Algorithm** 1:*

1) *Initialization:* Give weights $\boldsymbol{Q} \succeq \boldsymbol{0}$, $\boldsymbol{R} \succ \boldsymbol{0}$ for LQTP 1 and $\widetilde{\boldsymbol{Q}} \succeq \boldsymbol{0}$, $\widetilde{\boldsymbol{R}} \succ \boldsymbol{0}$ for transition, i.e. LQRP 2. Find a stabilizing $\boldsymbol{K}_0$ which determines a minimal degree of stability $\overline{\alpha}_0$, e.g. by following Remark 2. Choose a desired guaranteed degree of stability $0 \leq \alpha < \overline{\alpha}_0$.
2) *Exploration:* Run systems (2a) and (1a) excited by control input $\boldsymbol{u} = -\boldsymbol{K}_0\boldsymbol{x} + \boldsymbol{e}$ with exploration noise $\boldsymbol{e}$ selected as, e.g., sum of sinusoidal signals (cf. Remark 3). Simultaneously, collect the measurement data (21) and (28) based on measuring $\boldsymbol{x}$, $\boldsymbol{u}$, $\overline{\boldsymbol{x}}$ and integration over each interval $[t_i, t_{i+1}]$, see Section IV-A. Increment $i \in \mathbb{N}$ until rank condition (24) holds.
3) *Policy-Iteration:* Iteratively calculate $\boldsymbol{P}_k$ and $\boldsymbol{K}_{k+1}$, $k \in \mathbb{N}_0$, from (23) unless a typical convergence criterion, e.g. $\|\boldsymbol{P}_{\overline{k}} - \boldsymbol{P}_{\overline{k}-1}\|_{\mathrm{F}} \leq \epsilon_{\mathrm{P}}$, $\epsilon_{\mathrm{P}} \in \mathbb{R}^{>0}$, is satisfied at $k = \overline{k}$. Under premise that $\overline{k}$ is sufficiently large, we have $\boldsymbol{K}_{\overline{k}+1} \approx \boldsymbol{K}^*$ based on Theorem 1.

For an arbitrary $0 \leq k \leq \overline{k}$, we exactly obtain $\boldsymbol{B} = \boldsymbol{P}_k^{-1}\boldsymbol{K}_{k+1}^{\mathsf{T}}\boldsymbol{R}$ and $\boldsymbol{E}_\mathrm{d}$ from 3rd argument of (23).

4) *One-Step Calculation:* For an arbitrary $0 \leq k \leq \overline{k}$, calculate $\boldsymbol{\Pi}_x^*$ and $\boldsymbol{\Pi}_\phi^*$ from (34), which *exactly* solve (13) based on Theorem 2, and $\boldsymbol{F}^* = -\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{\Pi}_\phi^*$.
5) *Implementation:* To achieve C2-4) on $[t_\mathrm{E}, \infty)$, cf. Section I, apply $\boldsymbol{u}^* \approx -\boldsymbol{K}_{\overline{k}+1}(\boldsymbol{x} - \boldsymbol{\Pi}_x^*\overline{\boldsymbol{x}}) + \boldsymbol{F}^*\overline{\boldsymbol{x}}$.

Remark 5: As in [8], [9], [12], [17], [18], [20] and [21], we have to assume that all measurements are *not* corrupted by *unknown* sensor noise. In general, it would be of interest to make these methods more robust in this aspect.

In the succeeding simulation example of an over-actuated system, i.e. rank $(\boldsymbol{B}) >$ rank $(\boldsymbol{C})$, we implement Algorithm 1 and emphasize our contribution C4) by two observations. First, if output regulation is feasible, our approach is suited for *almost exact tracking*, cf. Section I. Second, the additional actuators are used more efficiently to save input-energy than in [8], [9] whereas our calculation procedure is less complicated at the same time.

Example: We consider a non-minimum phase and over-actuated LTI-system with $(\boldsymbol{C}^{\mathsf{T}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{E}_\mathrm{d}, \boldsymbol{D}_\mathrm{d}) :=$

$$\left(\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 2 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{0} \right).$$

The exosystem $(\overline{\boldsymbol{C}}, \overline{\boldsymbol{A}})$ and initial value $\overline{\boldsymbol{x}}_0$ are given such that $\overline{\boldsymbol{x}}(t)^{\mathsf{T}} = \begin{bmatrix} t & 1 & \cos(t)/4 & \sin(t)/4 \end{bmatrix}$ and scalar reference $\overline{y}(t) = {}^{(t-5)}/_4$ (- -) result, i.e. we regard a polynomial reference and sinusoidal disturbances.

