This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

Stochastic Adaptive Optimization with Dithers

Siyu Xie, Shu Liang, Le Yi Wang, *Fellow, IEEE,* George Yin, *Fellow, IEEE,* and Wen Chen

***Abstract*—Optimization methods are essential and have been used extensively in a broad spectrum of applications. Most existing literature on optimization algorithms does not consider systems that involve unknown system parameters. This paper studies a class of stochastic adaptive optimization problems in which identification of unknown parameters and search for the optimal solutions must be performed simultaneously. Due to a fundamental conflict between parameter identifiability and optimality in such problems, we introduce a method of adding stochastic dither signals into the system, which provide sufficient excitation for estimating the unknown parameters, leading to convergent adaptive optimization algorithms. Joint identification and optimization algorithms are developed and their simultaneous convergence properties of parameter estimation and optimization variable updates are proved. Under both noise-free and noisy observations, the corresponding convergence rates are established. The main results of this paper reveal certain fundamental relationships and trade-off among updating step sizes, dither magnitudes, parameter estimation errors, optimization accuracy, and convergence rates. Simulation case studies are used to illustrate the adaptive optimization algorithms and their main properties.**

***Index Terms*—Optimization, identification, adaptive optimization algorithm, stochastic dither, convexity**

I. Introduction

OPTIMIZATION techniques have been widely used in many disciplines, including engineering, statistics, economics, management sciences, computer science, and mathematics [1], [2]. Recent advances of industrial manufacturing, artificial intelligence, and machine learning have introduced a large variety of optimization methods and algorithms such as model predictive control (MPC), perception/inference, and expected/empirical risk minimization [3]. To deal with the underlying big data, large-scale structure, and distributed resources, control and optimization of cyber-physical systems have become highly prevalent, where physical plants such as collaborative robotics, unmanned vehicles, and mobile devices work cooperatively with network communications and computing to accomplish joint missions [4], [5].

Optimization in cyber-physical systems encounters fundamental challenges of modeling errors, system uncertainties, and disturbances. Such uncertainties on physical systems significantly affect optimization problems and introduce technical difficulties in algorithm development and convergence analysis. This paper teats constrained optimization problems with uncertainties in the physical plant. Model uncertainties and external disturbances are major concerns in control systems, and many related theories and techniques have been developed such as robust control [6], [7], system identification [8], stochastic approximation [9], and adaptive control [10], [11]. However, the corresponding counterpart in optimization has not been extensively explored, partly due to theoretical difficulties and lack of physical plant consideration. Common terminologies in optimization such as stochastic gradient methods [3], inexact oracle [12], robust optimization [13], [14], and adaptive optimization [15]–[18] from the optimization field have quite different meanings from control systems, since they do not consider uncertain physical plants. For example, Kosmatopoulos et al. [15]–[17] considered adaptive control problems and used optimization to construct an adaptive control law, based on system models. Since our system configuration involves physical systems in the loop, new methods are needed.

Optimization of physical systems with parameter uncertainties must identify unknown parameters and search for optimal solutions collaboratively in a cyber-physical setting. There is a fundamental conflict between parameter identifiability and optimality in such problems. In this paper, we introduce new adaptive optimization algorithms that consist of cyber updating of optimization variables, dithered signals for system probing, physical state observation with possible noise corruption, and parameter estimation. In particular, adding stochastic dither signals into the system provides sufficient excitation for estimating the unknown parameters. Many parameter estimation algorithms [9], [11], [19]–[23], including the least mean squares (LMS) algorithm, least squares (LS) algorithm, Kalman filtering algorithm, have been studied in system identification, control, signal processing and numerous other fields. Here, convergence results of LMS and LS algorithms in [23] and [11] are employed to establish the main convergence properties of our algorithms.

In this paper, joint identification and optimization algorithms are developed and their simultaneous convergence properties of parameter estimation and optimization variable updates are proved. Under both noise-free and noisy observations, the corresponding convergence rates are established. It is shown that under noise-free state observation, the algorithms converge to the optimal solution with $O(1/k)$ convergence rate under convexity assumption, and exponential convergence rate under strong convexity. Under noisy observations, the

Siyu Xie is with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, Michigan 48202, USA <syxie@wayne.edu>

Shu Liang is with the Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China <sliang@ustb.edu.cn>

Le Yi Wang is with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, Michigan 48202, USA <lywang@wayne.edu>

George Yin is with the Department of Mathematics, University of Connecticut, Storrs, CT 06269-1009, USA <gyin@uconn.edu>

Wen Chen is with the Division of Engineering Technology, Wayne State University, Detroit, MI 48202, USA <wchenc@wayne.edu>

This research was supported in part by the Air Force Office of Scientific Research under grant FA9550-18-1-0268.

