

Global Convergence of Policy Gradient Algorithms for Indefinite Least Squares Stationary Optimal Control

Jingjing Bu and Mehran Mesbahi

Abstract—We consider policy gradient algorithms for the indefinite least squares stationary optimal control, e.g., linear-quadratic-regulator (LQR) with indefinite state and input penalization matrices. Such a setup has important applications in control design with conflicting objectives, such as linear quadratic dynamic games. We show the global convergence of gradient, natural gradient and quasi-Newton policies for this class of indefinite least squares problems.

I. INTRODUCTION

Least squares stationary optimal control provides an effective synthesis procedure for linear control systems since Kalman’s original work in the 1960s [1]. This setting was later extended beyond positive semidefinite cost structure by Willems [2]. It is known that similar to standard LQR, this setup can be examined using the Algebraic Riccati Equation (ARE); DARE refers to the discrete analogue of this matrix equation. Historically, a large number of works have studied the solution of ARE and DARE, including approaches based on iterative algorithms [3],¹ algebraic solution methods [4], and semidefinite programming [5].

Although the cost function plays a fundamental role in the least squares optimal control, it is generally not “recommended” to *directly* compute the optimal gain (policy) using this cost without solving the associated Riccati equation.² This approach, in the meantime, is in sharp contrast to how one would typically go about minimizing a cost function over the variable of interest in introductory optimization, say, through gradient descent.³ Recently, there has been a surge of interest in constructing optimal control strategies directly, viewing control synthesis through the lens of first order methods.⁴ Adopting such a point of view has been partially inspired by the application of learning algorithms in control, such as Reinforcement Learning (RL), where using principles of (approximate) dynamic programming, one can devise real-time model-free methods for both continuous-time and discrete-time optimal control problems [6]–[13]. The RL perspective not only provides more insights into the synthesis problem, but also can be extended to model-free settings

by means of stochastic (zeroth-order) optimization [14], [15]. However, policy iteration is inherently prohibitive for an infinite horizon cost structure that is undiscounted and unbounded per stage [13].

The main contribution of this note is to extend policy based algorithms beyond positive (semi)definite cost structures considered in [16], [17]. More specifically, we show that under mild assumptions, even when the state and cost penalization matrices are indefinite in the least squares optimal control, gradient policy (respectively, natural gradient and quasi-Newton policies) converges to the global minimizer at a linear (respectively, linearly and Q -quadratic) rate. Along the way, we devise a distinct approach for arguing the stability of the iterative process as compared with those adopted in previous works.⁵

The note is organized as follows. In §II, we introduce the notation and preliminaries. §III is devoted to the LQR setup, analytical properties of the cost function, a “mild” assumption, and its implications. In §IV, we derive the corresponding stepsizes for gradient descent (GD), natural gradient descent (NGD), and quasi-Newton (QN) iterations; we then show the global linear (respectively, linear and Q -quadratic) convergence of gradient policy (respectively, natural gradient policy and quasi-Newton policy) under the proposed stepsizes. A numerical example is provided in §V. The note is concluded in §VI.

II. NOTATION AND PRELIMINARIES

We denote by $\mathbb{M}_{n \times m}(\mathbb{R})$ the set of $n \times m$ real matrices. \mathbb{R}^n denotes the n -dimensional real Euclidean space; when $n = 1$, this set is identified with the set of real numbers. Other notation include A^\top , $\rho(A)$, $\text{Tr}(A)$, representing the transpose, spectral radius, and trace of the matrix A , respectively. The real inner product between a pair of vectors x and y is denoted by $\langle x, y \rangle$. $\|A\|_2$ denotes the spectral (operator) norm of a square matrix A and $\|A\|_F$ denotes its Frobenius norm.⁶ Lastly, the notation $A \geq B$ for two symmetric matrices refers to the positive semi-definiteness of their difference $A - B$; analogously for positive definiteness of this difference we use $A > B$. We let $\lambda_i(A)$ denote the eigenvalues of a square matrix A . These eigenvalues are indexed in an increasing order with respect to their real parts, i.e.,

$$\text{Re}(\lambda_1(A)) \leq \dots \leq \text{Re}(\lambda_n(A)).$$

⁵The proposed technique also provides an alternative way to argue stability properties of the iterative process under standard LQR assumptions.

⁶2-norm is assumed when we use $\|\cdot\|$.

Submitted to IEEE Control Systems Letters. The research of the authors has been supported by AFOSR Grant No. FA9550-16-1-0022 and DARPA Lagrange Grant FA8650-18-2-7836.

The authors are with the University of Washington, Seattle, WA 98195, USA; Emails: bujing+mesbahi@uw.edu

¹In Hwer’s original work, Q and R are positive definite. However, the algorithm still converges even for the indefinite cost structure [4].

²In this note, feedback gain, feedback control and feedback policy are used interchangeably.

³This is essentially due to the dynamic nature of the constraint set.

⁴One might as well extrapolate that these methods provide a streamline recipe for learning optimal feedback gains in real-time.

