2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

Sequential Gradient-Descent Optimization of Data-Dependent, Rank-Deficient Cost Functions with a Unique Common Global Minimizer

Adam L. Bruce and Dennis S. Bernstein

Abstract— **Online system identification entails optimization of a sequence of cost functions that are updated by data as it becomes available. The goal is to obtain a sequence of minimizers that converge to the true parameters. If the true parameters are the unique common minimizer of all of the costs in the sequence, then convergence of a sequence of minimizers to the true parameters can be viewed as convergence of the state of a dynamical system to an equilibrium. This paper investigates global asymptotic stability of the equilibrium of a dynamical system defined by online gradient-based optimization of a sequence of quadratic cost functions that have a unique common minimizer, but may individually have multiple global minimizers due to rank deficiency. Under a weak persistency-type condition, it is shown that global asymptotic stability can be guaranteed for this class of costs. These results are specialized to the case of least squares costs and illustrated by examples.**

I. Introduction and Problem Statement

A. Background on Online Parameter Estimation

System identification typically requires online estimation of parameters from a linear regression model. Since data is obtained sequentially during online operation, the task of online identification leads to the problem of optimizing a sequence of costs that are updated at each step by the most recent data. For cumulative least squares costs, Recursive Least Squares (RLS) [1–8] is a well-established method, which includes sophisticated forgetting schemes [9–15] as well as techniques for avoiding divergence when the regressor lacks persistency of excitation [16–20]. However, RLS has the drawback of requiring the propagation of a covariance matrix and is restricted to a cumulative quadratic cost.

Gradient methods [21–23], [24, pp. 58–61] neither require covariance propagation nor assume a cumulative cost, and applications of gradient algorithms such as the stochastic gradient [25, 26], multi-innovation [27–32], and conjugate gradient [33–36] methods to system identification, adaptive control, and adaptive filtering have been studied extensively. Although stability conditions for particular gradient algorithms, such as the instantaneous and instantaneous normalized projected gradient methods [24, pp. 71-73], [37–40] are known, the stability of gradient-based identification methods for general quadratic costs is not well-studied. This can be compared to fixed-cost optimization, where the convergence criteria of gradient methods for large classes of costs are well-established [41, pp. 466-475], [42, pp. 28-35].

A third approach to sequential optimization is Online Convex Optimization (OCO) [43–45], which has emerged as a subfield of machine learning. The OCO literature has studied algorithms closely related to the gradient methods used in identification (cf. section I-E) and has treated problems similar to those of interest in system identification, such as finding a common global minimizer [46–48] or tracking a time-varying global minimizer [49–51] for sets of strongly convex costs. However, as shown in section I-E, the OCO objective of regret minimization may not be effective for system identification since the regret can be minimized without guaranteeing attractivity to the true parameters. This is especially true for non-strictly convex costs, which may have multiple global minimizers.

In this paper, we address the global asymptotic stability of gradient descent for the purpose of online system identification. That is, gradient descent with the objective of determining true system parameters using only sequentially available input/output data (cf. [23], [24, pp. 58–61]). We refer to this approach as *sequential-cost gradient descent* (SGD) to distinguish it from other problems that also use gradient descent (e.g., fixed-cost optimization). We restrict our attention to quadratic costs, which are the most relevant for identification, including possibly rank-deficient costs, which are convex but not strictly convex, and hence we allow for the existence of multiple global minimizers at each step.

Section I-B fixes notation and terms, Section I-C further describes the motivation of the main problem, Section I-D defines the main problem formally as **P1**, and section I-E discusses the relationship of the present work to OCO. In Section II, we present three fixed-point results that are subsequently used to prove global asymptotic stability conditions for SGD in Section III. Finally, these conditions are specialized to least squares costs in Section IV and illustrated with examples in Section V.

B. Notation and Terminology

We define $\mathbb{N} \overset{\triangle}{=} \{1, 2, 3, \dots\}$ and $\mathbb{N}_0 \overset{\triangle}{=} \{0\} \cup \mathbb{N}$. The symbols $\mathbf{S}^n$, $\mathbf{N}^n$, and $\mathbf{P}^n$ denote the sets of real $n \times n$ symmetric, positive-semidefinite, and positive-definite matrices, respectively. For $A \in \mathbf{S}^n$, $\lambda_i(A)$ denotes the ith largest eigenvalue of A, $\lambda_{\max}(A) \overset{\triangle}{=} \lambda_1(A)$, and $\lambda_{\min}(A) \overset{\triangle}{=} \lambda_n(A)$. For all $A \in \mathbb{R}^{n \times n}$, $\mathcal{R}(A)$ and $\mathcal{N}(A)$ denote the range and null space of A, respectively, and A^+ denotes the generalized inverse

Adam L. Bruce and Dennis S. Bernstein are with the Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI 48109, USA, admbruce@umich.edu, dsbaero@umich.edu.

978-1-6654-3658-8/21/$31.00 ©2021 IEEE

of A. The notation $|U|$ denotes the cardinality of the set U. The notation $(x_k)_{k\in\mathbb{N}_0} \subset U$ indicates that the components $x_0, x_1, \ldots$ of the sequence $(x_k)_{k\in\mathbb{N}_0}$ are elements of U. For convenience, we write (x_k) for $(x_k)_{k\in\mathbb{N}_0}$. The empty product and empty sum are defined to be 1 and 0, respectively. Given the n-tuple $\mathcal{J} = (j_1, \ldots, j_n)$ of indices and $r \times r$ matrices $A_{j_1}, \ldots, A_{j_n}$, we define $\prod_{j\in\mathcal{J}} A_j \stackrel{\triangle}{=} A_{j_n} A_{j_{n-1}} \cdots A_{j_1}$.