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

algorithms are shown to achieve $O(\log(k)/k)$ convergence rate under strong convexity conditions. Our algorithm design and analysis methods extend existing convex optimization algorithms to accommodate possible system uncertainties.

The main contributions of this paper are summarized as follows

1) A new adaptive optimization method for systems with parameter uncertainties is introduced by integrating system identification and optimization in a cyber-physical setting.
2) To achieve identification and optimization simultaneously, a method of adding stochastic dither signals into the system is introduced.
3) Adaptive optimization algorithms under noise-free and noisy state observations are developed, together with rigorous convergence analysis on joint procedures of parameter estimation and optimization.
4) Fundamental relationships and trade-off among updating step sizes, dither magnitudes, parameter estimation errors, optimization accuracy, and convergence rates are established quantitatively.

The remainder of the paper is organized as follows. In Section II, the optimization problem with unknown system parameters is formulated. The adaptive optimization algorithms without measurement noises are presented in Section III, together with convergence analysis on the parameter estimation and optimal solution. Section IV is focused on the adaptive optimization algorithms with measurement noise and their convergence properties. Simulation case studies are performed in Section V to illustrate the adaptive optimization algorithms and their main properties. Finally, some concluding remarks are provided in Section VI.

II. Problem Formulation

Consider a network of n subsystems, where u^i is the local control input of the ith subsystem, and x^{ij} is the link variable from node i to node j. The vectors containing all the control inputs and network states are denoted by u and x, respectively. The optimization problem for a physical system, subject to an unknown parameter vector θ, is

$$\begin{aligned} J^* &= \min_{u,x} J = f(u,x), \\ \text{s.t.} &\quad h(u,x,\theta)=0, \end{aligned} \tag{1}$$

where $f(u,x)$ is the performance index with respect to (w.r.t) the local control variable $u \in \mathbb{R}^m$ and the network state variable $x \in \mathbb{R}^n$. The function $h(u,x,\theta)=0$ characterizes the true physical system in which the true parameter $\theta \in \mathbb{R}^s$ is unknown and assumed to belong to a bounded uncertain set Θ. In our algorithms, we will use measured data on the physical system to estimate the unknown parameter θ and perform optimization simultaneously.

For simplicity and as the first step, assume that the constraint $h(u,x,\theta)=0$ entails the following explicit relationship:

$$x = \phi^\top(u)\theta \tag{2}$$

where $\phi^\top(u) = [h^1(u),\dots,h^s(u)]$ is the regression matrix which is a function of u. Note that (2) is linear w.r.t the unknown parameter θ, but may be nonlinear w.r.t the control variable u, exemplified by Hammerstein systems, and others. Consequently, we have a core mathematics problem as follows

$$\begin{aligned} J^* &= \min_{u} \ J_1(u,\theta) = f(u,\phi^\top(u)\theta), \\ \text{s.t.} &\quad x = \phi^\top(u)\theta. \end{aligned} \tag{3}$$

III. Adaptive Optimization without Measurement Noise

To establish the key convergence properties and develop viable algorithms, we start with the scenario of noise-free state observations. The prior information on the true unknown parameter θ is given by the bounded parameter set Θ. Select an initial parameter $\theta_0 \in \Theta$. The physical relationship between u and x is given by (2) with the true unknown parameter θ.

There is a fundamental conflict between optimality and identifiability in this problem: If u^* is the optimal solution, the regressor $\phi^\top(u)$ becomes a constant matrix and the observation relationship in (2)

$$x = \phi^\top(u^*)\theta$$

does not provide sufficient information for identifying θ, regardless how many measurements on x are collected. In other words, when an extreme-searching algorithm for u converges to its optimal solution, the identifiability for the unknown parameter θ is naturally lost. This fundamental conflict between identification and optimization implies that certain mechanisms should be developed so that searching for optimal solution and persistent identifiability can be simultaneously sustained.