If A is symmetric, the ordering becomes $\lambda_1(A) \leq \dots \leq \lambda_n(A)$. When $A \geq 0$, $\|A\| = \lambda_n(A)$ and we shall use these interchangeably. We use $C^\omega(U)$ to denote the set of real analytic functions over an open set $U \subseteq \mathbb{R}^n$. A square matrix $A \in \mathbb{M}_{n \times n}(\mathbb{R})$ is *Schur* if $\rho(A) < 1$. A pair (A, B) is stabilizable if there exists some K for which $A - BK$ is Schur. Given a pair of system matrices (A, B) , \mathcal{S} denotes the set of Schur stabilizing feedback gains,

$$\mathcal{S} = \{K \in \mathbb{M}_{m \times n}(\mathbb{R}) : \rho(A - BK) < 1\}.$$

For the pair (A, B) , we say that K is stabilizing if $A - BK$ is Schur; it is *marginally stabilizing* or *almost stabilizing* if $\rho(A - BK) = 1$. An eigenvalue λ of $A \in \mathbb{M}_{n \times n}(\mathbb{R})$ is called (C, A) -observable if

$$\text{rank} \left(\begin{pmatrix} A - \lambda I \\ C \end{pmatrix} \right) = n,$$

for a given $C \in \mathbb{M}_{p \times n}(\mathbb{R})$; p is the dimension of the output of a linear system.

III. PROBLEM SETUP

In the standard least squares (stationary) optimal control, we consider a (discrete-time) linear time invariant model of the form,

$$(1) \quad x_{k+1} = Ax_k + Bu_k,$$

where $A \in \mathbb{M}_{n \times n}(\mathbb{R})$, $B \in \mathbb{M}_{n \times m}(\mathbb{R})$ and (A, B) is stabilizable. The corresponding LQR problem is the optimization problem of devising a linear feedback gain $K \in \mathbb{M}_{m \times n}(\mathbb{R})$ for which $u_k = -Kx_k$, minimizing,⁷

$$J(x_0) = \sum_{k=0}^{\infty} [\langle x_k, Qx_k \rangle + \langle u_k, Ru_k \rangle],$$

where x_0 is the initial condition, and the quadratic cost is parameterized by $Q = Q^\top$ and $R = R^\top$; note that we *do not* require positive (semi-)definiteness of Q and R . Such a generalization is not only of theoretical interest but also has important applications in network synthesis and stability theory [2]. In order to update the feedback gain (policy) directly, it will conceptually be appealing to consider the cost as a matrix function over the set of feedback gains. With this aim in mind, we may define $J_{x_0}: \mathbb{M}_{m \times n}(\mathbb{R}) \rightarrow \mathbb{R}$ as,

$$(2) \quad J_{x_0}(K) = \sum_{j=0}^{\infty} [\langle (A - BK)^j x_0, (Q + K^\top RK)(A - BK)^j x_0 \rangle],$$

for some fixed initial condition $x_0 \in \mathbb{R}^n$. Note that the cost function J is a function of the policy K and initial condition x_0 . Since we are interested in *optimal policy* independent of initial conditions, naturally, we should reformulate the cost function to reflect this independence. Indeed, this point has been discussed in [17] where it is argued that such a formulation is necessary for the cost function to be well defined (see details in §III [17]). The independence with respect to the

⁷The condition that u_k has the form $-Kx_k$ is not set a priori in the LQR formulation; this feedback form is typically shown via the adoption of a dynamic programming step.

initial condition can be achieved by either sampling x_0 from a distribution with full-rank covariance [16], or choosing a spanning set $\{z_1, \dots, z_n\} \subseteq \mathbb{R}^n$ [17], and defining the value function over \mathcal{S} as,

$$(3) \quad f(K) = \sum_{i=1}^n J_{z_i}(K),$$

where $J_{z_i}(K)$ is the cost by choosing initial state x_0 as z_i and letting $u_k = Kx_k$. Note that over the set \mathcal{S} , f admits a compact form $f(K) = \text{Tr}(X\Sigma)$, where $\Sigma = \sum_{i=1}^n z_i z_i^\top$ and X is the solution to the Lyapunov equation,

$$(4) \quad (A - BK)^\top X(A - BK) + Q + K^\top RK = X.$$

How the cost function f behaves near the boundary $\partial\mathcal{S}$ is of paramount importance in the design of iterative algorithms for least squares optimal control problems. In the standard setting, the cost function diverges to $+\infty$ when the feedback gain approaches the boundary of this set (see [17] for details). In fact, this property guarantees stability of the obtained solution via first order iterative algorithms for the suitable choice of stepsize. However, the behavior of f on the boundary $\partial\mathcal{S}$ could be more intricate. For example, if $K \in \partial\mathcal{S}$, i.e., $\rho(A - BK) = 1$, then it is possible that the cost is still finite. This happens when an eigenvalue of $A - BK$ on the unit disk in the complex plane is not $(Q + K^\top RK, A - BK)$ -observable. To see this, we note that for every ω_i , the series

$$J_{\omega_i}(K) = \omega_i^\top \left(\sum_{j=0}^{\infty} ((A - BK)^\top)^j (Q + K^\top RK)(A - BK)^j \right) \omega_i.$$

is convergent to a finite (real) number if the marginally stable modes are not detectable. Even on $\bar{\mathcal{S}}^c$ (complement of closure of \mathcal{S}), f could be finite if all non-stable modes of $A - BK$ are not $(Q + K^\top RK, A - BK)$ -observable. The complication suggests that the function value is no longer a valid indicator of stability. We remark that such a situation does not occur in the LQ setting examined in [16], [17], as it has been assumed that Q is positive definite.