C. Identification Using Costs with Multiple Minimizers

Let $\mathcal{D} \subset \mathbb{R}^n$ and let $\mathcal{J}$ be a set of differentiable functionals $J\colon \mathcal{D} \to \mathbb{R}$. For all $J \in \mathcal{J}$, let M_J denote the set of global minimizers of J, and denote $M_{\mathcal{J}} \stackrel{\triangle}{=} \cap_{J\in\mathcal{J}} M_J$. It is frequently useful in system identification to consider sets $\mathcal{J}$ such that, for all $J \in \mathcal{J}$, $|M_J| > 1$ but $|M_{\mathcal{J}}| = 1$. For example, the set of instantaneous least squares costs $J_k(x) = \frac{1}{2}\|y_k - \phi_k x\|^2$, $k \geq 0$, where $(y_k) \subset \mathbb{R}^m$ and $(\phi_k) \subset \mathbb{R}^{m\times n}$ is a sequence of regressor matrices such that $\operatorname{rank}(\phi_k) < \min(m, n)$, satisfies this property in the case where $y_k = \phi_k x^*$ and $\cap_{k\geq 1}\mathcal{N}(\phi_k) = \{0\}$. The objective is to identify the single element $x^* \in M_{\mathcal{J}}$, which corresponds to the true system parameters, using only knowledge of the individual costs in $\mathcal{J}$. In particular, since each set M_J contains elements other than x^*, perhaps infinitely many, an algorithm for determining x^* must be capable of distinguishing between points that are only minimizers of a proper subset of the costs in $\mathcal{J}$ and the universal minimizer of every cost in $\mathcal{J}$. Section I-D proposes a simple gradient descent strategy for pursuing this objective.

D. sequential-cost gradient descent

Based on the discussion in the preceding paragraph, we make the following assumption:

A1. *There exists a set* $\mathcal{J} \stackrel{\triangle}{=} \{J\colon \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}\}$ *of differentiable functionals such that, for all* $J \in \mathcal{J}$, $|M_J| \geq 1$, *and* $|M_{\mathcal{J}}| = 1$.

Thus, we allow for the possibility that, for all $J \in \mathcal{J}$, $|M_J| > 1$, but need not assume this *a priori*.

A2. $\mathcal{J}$ *contains a sequence* (J_k) *such that* $\cap_k M_{J_k} = M_{\mathcal{J}}$.

We refer to a sequence satisfying **A2** as an *exhaustive sequence* in $\mathcal{J}$.

Since we are interested in online operation, where each cost J_k is not available until step k, we also make an assumption restricting the availability of costs.

A3. *At each step* k, *the only available cost is* J_k.

Although it is possible to use the previous costs $J_0, \ldots, J_{k-1}$ in addition to J_k at step k, we shall show that it is possible to identify x^* using only the current cost J_k, which is significantly more computationally efficient than holding multiple costs in memory.

Since the most important in system identification applications are quadratic functions, for simplicity in this initial research, we make the following final assumption:

A4. *For all* $k \geq 0$, J_k *is a quadratic function. That is, there exist* $A_k \in \mathbf{N}^n\backslash\{0\}$, $b_k \in \mathcal{R}(A_k)$, *and* $c_k \in \mathbb{R}$, *such that* $J_k(x) \stackrel{\triangle}{=} \frac{1}{2}x^{\mathrm{T}} A_k x + b_k^{\mathrm{T}} x + c_k$.

Note that by restricting attention to quadratic functions, we also assume that $\mathcal{D} = \mathbb{R}^n$. The assumption that $b_k \in \mathcal{R}(A_k)$ is necessary to ensure the existence of finite-norm minimizers, while the assumption that A_k is not necessarily full rank implies that J_k is convex, but not necessarily strictly convex, and hence that there may exist multiple global minimizers. Note that a quadratic cost sequence satisfies **A1-2** if and only if $|\cap_{k\geq 0}[-A_k^+ b_k + \mathcal{N}(A_k)]| = 1$.

Let $(\mu_k) \subset [0, \infty)$. Then the *gradient iteration* of (J_k) is defined as

$$x_{k+1} = x_k - \mu_k \nabla J_k(x_k), \tag{1}$$

$$x_0 \in \mathbb{R}^n, \tag{2}$$

where μ_k is the *step size* at step k. We refer to the use of the gradient iteration to determine x^* as sequential-cost gradient descent. Note that x^* is an equilibrium point of the gradient iteration (1)–(2). The main problem that we address in this paper may be stated as follows:

P1. *Under assumptions* ***A1–A4****, determine sufficient conditions such that* x^* *is a GAS equilibrium of* (1)–(2).

Guaranteeing GAS will prove *a fortiori* that the SGD estimates converge to x^* regardless of the initialization.

E. Relationship with Online Convex Opimization

The SGD algorithm is equivalent to Online Gradient Descent [45, pp. 9-11], [43, pp. 130-134], [44, pp. 179-183], [46] in OCO, possibly with the addition of a projection step, and thus we might initially consider using results from OCO to help answer **P1**. Unfortunately, since OCO is based on regret minimization, it cannot guarantee GAS when $\mathcal{J}$ has costs with multiple global minimizers.

To see this, assume that **A1–A4** hold and recall that the regret of an OCO algorithm is defined [45, pp. 1-2], [43, p. 112], [44, pp. 159-161] by

$$R_T \stackrel{\triangle}{=} \sum_{k=0}^{T} J_k(x_k) - \min_{x\in\mathbb{R}^n} \sum_{k=0}^{T} J_k(x), \tag{3}$$

where x_k is the estimated minimum of J_k at step k. In the OCO framework, the goal is to provide guarantees on the asymptotic growth of R_T, and the main figure of merit is how well R_T can be bounded (possibly asymptotically by a function of T). The ideal performance is $R_T = 0$ for all $T > 0$, but even the best OCO algorithms guarantee only sublinear growth of R_T, since the OCO framework allows J_k to be chosen adversarially [45, p. 6].