In the sequel, the data (4) of the dynamics is assumed to be **unknown**. Following Remark 2, we may obtain a stabilizing

$$\boldsymbol{K}_0 = \begin{bmatrix} 9.56 & 7.1 & 0.96 \\ 0.99 & 0.96 & 6.7 \end{bmatrix}.$$

Without knowing $\boldsymbol{A}$ and $\boldsymbol{B}$, diagonal weights such as $\widetilde{\boldsymbol{Q}} = \mathrm{diag}(10, 10, 5)$ and $\widetilde{\boldsymbol{R}} = \boldsymbol{I}_2$ seem appropriate with respect to LQRP 2. Here, the first two entries of $\widetilde{\boldsymbol{Q}}$ are equally chosen since both, 1st state x_1 and 2nd state x_2, contribute similarly to y based on the known $\boldsymbol{C}$. We cannot anticipate, however, that we penalized x_1 and its derivative $\dot{x}_1 = x_2$. This causes a counterbalance for the chosen $\widetilde{\boldsymbol{Q}}$ such that the real part of a closed-loop eigenvalue is bounded from the left by -1 for any $\boldsymbol{K}_k$, $k \geq 1$. In fact, this is even true when we scale $\widetilde{\boldsymbol{Q}}$ by an arbitrarily large positive scalar. Clearly, this indicates how useful our extension of the method by [12] can be as it guarantees in general that all closed-loop eigenvalues lie to the left of a defined $-\alpha$ independent of $\widetilde{\boldsymbol{Q}}$, $\widetilde{\boldsymbol{R}}$. In particular, we might know that $\alpha < 1.5$ satisfies the requirements in Lemma 3, e.g. by a study as in Remark 2. Indeed, we have $\max_{\lambda \in \sigma(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}_0)} \mathrm{Re}\,(\lambda) < -1.7$. By choosing $\alpha = 1$, all closed-loop eigenvalues lie certainly to the left of -1.

In view of tracking $\overline{y}(t)$ (- -), we consider $i = 1, 2$ different sets of weights: $\boldsymbol{Q}_1 = 200$, $\boldsymbol{R}_1 = {}^1/_{10} \cdot \boldsymbol{I}_2$ (—) and $\boldsymbol{Q}_2 = 0.5$, $\boldsymbol{R}_2 = \boldsymbol{R}_1$ (—). In the first case, we desire to track $\overline{y}(t)$ closely while in the second tracking errors are tolerated to save input energy. Since the data (4) is not available, for neither of the cases $i = 1, 2$ we are able to exactly calculate $\boldsymbol{u}_i^*(\cdot)$ as in Lemma 1. Clearly, this is in general severe considering 3) of Lemma 1. However, we may apply our developed Algorithm 1. The information required for *initialization* 1) has already been provided above. For 2), the *exploration* noise is defined as $\boldsymbol{e} = \sum_{\omega=2}^{6} \boldsymbol{r}_\omega \sin(\frac{\omega}{\mathrm{s}} t)$ with randomly chosen $\boldsymbol{r}_\omega \in \mathbb{R}^2$, $\boldsymbol{r}_\omega \neq \boldsymbol{0}$. This excitation by 5 distinct frequencies is *sufficiently rich*, cf. Remark 3. Hence, condition (24) with $\frac{1}{2}n(n+1)+nm+n\overline{n} = 24$ holds after $t_\mathrm{E} = 24 \cdot 0.3\,\mathrm{s} = 7.2\,\mathrm{s}$ of exploration for $t_{i+1} - t_i = 0.3\,\mathrm{s}$. The exploration phase (—) is illustrated in Fig. 1 for output y and the 3rd state x_3, where $\boldsymbol{x}_0^{\mathsf{T}} = \begin{bmatrix} 0 & 5 & 3 \end{bmatrix}$ and $\boldsymbol{x}(t_\mathrm{E})^{\mathsf{T}} = \begin{bmatrix} 0.089 & 0.092 & 0.232 \end{bmatrix}$.