A. Analytical properties of the indefinite cost function

In this section, we collect some useful analytic characterizations of $f(K)$. To simplify the notation, in the rest of this paper, we set,

$$A_K := A - BK, \quad \text{and} \quad \mathbf{N}_K := RK - B^\top X(A - BK);$$

when the context is clear, we will write \mathbf{N} instead of \mathbf{N}_K ; in describing the iterative process on the gain matrix (when K is updated), we shall denote \mathbf{N}_{K_j} as \mathbf{N}_j .

Proposition III.1. *The indefinite least squares optimal control problem (3) on the set of stabilizing feedback gains has the following properties:*

- a) *The set \mathcal{S} is regular open, contractible, and unbounded when $m \geq 2$ and the boundary $\partial\mathcal{S}$ is precisely the set $\mathcal{B} = \{K \in \mathbb{M}_{m \times n}(\mathbb{R}) : \rho(A - BK) = 1\}$.*
- b) *For the cost (3), one has $f \in C^\omega(\mathcal{S})$.*
- c) *The gradient of f (3) is given by*

$$\nabla f(K) = 2(RK - B^\top X A_K) Y_K,$$

where Y_K solves the Lyapunov matrix equation,

$$(5) \quad A_K Y A_K^\top + \Sigma = Y.$$

d) Let $K, \tilde{K} \in \bar{\mathcal{S}}$ ⁸; suppose that the corresponding Lyapunov matrix equations (4) have symmetric solutions X and \tilde{X} , respectively.⁹ Namely,

$$\begin{aligned} A_K^\top X A_K + Q + K^\top R K &= X, \\ A_{\tilde{K}}^\top \tilde{X} A_{\tilde{K}} + Q + \tilde{K}^\top R \tilde{K} &= \tilde{X}. \end{aligned}$$

Then we have

$$\begin{aligned} A_K^\top (X - \tilde{X}) A_{\tilde{K}} + (K - \tilde{K})^\top \mathbf{N}_K + \mathbf{N}_K^\top (K - \tilde{K}) \\ - (K - \tilde{K})^\top (R + B^\top X B) (K - \tilde{K}) &= X - \tilde{X}. \end{aligned}$$

e) Suppose that $K_* \in \arg \min_{K \in \mathcal{S}} f(K)$. Then

$$\tau_1 \|K - K_*\|_F^2 \leq f(K) - f(K_*) \leq \tau_2 \langle \mathbf{N}_K, \mathbf{N}_K \rangle,$$

where

$$\tau_1 = \lambda_1(Y) \lambda_1(R + B^\top X B), \quad \tau_2 = \frac{\lambda_n(Y_*)}{\lambda_1(R + B^\top X B)},$$

and Y_* solves the Lyapunov equation (5) with K_* .

The proofs of these results can be found in [17]. We emphasize that (e) holds only if $\arg \min_{K \in \mathcal{S}} f(K) \neq \emptyset$, namely, there exists $K_* \in \mathcal{S}$ such that $f(K) \geq f(K_*)$ for every $K \in \mathcal{S}$. In the next subsection, we shall elaborate on a “mild” assumption to ensure that this condition holds.

B. A key assumption and its consequences

Throughout the manuscript, we have the following standing assumption.

Assumption 1. *There exists a strict local minimizer of $f(K)$ over \mathcal{S} . In other words, there exists some $K_* \in \mathcal{S}$ and an open neighborhood $B_\delta(K_*) = \{K : \|K - K_*\|_F < \delta\}$, such that $f(K_*) < f(K)$ for every $K \in B_\delta(K_*) \cap \mathcal{S}$.*

Remark III.2. *The seminal work of Willems [2] explores many facets of the least squares optimal control with indefinite Q and R ,¹⁰ in particular, this work examines conditions for which the above assumption holds. We will not discuss these conditions and instead refer the reader to [2] and references therein.*

We observe several implications of this assumption.

Proposition III.3. *Suppose that K_* is the strict local minimizer of $f(K)$ over \mathcal{S} and X_* is the corresponding value matrix. Then,*

- a) $X_* = X_*^\top$,
- b) $R + B^\top X_* B > 0$,
- c) X_* solves the DARE (6),

$$(6) \quad X = A^\top X A + Q - A^\top X B (R + B^\top X B)^{-1} B^\top X A,$$

⁸ $\bar{\mathcal{S}}$ is the closure of \mathcal{S} .