The task given in **P1** is to prove GAS of x^*, and hence convergence of the estimate sequence to the true parameters. Hypothetically, if GAS could be guaranteed by bounding R_T, then the methods of OCO might be used to answer **P1**. Unfortunately, as the following example shows, if $\mathfrak{J}$ has costs with multiple global minimizers, then even achieving the ideal OCO performance of $R_T = 0$ for all $T > 0$ is insufficient to guarantee GAS, or even attractivity of x^*.

Example 1. Let $A \in \mathbf{N}^n \setminus \{0\}$, $\|A\| \le 1$, and $\operatorname{rank}(A) < n$, define $J_1(x) = x^T A x$ and $J_2 = x^T(I_n - A)x$, and let $\mathfrak{J} = \{J_1, J_2\}$ and $(J_k) = (J_1, J_2, J_1, J_2, \dots)$. Then $\mathfrak{J}$ satisfies **A1** and (J_k) satisfies **A2**. Since $\mathcal{N}(A)$, $\mathcal{N}(I_n - A) \ne \{0\}$, there exist $y_1 \in \mathcal{N}(A) \setminus \{0\}$ and $y_2 \in \mathcal{N}(I_n - A) \setminus \{0\}$, and thus, for all $\alpha \in \mathbb{R}$, the sequence

$$x_k \triangleq \begin{cases} \alpha^k y_1, & k = 0, 2, 4, \dots \\ \alpha^k y_2, & k = 1, 3, 5, \dots \end{cases} \tag{4}$$

is well-defined and has $R_T = 0$ for all $T > 0$. For all $\alpha > 1$, however, it follows that $\lim_{k\to\infty} \|x_k - x^*\| = \infty$. ◇

In Example 1, since both J_1 and J_2 have an infinite number of global minimizers, there are sequences with identically zero regret that diverge infinitely far from the true parameters. Hence, guarantees on R_T, such as those provided by OCO, do not guarantee attractivity, and thus cannot satisfactorily answer **P1**. In the remainder of the paper, we instead pursue a solution strategy based directly on the definition of stability.

II. Fixed-Point Theory

This section reviews three fixed-point results that are essential for the main results of the paper. Although straightforward, to the authors' knowledge, they have not appeared before in the system identification literature.

Let (M, d) be a metric space, and let $f\colon M \to M$. Then, f is *nonexpansive* if there exists a *nonexpansion coefficient* $q \in (0, 1]$ such that, for all $x, y \in M$, $d(f(x), f(y)) \le q d(x, y)$. The point $p \in M$ is a *fixed point* of f if $f(p) = p$.

Definition 1: Let (M, d) be a metric space, and let (f_k) be a sequence of functions on M. Then $p \in M$ is a *fixed point* of (f_k) if, for all $k \in \mathbb{N}_0$, $f_k(p) = p$. The set of all fixed points of (f_k) is denoted by $\operatorname{Fix}[(f_k)]$.

Proposition 1: For all $k \in \mathbb{N}_0$, let $f_k\colon M \to M$ be nonexpansive with nonexpansion coefficient q_k. Let $x_0 \in M$, define $x_{k+1} = f_k(x_k)$, and assume that p is a fixed point of (f_k). Then $\lim_{k\to\infty} d(x_k, p) \le \left(\prod_{k=0}^{\infty} q_k\right) d(x_0, p)$, and, furthermore, if $\prod_{k=0}^{\infty} q_k = 0$, then p is the only fixed point of (f_k) and $\lim_{k\to\infty} x_k = p$.

Proof. Since, for all $k \in \mathbb{N}_0$, f_k is nonexpansive, it follows that $d(x_{k+1}, p) = d(f_k(x_k), f_k(p)) \le q_k d(x_k, p)$, hence, $d(x_k, p) \le \prod_{\ell=0}^{k-1} q_\ell \; d(x_0, p)$, and thus $\lim_{k\to\infty} d(x_k, p) \le \left(\prod_{k=0}^{\infty} q_k\right) \; d(x_0, p)$. Setting $\prod_{k=0}^{\infty} q_k = 0$, it follows that $\lim_{k\to\infty} d(x_k, p) = 0$. Suppose that $p' \in \operatorname{Fix}[(f_k)]$. Then $d(p', p) \le \lim_{k\to\infty} d(p', x_k) + \lim_{k\to\infty} d(x_k, p) = 0$. □

Proposition 2: Let (f_k) be a sequence of nonexpansive functions on (M, d) with fixed point p. Then p is a Lyapunov stable equilibrium of the system

$$x_{k+1} = f_k(x_k), \tag{5}$$

$$x_0 \in M. \tag{6}$$

Proof. Let $\varepsilon > 0$ and $x_0 \in B_\varepsilon(p)$. Since (f_k) is nonexpansive, it follows that, for all $k \ge 0$, $d(x_k, p) = d(f_{k-1}(x_{k-1}), f_k(p)) \le d(x_{k-1}, p) \le d(x_0, p) < \varepsilon$. □

A *bounded interval* of $\mathbb{N}_0$ is a set $\{n, n+1, \dots, n+m\} \subset \mathbb{N}_0$, where $n, m \in \mathbb{N}_0$. A *bounded interval partition* of $\mathbb{N}_0$ is a partition of $\mathbb{N}_0$ whose elements are bounded intervals, and a *uniformly bounded interval partition* P of $\mathbb{N}_0$ is a bounded interval partition of $\mathbb{N}_0$ such that $\sup_{U\in P} |U| < \infty$.

Proposition 3: Let (f_k) be a sequence of nonexpansive functions on (M, d) with fixed point p, let (U_k) be a bounded interval partition of $\mathbb{N}_0$, and, for all $k \in \mathbb{N}_0$, define

$$F_k \triangleq f_{\max U_k} \circ \cdots \circ f_{\min U_k}. \tag{7}$$

Then the following statements hold:

i) p is a fixed point of (F_k).

ii) (F_k) is nonexpansive.

iii) For all $x \in M$, $\lim_{k\to\infty} F_k \circ \cdots \circ F_0(x) = p$ if and only if $\lim_{k\to\infty} f_k \circ \cdots \circ f_0(x) = p$.