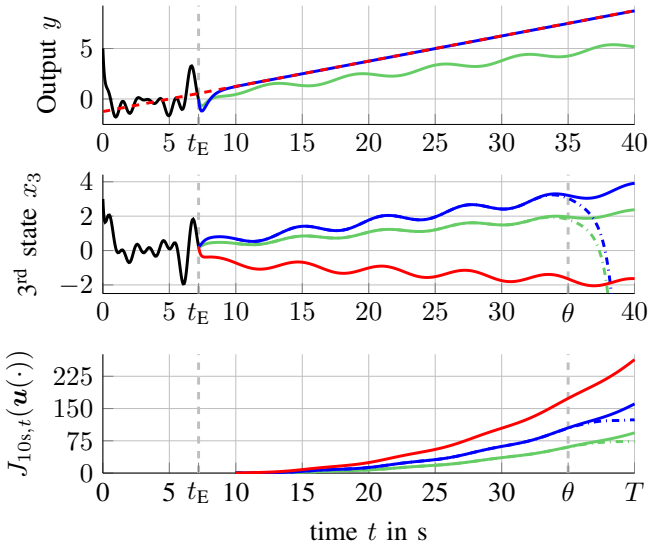

Fig. 1. Simulation results with exploration $[0, t_\mathrm{E}]$ (—) and application $[t_\mathrm{E}, T]$ for two different weights $\boldsymbol{Q}_1 = 200$ (—) and $\boldsymbol{Q}_2 = 0.5$ (—) for tracking of (- -). Comparison with each finite horizon $T = 40\,\mathrm{s}$ optimal solution (- - -), (- · -) and "optimal" output regulation (—) of [8], [9].

Subsequently, the *policy iteration* 3) is performed. Reminding Section I, [17] and [18] carry out this step for the augmented $n + \overline{n}$-dimensional system in order to obtain $\boldsymbol{u}_\mathrm{aug} = -\boldsymbol{R}^{-1}\begin{bmatrix} \boldsymbol{B}^{\mathsf{T}} & \boldsymbol{0} \end{bmatrix} \boldsymbol{P}_{\mathrm{aug},k} \begin{bmatrix} \boldsymbol{x}^{\mathsf{T}} & \overline{\boldsymbol{x}}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$. This also determines the tracking performance which obviously depends on the accuracy of $\boldsymbol{P}_{\mathrm{aug},k}$. In contrast, our tracking performance does not depend on the accuracy of $\boldsymbol{P}_k$. Hence, $\epsilon_\mathrm{P} = 0.05$ serves our needs for which we obtain $\|\boldsymbol{P}_{\overline{k}} - \boldsymbol{P}\|_\mathrm{F} = 0.003$ and $\|\boldsymbol{K}_{\overline{k}+1} - \boldsymbol{K}^*\|_\mathrm{F} = 0.0006$ after only $\overline{k} = 3$ steps. By 3) and our *one-step calculation* 4) for $i = 1, 2$, we calculate $\boldsymbol{B}$, $\boldsymbol{E}_\mathrm{d}$ and $\boldsymbol{\Pi}_{x,i}^*$, $\boldsymbol{\Pi}_{\phi,i}^*$. The Frobenius norm of each deviation from its true value is smaller than $3.6 \cdot 10^{-12}$ and, thus, lies in range of the numerical precision of a calculation by (13).

In summary, though not having knowledge of (4) we have found $\boldsymbol{u}^*(\cdot)$ such that the specifications of an optimal transition by LQRP 2 are approximatively met with a guaranteed degree of stability and these of stationarily optimal tracking by LQTP 1 with unbounded cost are exactly met. That is, the desired properties in Lemma 1 and 2 are achieved.

The phase of application $[t_\mathrm{E}, T]$ with $T = 40\,\mathrm{s}$ is displayed in Fig. 1. Obviously, $\boldsymbol{u}_2^*(\cdot)$ (—) with $\boldsymbol{Q}_2 = 0.5$ tracks the reference $\overline{y}$ (- -) less closely and attenuates disturbances