⁹Note that the assumption clearly holds if $K, \tilde{K} \in \mathcal{S}$. It will also hold if $K \in \partial \mathcal{S}$ and the eigenvalues of $A - BK$ on the unit disk are not $(Q + K^\top R K, A - BK)$ -observable.

¹⁰An our adopted terminology is in his honor.

d) The minimizer K_* is the unique global minimizer,

e) X_* is the maximal solution to DARE (6) and is unique among all almost stabilizing solutions of (6).

Proof. Part (a) follows from having X_* solve the Lyapunov matrix equation (4) with $K = K_*$ and the fact that $Q + K^\top R K$ is symmetric. For parts (b) and (c), we first note that if K_* is a strict local minimizer in \mathcal{S} , since $f \in C^\omega(\mathcal{S})$, first-order and second-order optimality conditions imply $\nabla f(K_*) = 0$ and $\nabla^2 f(K_*) > 0$. By the Hessian formula in [17], we have $R + B^\top X_* B > 0$, i.e., (b) holds. Further, since $\nabla f(K_*) = \mathbf{N}_{K_*} Y_{K_*}$ and $Y_{K_*} > 0$, it follows that $\mathbf{N}_{K_*} = 0$. Namely, $R K_* - B^\top X_* A_{K_*} = 0$. Substituting $K_* = (R + B^\top X_* B)^{-1} B^\top X_* A$ into the Lyapunov equation (4), we have that X_* solves the DARE (6). For part (d), it suffices to observe that K_* is the unique stationary point. To this end, suppose that there exist $K_{*,1}$ and $K_{*,2}$ such that the gradient vanishes at both points, namely $\mathbf{N}_{K_{*,1}} = \mathbf{N}_{K_{*,2}} = 0$ ¹¹. By part (d) in Proposition III.1, we have

$$\begin{aligned} X_{*,1} - X_{*,2} &= A_{K_{*,2}}^\top (X_{*,1} - X_{*,2}) A_{K_{*,2}} \\ &\quad - (K_{*,1} - K_{*,2})^\top (R + B^\top X_{*,1} B) (K_{*,1} - K_{*,2}). \end{aligned}$$

As $A_{K_{*,2}}$ is Schur, it follows that $X_{*,1} \geq X_{*,2}$ and similarly $X_{*,2} \geq X_{*,1}$. Hence, the stationary point is unique. Part (e) follows from standard DARE theory (see Chapters 12 and 13 in [4] for details.) \square

IV. GLOBAL CONVERGENCE OF POLICY GRADIENT ALGORITHMS

In this section, we show the global convergence of gradient descent (GD), natural gradient descent (NGD), and quasi-Newton (QN) iterations for indefinite least squares optimal control. In particular, under Assumption 1, it is shown that gradient descent (respectively, natural gradient descent and quasi-Newton) converges to the maximal solution of the DARE at a linear (respectively, linear and quadratic) rate. In this direction, first recall that the gradient, natural gradient and quasi-Newton directions [17] are given by,

$$\begin{aligned} \mathbf{g}(K) &:= 2(RK - B^\top X A_K)Y, \\ \mathbf{n}(K) &:= 2(RK - B^\top X A_K), \\ \mathbf{qn}(K) &:= 2(R + B^\top X B)^{-1}(RK - B^\top X A_K); \end{aligned}$$

GD, NGD and QN now refer to following update rules:

$$\begin{aligned} (7) \quad \text{GD} : \quad & K_{j+1} = K_j - \eta_j \mathbf{g}(K_j), \\ (8) \quad \text{NGD} : \quad & K_{j+1} = K_j - \eta_j \mathbf{n}(K_j), \\ (9) \quad \text{QN} : \quad & K_{j+1} = K_j - \eta_j \mathbf{qn}(K_j), \end{aligned}$$

where η_j 's are stepsizes to be determined. We provide the convergence analysis for the case of natural gradient descent.

Theorem IV.1 (Natural Gradient Analysis). *Consider the iterates $\{K_j\}$ generated by NGD (8), with stepsize $\eta_j = 1/(2\lambda_n(R + B^\top X_j B))$, where $\{X_j\}$ solve the corresponding Lyapunov equations (4). Then both the function values and*

¹¹This follows from $Y_K > 0$ for every $K \in \mathcal{S}$.

gain iterates converge to their corresponding global minima at a linear rate. That is,

$$\begin{aligned} f(K_j) - f(K_*) &\leq q_1^j (f(K_0) - f(K_*)), \\ \|K_j - K_*\|_F^2 &\leq c_1 q_1^j \|K_0 - K_*\|_F^2, \end{aligned}$$

for some $q_1 \in (0, 1)$ and $c_1 > 0$.

Proof. The analysis provided in [17] for the one-step progression of NGD holds here; thus the convergence rate would remain the same if we can prove that the iterates remain stabilizing.