Proof. *i)* and *ii)* are immediate from (7). To prove *iii)*, let $z \in M$, and define $(y_k), (x_k) \subset M$ by $x_0 = z$, $y_0 = z$, and, for all $k \in \mathbb{N}_0$, by $x_{k+1} = f_k(x_k)$ and $z_{k+1} = F_k(z_k)$. Suppose that $\lim_{k\to\infty} x_k = p$. Since (z_k) is a subsequence of (x_k), it follows that $\lim_{k\to\infty} z_k = p$. Conversely, suppose that $\lim_{k\to\infty} z_k = p$, and let $\varepsilon > 0$. Then there exists $k \in \mathbb{N}_0$ such that, for all $k \ge K$, $d(z_k, p) < \varepsilon$. Since (z_k) is a subsequence of (x_k), it follows that there exists $M \ge 0$ such that $z_K = x_M$. Let $m > M$. Then, since f_k is nonexpansive, it follows that

$$\begin{aligned} d(x_m, p) &= d(f_m^{n_m} \circ \cdots \circ f_M^{n_M}(x_M), f_m^{n_m} \circ \cdots \circ f_M^{n_M}(p)) \\ &\le d(x_M, p) = d(z_K, p) < \varepsilon, \end{aligned}$$

and thus $\lim_{k\to\infty} x_k = p$. □

III. Global Asymptotic Stability

In this section, we state and prove sufficient conditions for GAS of SGD in Theorem 1, answering **P1**. This is our main result, the proof of which requires the following three lemmas.

Lemma 1: Let (U_k) be a bounded interval partition of $\mathbb{N}_0$ and, for all $k \ge 0$, let $A_k \in \mathbf{N}^n$, with $\|A_k\| \le 1$ and $\lim_{k\to\infty} \left\|\prod_{j\in U_k} A_j\right\| = 1$. Then $\lim_{k\to\infty} \|A_k\| = 1$.

Proof. Let $\varepsilon > 0$. Since $\lim_{k\to\infty} \left\|\prod_{j\in U_k} A_j\right\| = 1$ there exists $K \ge 0$ such that, for all $k \ge K$ and $j \in U_k$, $1 - \varepsilon <$

$\left\|\prod_{j\in U_k} A_j\right\| \le \prod_{j\in U_k} \|A_j\| \le \|A_j\| \le 1$. Thus, for all $j \ge \min U_K$, $1-\varepsilon \le \|A_j\| \le 1$.

Lemma 2: Let (U_k) be a uniformly bounded interval partition of $\mathbb{N}_0$ and let $(a_k) \subset \mathbb{R}$ be a sequence such that $\lim_{k\to\infty} a_k = 0$. Then $\lim_{k\to\infty} \sum_{j\in U_k} a_j = 0$.

Proof. Let $\varepsilon > 0$ and $\sup_{k\in\mathbb{N}_0} |U_k| = M$. Since $\lim_{k\to\infty} a_k = 0$, it follows that there exists $K \in \mathbb{N}_0$ such that, for all $k \ge K$, $|a_k| < \varepsilon/M$. Since (U_k) is an interval partition, it follows that there exists $K_1 \ge 0$ such that, for all $k \ge K_1$, $\min U_k > K$. Let $k \ge K_1$. Since, for all $j \in U_k$, $j \ge K$, it follows that $\left|\sum_{j\in U_k} a_j\right| \le \sum_{j\in U_k} |a_j| < \frac{\varepsilon}{M}|U_k| \le \frac{\varepsilon}{M} \sup_{k\in\mathbb{N}_0} |U_k| = \varepsilon$. □

To see that the assumption of uniform boundedness is essential, let $U_0 = \{0\}$, $U_1 = \{1\}$, $U_3 = \{2,3\}$, $U_4 = \{4,5,6\}$, $U_5 = \{7,8,9,10\}$, ... Then (U_k) is a bounded, but not uniformly bounded interval partition. For all $k \in \mathbb{N}_0$, let $\{a_j\}_{j\in U_k} = \{\frac{1}{k+1}, \ldots, \frac{1}{2k}\}$. Then $\sum_{j\in U_k} a_j = \ln(2) + \varepsilon_{2k} - \varepsilon_k$, where $\lim_{k\to\infty} \varepsilon_k = 0$. Thus $\lim_{k\to\infty} \sum_{j\in J_k} a_j = \ln(2)$ even though $\lim_{k\to\infty} a_k = 0$.

Definition 2: Let $(S_k) \subset \mathbf{N}^n$. Then (S_k) is *ultimately positive* if $\liminf_{k\to\infty} \lambda_{\min}(S_k) > 0$ and *weakly ultimately positive* if there exists a uniformly bounded interval partition (U_k) of $\mathbb{N}_0$ such that $\left(\sum_{j\in U_k} S_j\right)$ is ultimately positive. The sequence $(\phi_k) \subset \mathbb{R}^{n\times m}$ is ultimately positive, or weakly ultimately positive if $(\phi_k^{\mathrm{T}}\phi_k)$ is ultimately positive, or weakly ultimately positive, respectively.

Lemma 3: For all $k \in \mathbb{N}_0$, let $A_k \in \mathbf{N}^n\backslash\{0\}$,

$$\mu_k \in [0, \lambda_{\max}^{-1}(A_k)], \tag{8}$$

and define $q_k \overset{\triangle}{=} \left\|\prod_{j\in U_k} (I_n - \mu_j A_j)\right\|$. Furthermore, assume that $(\mu_k A_k)$ is weakly ultimately positive. Then $\limsup_{k\to\infty} q_k < 1$.