less than $\boldsymbol{u}_1^*(\cdot)$ (—) with $\boldsymbol{Q}_1 = 200$. Thus, $\boldsymbol{u}_2^*(\cdot)$ saves $58.6\,\%$ input-energy compared with $\boldsymbol{u}_1^*(\cdot)$. But on the other hand, $\boldsymbol{u}_1^*(\cdot)$ leads to almost exact tracking with stationary tracking error only about $1‰$. When the transition is over: $\widetilde{\boldsymbol{x}}(t) \approx \boldsymbol{0}$ for $t \geq 10\,\mathrm{s}$, we are interested in the stationary cost $J_{10\,\mathrm{s},t}(\boldsymbol{u}(\cdot))$. For comparison, we consider exact tracking by "optimal" output regulation (OR) given in [8], [9]. There, the stationary state $\boldsymbol{\Pi}_{x,\mathrm{OR}}\overline{\boldsymbol{x}}$ and control $\boldsymbol{u}_{\mathrm{OR}}(\cdot) = \boldsymbol{\Gamma}_{\mathrm{OR}}\overline{\boldsymbol{x}}$ (—) are the solution of

$$\min_{\boldsymbol{\Pi}_x,\,\boldsymbol{\Gamma}} \ \mathrm{trace}\left(\boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{R}_1\boldsymbol{\Gamma}\right)$$
$$\text{s.t.:} \quad \boldsymbol{\Pi}_x\overline{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{\Pi}_x + \boldsymbol{B}\boldsymbol{\Gamma} + \boldsymbol{E}_{\mathrm{d}} \ \text{ and } \ \boldsymbol{C}\boldsymbol{\Pi}_x + \boldsymbol{D}_{\mathrm{d}} = \overline{\boldsymbol{C}}.$$

This optimization problem (OP) was introduced by [14] in case that the solution of the regulator equations in [19], i.e. the constraints, is not unique due to over-actuation. It intends to provide an optimized solution in view of a cost functional such as (3). But since the OP is generally not equivalent it results in sub-optimal solutions as already indicated in [4].

As a consequence, while $\boldsymbol{u}_{\mathrm{OR}}(\cdot)$ (—) and $\boldsymbol{u}_1^*(\cdot)$ (—) lead to an equivalent tracking performance $\overline{y}^*(t) \approx \overline{y}_{\mathrm{OR}}(t)$, $\boldsymbol{u}_{\mathrm{OR}}(\cdot)$ requires stationarily $64\,\%$ more input energy than $\boldsymbol{u}_1^*(\cdot)$; note that cost $J_{10\,\mathrm{s},t}(\boldsymbol{u}(\cdot))$ equals the input-energy if $y(t) - \overline{y}(t) \approx 0$. Indeed, $\boldsymbol{u}_1^*(\cdot)$ outperforms $\boldsymbol{u}_{\mathrm{OR}}(\cdot)$ which is approved by the observations in [4]. This is reasoned as follows. Most of the entries in $\boldsymbol{\Pi}_x^* = \left[\pi_{x,i,j}^*\right]$ and $\boldsymbol{\Pi}_{x,\mathrm{OR}} = \left[\pi_{x,\mathrm{OR},i,j}\right]$ approximatively coincide. More precisely, we have indeed $\lim_{\boldsymbol{Q}_1\to\infty} \pi_{x,i,j}^* = \pi_{x,\mathrm{OR},i,j}$, besides $\pi_{x,3,1}^* \not\to \pi_{x,\mathrm{OR},3,1}$ and $\pi_{x,3,2}^* \not\to \pi_{x,\mathrm{OR},3,2}$. Hence, it results an essentially different stationary behavior of 3rd state x_3 for $\boldsymbol{u}_1^*(\cdot)$ and $\boldsymbol{u}_{\mathrm{OR}}(\cdot)$, see Fig. 1. Under the exact tracking constraint, this reflects the degrees of freedom available due to the over-actuation. These are clearly more efficiently exploited by $\boldsymbol{u}_1^*(\cdot)$. In addition, to obtain $\boldsymbol{u}_{\mathrm{OR}}(\cdot)$ in [8], [9], the optimization problem above has to be solved online whereas we only need to solve a single set of equations (30) in step 4) to derive $\boldsymbol{F}^*$.