By induction, it suffices to argue that with the chosen stepsize, K_j is stabilizing provided that K_{j-1} is. Consider the ray $\{K_t = K_{j-1} - t\mathbf{n}(K_{j-1}) : t \geq 0\}$. Note that by openness of \mathcal{S} and continuity of eigenvalues, there is a maximal interval $[0, \zeta]$ ¹² such that $K_{j-1} + t\mathbf{n}(K_{j-1})$ is stabilizing for $t \in [0, \zeta)$ and $K_{j-1} + \zeta\mathbf{n}(K_{j-1})$ is marginally stabilizing. Now suppose that $\zeta \leq 1/(2\lambda_n(R_1 + B_1^\top X_{i-1} B_1))$; take a sequence $t_l \in [0, \zeta)$ such that $t_l \rightarrow \zeta$. Consider the sequence of value matrices $\{X_{t_l}\}$ and denote by \mathcal{L} as the set of all limit points of $\{X_{t_l}\}$. Observe that $X_* \leq X_{t_l} \leq X_{j-1}$. By Bolzano-Weierstrass [18], \mathcal{L} is nonempty.¹³ By continuity, any $Z \in \mathcal{L}$ solves,

$$Z = (A - BK_\zeta)^\top Z (A - BK_\zeta) + Q + K_\zeta^\top R K_\zeta.$$

Now by part (d) in Proposition III.1, we have

$$\begin{aligned} Z - X_* &= (A - BK_\zeta)^\top (Z - X_*) (A - BK_\zeta) \\ &\quad + (K_\zeta - K_*)^\top (R + B^\top X_* B) (K_\zeta - K_*). \end{aligned}$$

Suppose that (λ, v) is the eigenvalue-eigenvector pair of $A - BK_\zeta$ such that $(A - BK_\zeta)v = \lambda v$ and $|\lambda| = 1$. Then it follows that,

$$\begin{aligned} v^\top (Z - X_*) v &= v^\top (A - BK_\zeta)^\top (Z - X_*) (A - BK_\zeta) v \\ &\quad + v^\top (K_\zeta - K_*)^\top (R + B^\top X_* B) (K_\zeta - K_*) v. \end{aligned}$$

Thereby $(K_\zeta - K_*)v = 0$ and $K_\zeta v = K_* v$. Consequently, $(A - BK_*)v = (A - BK_\zeta)v$. But this is a contradiction to the assumption that K_* is a stabilizing solution.

Hence $\{X_j\}$ is a monotonically non-increasing sequence bounded below by X_* . As such, the sequence of iterates $\{K_j\}$ and the sequence of function values $\{f(K_j)\}$ converge linearly to K_* and $f(K_*)$ following the arguments in [17]. \square

We mention that the above stability argument can be applied for the sequence generated by the quasi-Newton iteration as well. The quadratic convergence rate for such a sequence would then follow from the proof in [17].

Theorem IV.2 (Quasi-Newton Analysis). *Suppose Assumption 1 holds. Consider the iterates $\{K_j\}$ generated by QN (9) with stepsize $\eta_j = 1/2$. Then both the function values and*

¹²We suppose ζ is finite; if ζ is infinite, there is nothing needed to be shown.

¹³Note that it is not guaranteed that X_{t_j} is convergent. The limit points are also not necessarily well-ordered in the ordering induced by the p.s.d. cone.

iterates converge to their respective global minima at a Q -quadratic rate. That is,

$$\begin{aligned} f(K_j) - f(K_*) &\leq q_2 (f(K_{j-1}) - f(K_*))^2, \\ \|K_j - K_*\|_F^2 &\leq c_2 q_2 \|K_{j-1} - K_*\|_F^4, \end{aligned}$$

for some $q_2 > 0$ and $c_2 > 0$.

The gradient policy analysis requires more work since the stepsize developed in [17] involves the smallest eigenvalue $\lambda_1(Q)$. However by carefully replacing “ $\lambda_1(Q)$ -related quantities” in [17], one can still prove the global linear convergence rate as follows.

Theorem IV.3 (Gradient Analysis). *Suppose Assumption 1 holds. Consider the iterate $\{K_j\}$ generated by GD (7) with stepsize η_j specified in Theorem A.3. Then both the function values and iterates converge to their respective global minima at a linear rate. That is,*

$$\begin{aligned} f(K_j) - f(K_*) &\leq q_3^j (f(K_0) - f(K_*)), \\ \|K_j - K_*\|_F^2 &\leq c_3 q_3^j \|K_0 - K_*\|_F^2, \end{aligned}$$

for some $q_3 \in (0, 1)$ and $c_3 > 0$.

In [17], the compactness of sublevel sets have been used to devise the stepsize rule to guarantee a sufficient decrease in the cost and stability of the iterates. The proof of compactness in [17] however, relies on the positive definiteness of Q and R .¹⁴ But, we can show that a perturbation bound can be employed to derive a suitable constant stepsize for the indefinite cost structure as well. The details of this observation are deferred to the Appendix A.