Proof. From (8), it follows that $\limsup_{k\to\infty} q_k \le 1$. Hence, suppose for contradiction that $\limsup_{k\to\infty} q_k = 1$. From Lemma 1, it follows that $\limsup_{k\to\infty} \|I_n - \mu_k A_k\| = 1$, and thus $\limsup_{k\to\infty} (1 - \|I_n - \mu_k A_k\|) = 0$. For all $k \in \mathbb{N}_0$, let x_k be the unit eigenvector of A_k corresponding to $\lambda_{\max}(A_k)$ and let $\xi \in B_1(0)$. Since $\mu_k A_k x_k = (1 - \|I - \mu_k A_k\|) x_k$, Lemma 2 implies that

$$\begin{aligned}
\liminf_{k\to\infty} \left[\lambda_{\min}\left(\sum_{j\in U_k} \mu_j A_j\right)\right] &\le \liminf_{k\to\infty} \left\|\left[\sum_{j\in J_k} \mu_j A_j\right]\xi\right\| \\
&\le \liminf_{k\to\infty} \left\|\sum_{j\in J_k} \mu_j A_j x_j\right\| \le \liminf_{k\to\infty} \sum_{j\in J_k} \|\mu_j A_j x_j\| \\
&= \liminf_{k\to\infty} \sum_{j\in J_k} (1 - \|I_n - \mu_j A_j\|) \\
&\le \limsup_{k\to\infty} \sum_{j\in J_k} (1 - \|I_n - \mu_j A_j\|) = 0,
\end{aligned}$$

which contradicts the assumption that $(\mu_k A_k)$ is weakly ultimately positive. □

Theorem 1: Under the notation and assumptions **A1–A4**, let $(\mu_k) \subset [0,\infty)$ satisfy (8) and assume that $(\mu_k A_k)$ is weakly ultimately positive. Then $x^* \in M_{\mathcal{J}}$ is the globally asymptotic stable equilibrium of (1)–(2).

Proof. The point x^* is an equilibrium of (1)–(2) by definition. Consider the sequence $(f_k\colon \mathbb{R}^n \to \mathbb{R}^n)$ defined for all $k \ge 0$ by $f_k(x) \overset{\triangle}{=} x - \mu_k \nabla J_k(x)$, where $(J_k) \subset \mathcal{J}$ is exhaustive. Since (8) implies that $\|I - \mu_k A_k\| = 1 - \mu_k \lambda_{\min}(A_k) \in [0,1]$, it follows that, for all $x, y \in \mathbb{R}^n$, $\|f_k(x) - f_k(y)\| \le \|I - \mu_k A_k\| \|x - y\| \le \|x - y\|$, and thus (f_k) is nonexpansive. From Proposition 2, it follows that x^* is Lyapunov stable.

To prove attractivity, let (U_k) be a uniformly bounded interval partition for which $(\mu_k A_k)$ is weakly ultimately positive, for all $k \in \mathbb{N}_0$, define F_k by (7) with f_k given as in the previous paragraph, and define $q_k \overset{\triangle}{=} \left\|\prod_{j\in U_k} (I_n - \mu_j A_j)\right\|$. Using (7), it follows that, for all $k \in \mathbb{N}_0$, $\|F_k(x) - F_k(y)\| = \left\|\left[\prod_{j\in J_k} (I_n - \mu_j A_j)\right](x-y)\right\| \le q_k \|x-y\|$. Finally, $\|I_n - \mu_k A_k\| \le 1$ implies that $q_k \in [0,1]$. Next, since $(\mu_k A_k)$ is weakly ultimately positive, it follows that Lemma 3 holds, and thus (3) implies that $1 - \limsup_{k\to\infty} q_k \in (0,1)$. Hence, let $\alpha \in (0, 1 - \limsup_{k\to\infty} q_k)$, and, for all $k \in \mathbb{N}_0$, define

$$\tilde{q}_k \overset{\triangle}{=} \begin{cases} q_k, & q_k \neq 0, \\ \alpha, & q_k = 0. \end{cases}$$

Since, for all $k \in \mathbb{N}_0$, $\tilde{q}_k \in (0,1]$ and, for all $x, y \in \mathbb{R}^n$, $\|F_k(x) - F_k(y)\| \le \tilde{q}_k \|x - y\|$, it follows that $\tilde{q}_k$ is a nonexpansion coefficient for f_k. Since $\tilde{q}_k \in (0,1]$, it follows that $\prod_{k=0}^{\infty} \tilde{q}_k$ is either zero or positive. Suppose for contradiction that $\prod_{k=0}^{\infty} \tilde{q}_k$ is positive. Then $\lim_{k\to\infty} \tilde{q}_k = 1$, which implies that q_k is ultimately positive, and hence $\lim_{k\to\infty} q_k = \lim_{k\to\infty} \tilde{q}_k = 1$, contradicting Lemma 3. Thus $\prod_{k=0}^{\infty} \tilde{q}_k = 0$.

From **A1**, it follows that x^* is the unique fixed point of (f_k). From Proposition 3, *i)* and 1 it follows that x^* is the unique fixed point of (F_k), and, for all $x \in \mathbb{R}^n$, $\lim_{k\to\infty} F_k \circ \cdots \circ F_0(x) = x^*$. Thus, it follows from Proposition 3, *iii)* that $\lim_{k\to\infty} x_k = x^*$. Since the initialization is arbitrary, it follows that x^* is GAS. □

Note that Theorem 1 combined with the uniqueness of the universal fixed point in Proposition 1 implies that if (μ_k) satisfies (8), $(\mu_k A_k)$ is weakly ultimately positive, and $\cap_{k\ge 0}[-A_k^+ b_k + \mathcal{N}(A_k)] \neq \varnothing$, then, moreover, $|\cap_{k\ge 0} [-A_k^+ b_k + \mathcal{N}(A_k)]| = 1$ and $x^* \in \cap_{k\ge 0}[-A_k^+ b_k + \mathcal{N}(A_k)]$ is the GAS equilibrium of (1)–(2).