In view of Lemma 2 with $T_1 = 10\,\mathrm{s}$, we compare the stationarily agreeable plans $\boldsymbol{u}_1^*(\cdot)$ (—) and $\boldsymbol{u}_2^*(\cdot)$ (—) with their corresponding finite horizon $[T_1, 40\,\mathrm{s}]$ optimal solutions $\boldsymbol{u}_{T,1}^*(\cdot)$ (- - -) and $\boldsymbol{u}_{T,2}^*(\cdot)$ (- - -), respectively. As we see from Fig. 1, we can choose $\theta = 35\,\mathrm{s}$ to guarantee a relative cost increase of less than $1\,\%$ for $i = 1, 2$ on $[10\,\mathrm{s}, 35\,\mathrm{s}]$. Switching from $\boldsymbol{u}_i^*(\cdot)$ to $\boldsymbol{u}_{T,i}^*(\cdot)$ on $[35\,\mathrm{s}, 40\,\mathrm{s}]$ would allow us to save input-energy. However, the unstable characteristics of x_3 for $\boldsymbol{u}_{T,i}^*(\cdot)$ on $[\theta, T]$ are unacceptable. To prevent this, one typically considers a quadratic end-cost for penalizing $|x_3(T)|$ in (3). But then, the mentioned performance gain of $\boldsymbol{u}_{T,i}^*(\cdot)$ on $[\theta, T]$ drops significantly and we may simply stick to $\boldsymbol{u}_i^*(\cdot)$ which is considerably easier to implement.

VI. CONCLUSION

For the first time, we derived a method to determine an LQ optimal tracking control for problems with unbounded cost without the need to know the system dynamics in advance. The control consists of two independent parts. A state feedback guarantees an optimal transition with specified degree of stability. It solves an LQR problem with individual weights and is obtained by extending the iterative ADP approach in [9], [12]. The second part is a static pre-filter resulting in optimal stationary tracking and is provided by a new one-step calculation. It adequately solves an infinite horizon LQT problem with generally unbounded cost and approximates finite horizon optimal solutions. For over-actuated systems, we showed that the new approach leads to a less complicated calculation and a more efficient operation in comparison with [8], [9].

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. Dover Publications, Inc., 2007.

[2] Z. Artstein and A. Leizarowitz, "Tracking periodic signals with the overtaking criterion," *IEEE Transactions on Automatic Control*, vol. 30, no. 11, pp. 1123–1126, 1985.

[3] S. Bernhard, "Time-invariant control in LQ optimal tracking: An alternative to output regulation," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 4912 – 4919, 2017, 20th IFAC World Congress.

[4] S. Bernhard and J. Adamy, "Static optimal decoupling control for linear over-actuated systems regarding time-varying references," in *2017 American Control Conf. (ACC)*, 2017, pp. 1049–1055.

[5] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. SIAM, 1994.

[6] D. A. Carlson, "Uniformly overtaking and weakly overtaking optimal solutions in infinite-horizon optimal control: When optimal solutions are agreeable," *Journal of Optimization Theory and Applications*, vol. 64, no. 1, pp. 55–69, 1990.

[7] D. A. Carlson and A. Haurie, *Infinite Horizon Optimal Control*. Springer-Verlag Berlin Heidelberg, 1987.

[8] W. Gao and Z.-P. Jiang, "Linear optimal tracking control: An adaptive dynamic programming approach," in *2015 American Control Conference (ACC)*, 2015, pp. 4929–4934.

[9] ——, "Adaptive dynamic programming approach and adaptive optimal output regulation of linear systems," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 4164–4169, 2016.

[10] H. Halkin, "Necessary conditions for optimal control problems with infinite horizons," *Econometrica*, vol. 42, no. 2, pp. 267–272, 1974.

[11] P. Ioannou and J. Sun, *Robust Adaptive Control*. Prentice Hall, 1996.

[12] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[13] D. L. Kleinman, "On an iterative tecnique for Riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114–115, 1968.

[14] A. J. Krener, "The construction of optimal linear and nonlinear regulators," *Systems, Models and Feedback: Theory and Applications*, vol. 12, pp. 301–322, 1992.

[15] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Control Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[16] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control," *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 76–105, 2012.

[17] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 3051–3056, 2014.

[18] C. Qin, H. Zhang, and Y. Luo, "Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming," *International Journal of Control*, vol. 87, no. 5, pp. 1000–1009, 2014.

[19] H. Trentelman, A. A. Stoorvogel, and M. Hautus, *Control Theory for Linear Systems*. Springer-Verlag London, 2001.

[20] K. G. Vamvoudakis, "Optimal trajectory output tracking control with a Q-learning algorithm," in *2016 American Control Conference (ACC)*, 2016, pp. 5752–5757.

[21] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear-systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.