V. A NUMERICAL EXAMPLE

In this section, we show the proposed convergence results by a numerical example. The system parameters are $A = 0.5I$, $B = I$, $R = I$ and

$$Q = \begin{pmatrix} 1.62370842 & 0.36712592 & -1.31209102 & 1.97803823 & -0.49297266 \\ 0.36712592 & 2.21878741 & 0.47525552 & -1.07142839 & 1.04343275 \\ -1.31209102 & 0.47525552 & 1.90887732 & -0.83057818 & 0.3818043 \\ 1.97803823 & -1.07142839 & -0.83057818 & 0.93847322 & -0.90779531 \\ -0.49297266 & 1.04343275 & 0.3818043 & -0.90779531 & -1.06295748 \end{pmatrix}.$$

Note that Q is indefinite and its (rounded) eigenvalues are 4.75, 2.55, 0.96, -1.1, -1.53. Figures 1-2 show the global linear convergence of the gradient policy update. The global linear convergence of natural gradient policy are demonstrated in Figures 3-4. Figures 5-6 show the Q -quadratic convergence for the quasi-Newton policy update.

VI. CONCLUDING REMARKS

This note considers policy gradient algorithms for the indefinite least squares stationary optimal control, e.g., indefinite LQR. We show the global linear (respectively, linear and Q -quadratic) convergence of gradient policy (respectively, natural gradient and quasi-Newton policies.) Although these results are presented assuming the knowledge of the system matrices, gradient and natural gradient policies can be extended to model-free case by means of stochastic (zeroth

¹⁴Or the observability of (Q, A) .

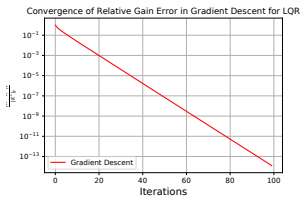


Fig. 1: Convergence of the relative error for the feedback gain under gradient descent iteration

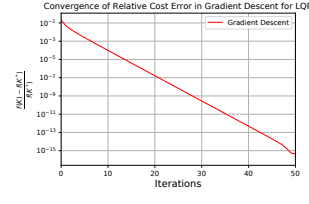


Fig. 2: Convergence of the relative error for indefinite LQR cost under gradient descent iteration

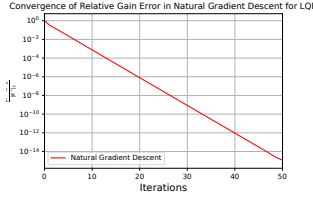


Fig. 3: Convergence of the relative error for the feedback gain under natural gradient descent iteration

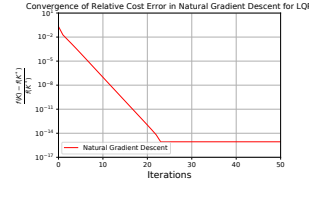


Fig. 4: Convergence of the relative error for indefinite LQR cost under natural gradient descent iteration

order) optimization (see [16] for details). As such, this note extends the results reported in [16], [17] for indefinite LQR. These extensions have important implications for optimal control, stability analysis and LQ games. Indeed, some of these observations have been utilized to show global convergence of sequential policy updates in LQ dynamic games [19].

ACKNOWLEDGEMENTS

The authors thank Henk J. van Waarde for many helpful discussions.

REFERENCES

- [1] R. E. Kalman *et al.*, “Contributions to the theory of optimal control,” *Bol. soc. mat. mexicana*, vol. 5, no. 2, pp. 102–119, 1960.
- [2] J. Willems, “Least squares stationary optimal control and the algebraic riccati equation,” *IEEE Transactions on Automatic Control*, vol. 16, no. 6, pp. 621–634, 1971.
- [3] G. Hewer, “An iterative technique for the computation of the steady state gains for the discrete optimal regulator,” *IEEE Transactions on Automatic Control*, vol. 16, no. 4, pp. 382–384, 1971.
- [4] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*. New York, NY: Oxford University Press, 1995.
- [5] V. Balakrishnan and L. Vandenberghe, “Semidefinite programming duality and linear time-invariant systems,” *IEEE Transactions on Automatic Control*, vol. 48, no. 1, pp. 30–41, 2003.
- [6] 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.
- [7] J. Y. Lee, J. B. Park, and Y. H. Choi, “Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems,” *Automatica*, vol. 48, no. 11, pp. 2850–2859, 2012.
- [8] D. Lee and J. Hu, “Primal-dual Q-learning framework for LQR design,” *IEEE Transactions on Automatic Control*, pp. 1–1, 2018.
- [9] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, “Adaptive linear quadratic control using policy iteration,” in *Proceedings of 1994 American Control Conference*, vol. 3, 1994, pp. 3475–3479.
- [10] F. L. Lewis and D. Vrabie, “Reinforcement learning and adaptive dynamic programming for feedback control,” *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.