IV. Least Squares Costs

Least squares costs form a significant subset of the quadratic costs commonly used in practice. For this special case, the results in the previous section can be specialized.

Definition 3: The sequence (J_k) of quadratic cost functions is a *least squares sequence* if, for all $k \in \mathbb{N}_0$, there exist $\phi_k \in \mathbb{R}^{p\times n}\backslash\{0\}$, $y_k \in \mathbb{R}^p$, $\ell_k \in \{0,\dots,k\}$ and, for all $1 \le i \le \ell_k$, $W_{k,i} \in \mathbf{P}^p$ such that

$$J_k(x) = \frac{1}{2}\sum_{i=0}^{\ell_k}(y_{k-i} - \phi_{k-i}x)^{\mathrm{T}} W_{k,i}(y_{k-i} - \phi_{k-i}x). \quad (9)$$

Note that in the case where $\sum_{i=0}^{\ell_k}\phi_{k-i}^{\mathrm{T}} W_{k,i}\phi_{k-i}$ is rank-deficient, J_k has an infinite number of global minimizers.

Theorem 2: Let $(\phi_k) \subset \mathbb{R}^{p\times n}$ be weakly ultimately positive, and, for all $k \in \mathbb{N}_0$, define $y_k = \phi_k x^*$, where $x^* \in \mathbb{R}^n$, let $0 \le \ell_k \le k$, and assume that, for all $1 \le i \le \ell_k$, $W_{k,i} \in \mathbf{P}^n$ and $W_{k,i} \ge \xi I_n$, where $\xi > 0$. Finally, let $(\mu_k) \subset [0,\infty)$ be an ultimately positive sequence such that, for all $k \in \mathbb{N}_0$,

$$\mu_k \le \frac{1}{\lambda_{\max}\left(\sum_{i=0}^{\ell_k}\phi_{k-i}^{\mathrm{T}} W_{k,i}\phi_{k-i}\right)}, \quad (10)$$

Then x^* is the globally asymptotically stable equilibrium of (1)–(2) with J_k given, for all $k \ge 0$, by (9).

Proof. Note that (9) can be written as $J_k(\tilde{x}) \overset{\triangle}{=} \frac{1}{2}\tilde{x}^{\mathrm{T}} A_k \tilde{x}$, where $A_k = \sum_{i=0}^{\ell_k}\phi_{k-i}^{\mathrm{T}} W_{k,i}\phi_{k-i}$ and $\tilde{x} = x^* - x$. Then μ_k satisfies (8) by construction. To show that $(\mu_k A_k)$ is weakly ultimately positive, let (U_k) be a uniformly bounded interval partition with respect to which (ϕ_k) is weakly ultimately positive, and for all $k \in \mathbb{N}_0$, define $\mu_k^- = \min_{j\in U_k}\mu_j$. Since (μ_k) is ultimately positive, (μ_k^-) is ultimately positive, and thus

$$\begin{aligned}
&\liminf_{k\to\infty}\lambda_{\min}\left(\sum_{j\in U_k}\mu_j A_j\right)\\
&\ge \xi \liminf_{k\to\infty}\mu_k^- \lambda_{\min}\left(\sum_{j\in U_k}\sum_{i=0}^{\ell_j}\phi_{j-i}^{\mathrm{T}}\phi_{j-i}\right)\\
&\ge \xi \liminf_{k\to\infty}\mu_k^- \lambda_{\min}\left(\sum_{j\in U_k}\phi_j^{\mathrm{T}}\phi_j\right) > 0. \qquad \square
\end{aligned}$$

V. Examples

The following examples illustrate Theorems 1 and 2.

Example 2. Considering the cost set and sequence of Example 1, for all $k \in \mathbb{N}_0$, let $\mu_k = \min(1, \lambda_{\max}^{-1}(A))$, and let $(U_k) = (\{0,1\},\{2,3\},\{4,5\},\dots)$. Then Theorem 2 implies that $x^* = 0$ is the GAS equilibrium of (1)–(2). $\diamond$

Example 3. Let $\theta = [1\ 2]^{\mathrm{T}}$, and, for all $k \in \mathbb{N}_0$, define

$$\phi_k = \begin{cases} \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{k} \end{bmatrix}, & k = 0,2,4,\dots, \\ \begin{bmatrix} \frac{1}{\sqrt{k}} & 0 \\ 0 & 0 \end{bmatrix}, & k = 1,3,5,\dots, \end{cases} \quad (11)$$

$y_k = \phi_k\theta$, and let $J_k(x) = \frac{1}{2}\|y_k - \phi_k x\|^2$. Let $\mu_k = k^2$ and consider the uniformly bounded interval partition given in Example 2. Then Theorem 2 implies that θ is the GAS equilibrium of (1)–(2). Note that θ is GAS even though μ_k is unbounded. $\diamond$

Example 4. Let $\theta \in \mathbb{R}^n$, let $N_w \ge 0$, let $(\phi_k) \subset \mathbb{R}^{p\times n}$ be weakly ultimately positive, and assume that $\beta \overset{\triangle}{=} \sup_{j\ge 0}\lambda_{\max}\left(\phi_j^{\mathrm{T}}\phi_j\right) < \infty$. For all $k \in \mathbb{N}_0$, let $y_k = \phi_k\theta$, and define $J_k(\hat{\theta}) = \frac{1}{2}\sum_{i=0}^{N_w}\|y_{k-i} - \phi_{k-i}\hat{\theta}\|^2$. Finally, let (μ_k) be an ultimately positive sequence such that $\mu_k \in [0, \frac{1}{N_w\beta}]$. Then (μ_k) satisfies (10), since $\lambda_{\max}\left(\sum_{i=0}^{N_w}\phi_{k-i}^{\mathrm{T}}\phi_{k-i}\right) \le N_w\beta$, and Theorem 2 implies that θ is the GAS equilibrium of (1), (2). $\diamond$

VI. Conclusion

Sufficient conditions were given under which the sequential-cost gradient descent is GAS with an equilibrium corresponding to the true system parameters. GAS was obtained regardless of whether or not the individual costs have a unique minimizer (that is, are strictly convex). Since quadratic costs are the most common type in system identification, we restricted attention to this important class of costs. Specialized conditions were given for least squares costs, including rank-deficient least squares costs with an infinite number of global minimizers. Future work will consider extensions to sequences of nonquadratic convex cost functions, the effect of noise, and extension to costs without a common global minimizer.

Acknowledgments

This research was partially supported by AFOSR under grant FA9550-20-1-0028.

References

[1] R. L. Plackett, “Some Theorems in Least Squares,” *Biometrika*, vol. 37, pp. 149–157, 1950.

[2] A. Albert and R. W. Sittler, “A Method for Computing Least Squares Estimators that Keep Up with the Data,” *SIAM J. Contr.*, vol. 3, no. 3, pp. 384–417, 1965.

[3] T. Kailath, “An Innovations Approach to Least-Squares Estimation Part I: Linear Filtering in Additive White Noise,” *IEEE Trans. Automatic Control*, vol. AC-13, no. 6, pp. 646–655, 1968.

[4] R. A. Geesey and T. Kailath, “An Innovations Approach to Least-Squares Estimation Part IV: Recursive Estimation Given Lumped Covariance Functions,” *IEEE Trans. Automatic Control*, vol. AC-16, no. 6, pp. 217–226, 1971.

[5] K. S. Miller, “Complex Linear Least Squares,” *SIAM Review*, vol. 15, no. 4, pp. 706–726., 1973.

[6] K. J. Astrom, *Adaptive Control*, 2nd ed. Reading, MA: Addison-Wesley, 1995.

[7] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification*. The MIT Press, 1983.

[8] S. A. U. Islam and D. S. Bernstein, “Recursive Least Squares for Real-Time Implementation,” *IEEE Cont. Sys. Mag.*, vol. 39, no. 3, pp. 82–85, 2019.

[9] S.-H. Leung and C. F. So, “Gradient-Based Variable Forgetting Factor RLS Algorithm in Time-Varying Environments,” *IEEE Trans. Sig. Proc.*, vol. 53, no. 8, pp. 3141–3150, 2005.

[10] S. Song, J.-S. Lim, S. J. Baek, and K.-M. Sung, “Gauss Newton variable forgetting factor recursive least squares for time varying parameter tracking,” *Electron. Lett.*, vol. 36, no. 11, pp. 988–990, 2000.

[11] D. J. Park, B. E. Jun, and K. J. H., “Fast tracking RLS algorithm using novel variable forgetting factor with unity zone,” *Electron. Lett.*, vol. 27, no. 23, pp. 2150–2151, 1991.

[12] T. R. Fortescue, L. S. Kershenbaum, and B. E. Ydstie, “Implementation of Self-tuning Regulators with Variable Forgetting Factors,” *Automatica*, vol. 17, no. 6, pp. 831–835, 1981.

[13] C. Paleologu, J. Benesty, and C. Silviu, “A Robust Variable Forgetting Factor Recursive Least-Squares Algorithm for System Identification,” *IEEE Sig. Proc. Lett.*, vol. 15, pp. 597–600, 2008.

[14] R. M. Johnstone, C. R. J. Johnson, R. R. Bitmead, and B. D. O. Anderson, “Exponential convergence of recursive least squares with exponential forgetting factor,” *21st IEEE Conference on Decision and Control*, pp. 994–997, 1982.

[15] A. L. Bruce, A. Goel, and D. S. Bernstein, “Convergence and consistency of recursive least squares with variable-rate forgetting,” *Automatica*, vol. 119, p. 109052, 2020.

[16] K. Narendra and A. Annaswamy, “A new adaptive law for robust adaptation without persistent excitation,” *IEEE Trans. Automatic Control*, vol. 32, no. 2, pp. 134–145, 1987.

[17] G. Chowdhary, T. Yucelen, M. Mühlegg, and E. N. Johnson, “Concurrent learning adaptive control of linear systems with exponentially convergent bounds,” *Int. J. Adapt. Contr. Sig. Proc.*, vol. 27, no. 4, pp. 280–301, 2013.

[18] G. Chowdhary, M. Mühlegg, and E. N. Johnson, “Exponential parameter and tracking error convergence guarantees for adaptive controllers without persistency of excitation,” *Int. J. Contr.*, vol. 87, no. 8, pp. 1583–1603, 2014.

[19] A. Parikh, R. Kamalapurkar, and W. E. Dixon, “Integral concurrent learning: Adaptive control with parameter convergence using finite excitation,” *Int. J. Adapt. Contr. Sig. Proc.*, vol. 33, no. 12, pp. 1775–1787, 2019.

[20] A. Goel, A. L. Bruce, and D. S. Bernstein, “Recursive least squares with variable-direction forgetting–compensating for the loss of persistency,” *IEEE Contr. Sys.*, vol. 40, pp. 80–102, 2020.

[21] J. M. Mendel, “Gradient identification for linear systems,” in *Adaptive, Learning and Pattern Recognition Systems: Theory and Applications*, ser. Mathematics in Science and Engineering, J. M. Mendel and K. S. Fu, Eds. Cambridge, Massachusetts: Academic Press, 1970, vol. 66, ch. 6, pp. 209–242.

[22] R. M. Laurenson and J. R. Baumgarten, “Application of gradient search procedures for the identification of unknown system parameters from system response observations,” *J. Eng. Ind.*, vol. 94, no. 1, pp. 109–114, 1972.