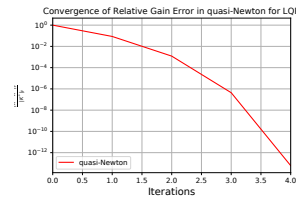


Fig. 5: Convergence of the relative error for the feedback gain under quasi-Newton iteration with constant stepsize 1/2

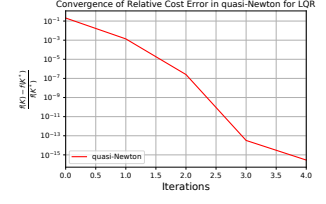


Fig. 6: Convergence of the relative error for indefinite LQR cost under quasi-Newton iteration with constant stepsize 1/2

- [11] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, “Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers,” *IEEE Control Systems*, vol. 32, no. 6, pp. 76–105, 2012.
- [12] T. Y. Chun, J. Y. Lee, J. B. Park, and Y. H. Choi, “Stability and monotone convergence of generalised policy iteration for discrete-time linear quadratic regulations,” *International Journal of Control*, vol. 89, no. 3, pp. 437–450, 2016.
- [13] D. Bertsekas, *Dynamic Programming and Optimal Control*, 4th ed. Athena Scientific, 2012, vol. II.
- [14] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to derivative-free optimization*. Siam, 2009, vol. 8.
- [15] “Robust stochastic approximation approach to stochastic programming.”
- [16] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, “Global convergence of policy gradient methods for the linear quadratic regulator,” in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1467–1476.
- [17] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, “LQR through the lens of first order methods: Discrete-time case,” *arXiv preprint arXiv:1907.08921*, 2019.
- [18] W. Rudin, *Principles of Mathematical Analysis*, ser. International series in pure and applied mathematics.
- [19] J. Bu, L. J. Ratliff, and M. Mesbahi, “Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games,” *arXiv preprint arXiv:1911.04672*, 2019.

APPENDIX

A. Gradient Policy Analysis for Nonstandard LQR

This section is devoted to the proof of Theorem IV.3. As it was pointed out previously, the strategy adopted in [16], [17] are no longer viable for an indefinite cost structure. However, as we will show, a perturbation bound would circumvent this issue and allows deriving the required stepsize, guaranteeing a decrease in function values while ensuring stabilization.

In the following, we shall drop all the subscripts as the stepsize will be valid for every iterate. Suppose now that we have a stabilizing policy K and the gradient direction is given by $\mathbf{g}(K) = 2NY$.¹⁵ The main object that we work with in this section is the ray starting at K along the gradient direction,

$$\{K_\eta : K - \eta \mathbf{g}(K), \eta \geq 0\}.$$

We shall further denote $A_\eta = A - BK_\eta = A - B(K - \eta 2NY)$.

Here is an outline of our proof strategy:

- a) By the openness of \mathcal{S} and continuity of eigenvalues, there exists a maximal interval $[0, c)$ such that K_η is stabilizing for every $\eta < c$ and K_c is marginally stabilizing; such a c could be either finite or infinite.

¹⁵Note the subscripts are dropped; N and Y are both dependent on K

b) Now suppose that c above is known. Then for every $\eta < c$, $f(K_\eta)$ is well-defined and we can compute the difference,

$$f(K) - f(K_\eta) = 4\eta \text{Tr}(\mathbf{N}^\top \mathbf{N}(Y Y_\eta - \eta a Y Y_\eta Y)),$$

where $a = \lambda_n(R + B^\top X B)$, and Y_η solves the Lyapunov matrix equation,

$$(A - BK_\eta)Y_\eta(A - BK_\eta)^\top + \Sigma = Y_\eta.$$

c) Next we define a univariate function $\phi : [0, c) \rightarrow \mathbb{R}$ by,

$$\phi(\eta) = \text{Tr}(\mathbf{N}^\top \mathbf{N}(Y Y_\eta - \eta a Y Y_\eta Y)).$$

Note that $\phi(0) > 0$ if the gradient does not vanish at K . Now our goal is to characterize a step size $0 < \eta' < c$ such that $\phi(\eta') > 0$.

It is clear that the knowledge of c and characterizing η' above are crucial for stepsize analysis. We shall demonstrate that characterizing η' will suffice to provide a stepsize; the quadratic cost structure will implicitly enforce stabilization.

To begin, we observe a perturbation bound on Y_η , assuming that K_η is stabilizing.

Proposition A.1. Put $\mu_1 = \|Y\|_2 \|B_1 \mathbf{N} Y\|_2^2 / \lambda_1(\Sigma)$ and $\mu_2 = \|Y\|_2 \|B_1 \mathbf{N} Y\|_2 \|A_K\|_2 / \lambda_1(\Sigma)$, and let

$$\eta_0 = \frac{\sqrt{\mu_1 + \mu_2^2}}{4\mu_1} - \frac{\mu_2}{4\mu_1};$$

suppose that A_η is Schur stable for every $\eta \leq \eta_0$. Then for all $\eta \leq \eta_0$,

$$\|Y_\eta\|_2 \leq \beta_0 \|Y\|_2,$$

where $\beta_0 = 1/(1 - 4\mu_1\eta_0^2 - 4\mu_2\eta_0) > 0$.