[23] J. U. T. Cate and H. J. Verbruggen, “least-squares like gradient method for discrete process identification,” *Int. J. Control*, vol. 28, no. 6, pp. 933–952, 1978.

[24] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence, and Robustness*. Prentice-Hall, 1989. Dover Republication, 2011.

[25] R. D. Poltmann, “Stochastic gradient algorithm for system identification using adaptive fir-filters with too low number of coefficients,” *IEEE Trans. Circuits and Systems*, vol. 35, no. 2, pp. 247–250, 1988.

[26] D. Levanony and N. Berman, “Recursive nonlinear system identification by a stochastic gradient algorithm: stability, performance, and model nonlinearity considerations,” *EEE Trans. Sig. Proc.*, vol. 52, no. 9, pp. 2540–2550, 2004.

[27] F. Ding and T. Chen, “Performance analysis of multi-innovation gradient type identification methods,” *Automatica*, vol. 43, no. 1, pp. 1–14, 2007.

[28] Y. Liu, L. Yu, and F. Ding, “Multi-innovation extended stochastic gradient algorithm and its performance analysis,” *Circuits, Systems, and Signal Processing*, vol. 29, no. 4, pp. 649–667, 2010.

[29] J. Chen and F. Ding, “Least squares and stochastic gradient parameter estimation for multivariable nonlinear box-jenkins models based on the auxiliary model and the multi-innovation identification theory,” *Engineering Computations*, vol. 29, no. 8, pp. 907–921, 2012.

[30] Y. Mao and F. Ding, “Multi-innovation stochastic gradient identification for hammerstein controlled autoregressive systems based on the filtering technique,” *Nonlinear Dynamics*, vol. 79, pp. 1745—1755, 2015.

[31] X. Wang and F. Ding, “Convergence of the auxiliary model-based multi-innovation generalized extended stochastic gradient algorithm for box–jenkins systems,” *Nonlinear Dynamics*, vol. 82, pp. 269—-280, 2015.

[32] Q. Liua, D. F., Z. Q., and T. T. Haya, “Two-stage multi-innovation stochastic gradient algorithm for multivariate output-error arma systems based on the auxiliary model,” *Int. J. Systems Science*, vol. 50, no. 15, p. 2870–2884, 2019.

[33] G. K. Boray and M. Srinath, “Conjugate gradient techniques for adaptive filtering,” *IEEE Trans. Circuits and Systems I: Fundamental Theory and Applications*, vol. 39, no. 1, pp. 1–10, 1992.

[34] T. Bose and M.-Q. Chen, “Conjugate gradient method in adaptive bilinear filtering,” *IEEE Trans. Sig. Proc.*, vol. 43, no. 6, pp. 1503–1508, 1995.

[35] P. S. Chang and A. N. Willson, “Analysis of conjugate gradient algorithms for adaptive filtering,” *IEEE Trans. Sig. Proc.*, vol. 48, no. 2, pp. 409–418, 2000.

[36] Z. Shengkui, M. Zhihong, and N. K. Suiyang, “Conjugate gradient algorithm design with rls normal equation,” in *Proc. 6th Int. Conference on Information, Communications, & Signal Processing*, 10–13 Dec. 2007.

[37] M. M. Sondhi and D. Mitra, “New results on the performance of a well-known class of adaptive filters,” *Proc. IEEE*, vol. 64, no. 11, pp. 1583–1597, 1976.

[38] A. P. Morgan and K. S. Narendra, “On the uniform asymptotic stability of certain linear nonautonomous differential equations,” *SIAM J. Cont. Opt.*, vol. 15, no. 1, pp. 163–176, 1977.

[39] B. D. O. Anderson, “Exponential stability of linear equations arising in adaptive identification,” *IEEE Trans. Aut. Contr.*, vol. AC-22, no. 1, pp. 83–88, 1977.

[40] G. Kreisselmeier, “Adaptive estimators with exponential rate of convergence,” *IEEE Trans. Aut. Contr.*, vol. AC–22, no. 1, pp. 2–8, 1977.

[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[42] Y. Nesterov, *Lectures on Convex Optimization, 2nd Edition*. Cham, Switzerland: Springer Nature Switzerland, 2004.

[43] S. Shalev-Shwartz, “Online learning and online convex optimization,” *Foundations and trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[44] E. Hazan, “Introduction to online convex optimization,” *Foundations and Trends in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[45] F. Orabona, “A modern introduction to online learning,” 2020, [Online; accessed 15-March-2021]. [Online]. Available: https://open.bu.edu/handle/2144/40900

[46] P. Bartlett, E. Hazan, and A. Rakhlin, “Adaptive online gradient descent,” Electrical Engineering and Computer Sciences, University of California at Berkele, Berkeley, CA, Tech. Rep. Technical Report No. UCB/EECS-2007-82, June 2007.

[47] R. Dixit, A. S. Bedi, R. Tripathi, and R. Rajawat, “Online learning with inexact proximal online gradient descent algorithms,” *IEEE Transactions on Signal Processing*, vol. 67, no. 5, pp. 1338–1352, 2019.

[48] A. S. Bedi, P. Sarma, and K. Rajawat, “Tracking moving agents via inexact online gradient descent algorithm,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 202–216, 2018.

[49] Y. Ding, J. Lavaei, and M. Arcak, “Escaping spurious local minimum trajectories in online time-varying nonconvex optimization,” *arXiv:1912.00561v1*, 2019.

[50] S. Park, J. Mulvaney-Kemp, M. Jin, and J. Lavaei, “Diminishing regret for online nonconvex optimization,” https://lavaei.ieor.berkeley.edu/regret_ONO_2020_1.pdf, 2020, [Online; accessed 18-September-2020].

[51] H. Feng and J. Lavaei, “Analysis of the landscape of time-varying non-convex optimization problems via linear operators,” https://lavaei.ieor.berkeley.edu/Online_opt_2020_1.pdf, 2020, [Online; accessed 18-September-2020].