Proof. Taking the difference of the corresponding Lyapunov equations, we have

$$\begin{aligned} Y_\eta - Y - A_K(Y_\eta - Y)A_K^\top &= 2\eta(A_K Y_\eta(B_1 \mathbf{N} Y)^\top + B \mathbf{N} Y Y_\eta A_K^\top) \\ &\quad + 4\eta^2 B \mathbf{N} Y Y_\eta (B \mathbf{N} Y)^\top \\ &\leq \|Y_\eta\|_2 (4\eta \|B \mathbf{N} Y\|_2 \|A_K\|_2 + 4\eta^2 \|B \mathbf{N} Y\|_2^2) I \\ &\leq \|Y_\eta\|_2 (4\eta \|B \mathbf{N} Y\|_2 \|A_K\|_2 + 4\eta^2 \|B \mathbf{N} Y\|_2^2) \frac{\Sigma}{\lambda_1(\Sigma)}. \end{aligned}$$

It thus follows that,

$$Y_\eta - Y \leq \frac{\|Y_\eta\|_2 (4\eta \|B \mathbf{N} Y\|_2 \|A_K\|_2 + 4\eta^2 \|B \mathbf{N} Y\|_2^2)}{\lambda_1(\Sigma)} Y.$$

Hence,

$$\|Y_\eta\|_2 \left(1 - \frac{\|Y\|_2 (4\eta \|B \mathbf{N} Y\|_2 \|A_K\|_2 + 4\eta^2 \|B \mathbf{N} Y\|_2^2)}{\lambda_1(\Sigma)} \right) \leq \|Y\|_2.$$

The proof is completed by a direct computation showing that $1/\beta_0 = 1 - \mu_1\eta_0^2 - 4\mu_2\eta_0 > 0$ with the choice of η_0 and for every $\eta \leq \eta_0$,

$$1 - 4\mu_1\eta^2 - 4\mu_2\eta \geq 1 - 4\mu_1\eta_0^2 - 4\mu_2\eta_0.$$

□

The next lemma shows that if c is known, a positive stepsize can be chosen.

Lemma A.2. Let c be the largest real positive number such that A_t is Schur stable for every $t \in [0, c)$ and A_c is marginally Schur stable.¹⁶ Let

$$a_1 = a\beta_0 \|Y\|_2 + 4\|\mathbf{N}\|_2 \beta_0 \|Y\|_2^2, \quad a_2 = a4\|\mathbf{N}\|_2 \beta_0 \|Y\|_2^2;$$

then with $\eta_1 \leq \min(c - \varepsilon, \eta_0, c_0)$, where $\varepsilon > 0$ is an arbitrary positive real number and

$$c_0 < \sqrt{\frac{1}{a_2} + \frac{a_1^2}{4a_2^2}} - \frac{a_1}{2a_2},$$

one has $\phi(\eta_1) \geq 0$.

Proof. The computation follows a similar method used in [17] by replacing the estimate of $Y(\theta)$ by the bound in the above proposition (see details in Lemma 5.5 in [17]). □

Finally, we show that $c > \min(\eta_0, c_0)$. This would then imply that one can choose the stepsize as $\eta = \min(\eta_0, c_0)$.

Theorem A.3. With the stepsize $\eta = \min(\eta_0, c_0)$, M_η remains stabilizing and $\phi(\eta) \geq 0$.

Proof. Let $\eta = \min(\eta_0, c_0)$. It suffices to prove that for every $t \in [0, \eta]$, A_t is Schur stabilizing and $\phi(t) \geq 0$. We prove this by contradiction. Suppose that this is not the case. Then by continuity of eigenvalues, there exists a number $\eta' \leq \eta$ such that A_s is stabilizing for every $s \in [0, \eta')$ and $K_{\eta'}$ is marginally stabilizing. If this is the case, the choice of η_0, c_0 guarantees that for every $s \in [0, \eta')$, $\phi(s)$ is well-defined and $\phi(s) \geq 0$. Now take a sequence $t_i \rightarrow \eta'$ and consider the corresponding sequence of value matrices $\{X_{t_i}\}$. Note that the sequence of function values $\text{Tr}(X_{t_i} \Sigma)$ satisfies,

$$\text{Tr}(X_* \Sigma) \leq \text{Tr}(X_{t_i} \Sigma) \leq \text{Tr}(X \Sigma),$$

since $\phi(t) \geq 0$. But this implies that $\{X_{t_i}\}$ is a bounded sequence (note that the above inequality on function values does not guarantee the boundedness of the sequence; it is crucial that $X_{t_i} \geq X_*$). Hence by a similar argument adopted in the proof of Theorem IV.1, these observations establish a contradiction; as such, the proposed stepsize guarantees stabilization. □

It is now straightforward to conclude the convergence rate of Theorem IV.3 by similar arguments as in [17].¹⁷

¹⁶Here we have assumed that c is not $+\infty$. Of course, if $c = +\infty$, then any stepsize would lead to a stabilizing update.

¹⁷Strictly speaking, we need to show our proposed stepsizes are bounded away from 0. Namely, that there is some constant $d > 0$ such that $\eta_j > d$ for every j . The computations are omitted here due to space limitation. In the meantime, one can be convinced of this fact by checking the asymptotics of η_0 and c_0 .