In this paper, we introduce a method of adding small stochastic dither signals to resolve this conflict. Here we assume that $J_1(u,\theta)$ is differentiable w.r.t u, and denote the gradient of $J_1(u,\theta)$ w.r.t u as $\nabla_u J_1(u,\theta)$. The word "adaptation" means that the optimal solution changes with the true parameter, and as such our algorithm must follow the unknown parameter θ. Consequently, the algorithms for estimation of the unknown parameter and optimization must be integrated, leading to the following "adaptive optimization" algorithm.

In Algorithm 1, μ and η are the step sizes that will be selected for convergence, L_k is the adaptation gain to be specified later, and $\delta_k \in \mathbb{R}^m$ is an excitation dither added by the designer for estimation of θ and satisfies:

$$\delta_k \sim U(-\delta,\delta), \tag{5}$$

which means that $\{\delta_k\}$ is a sequence of independent and identically distributed random variables with uniform distribution $\delta_k^i \sim U(-\delta,\delta)$. Note that the dither magnitude $\delta > 0$ is usually chosen to be small so that its disturbance to the optimal solution can be reduced.

Remark 3.1: We would like to point out that deterministic periodic dithers can also be used to fulfill the estimation task, with one major drawback: We must know the system order (or the number of unknown parameters) *a priori*. In contrast, the stochastic dither we use is independent of the system order. Consequently, we may adaptively identify the order of the

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

```
**Algorithm 1** Adaptive Optimization Algorithm
**Initialization**: Select an initial control input and initial parameter estimate as
    $u_0 \in \mathbb{R}^m, \quad \theta_0 \in \mathbb{R}^s.$
**Update flows**: At each time instant $k \geq 0$:
  • Cyber update
        $u_{k+1} = u_k - \mu \nabla_u J_1(u_k, \theta_k)$    (4a)
  • Add a dither $\delta \in \mathbb{R}^m$ to the computed $u$
        $v_k = u_k + \delta_k$    (4b)
  • Apply the dithered input to the physical system
        $x_k = \phi^\top(v_k)\theta$    (4c)
  • Estimate parameter
        $\theta_{k+1} = \theta_k + \eta L_k(x_k - \phi^\top(v_k)\theta_k)$    (4d)
```

system and its parameters simultaneously. While this topic is beyond the scope of this paper, the framework is set up to accommodate future development.

Remark 3.2: Note that for some common network structures and performance indices, Algorithm 1 will lead to a strictly distributed algorithm, i.e., the updates of u^i and θ^{ij} involve only local variables. For example, if the performance index is separable

$$J = f(u, x) = \sum_{i \in \mathcal{V}} f_u^i(u^i) + \sum_{(i,j) \in \mathcal{E}} f_x^{ij}(x^{ij}),$$

and the constraint is local

$$x^{ij} = \phi^\top(u^i, u^j)\theta^{ij},$$

then the corresponding algorithm is strictly distributed. Indeed, when u^i and θ^{ij} use their own step sizes, the updates of u^i and θ^{ij} will only use local information. We emphasize that although the performance index is separated in this example, the physical variables are coupled via the constraint.

A. Convergence analysis of θ_k

Due to their simplicity and robustness, the LMS-like algorithms [9], [19]–[23] are broadly used in signal processing, system identification, adaptive filtering, adaptive control, etc. Here we employ the following normalized LMS algorithm to estimate the unknown parameter θ:

$$\theta_{k+1} = \theta_k + \eta \frac{\phi(v_k)}{1 + \|\phi(v_k)\|^2}(x_k - \phi^\top(v_k)\theta_k), \tag{6}$$

i.e., $L_k = \frac{\phi(v_k)}{1+\|\phi(v_k)\|^2}$, where $0 < \eta < 1$ is the step size.

Define $\widetilde{\theta}_k = \theta_k - \theta$ as the estimation error, and $x_k = \phi^\top(v_k)\theta$. We have the following estimation error equation:

$$\begin{aligned}\widetilde{\theta}_{k+1} =& \widetilde{\theta}_k - \eta \frac{\phi(v_k)\phi^\top(v_k)}{1 + \|\phi(v_k)\|^2}\widetilde{\theta}_k \\ =& \left(I_s - \eta \frac{\phi(v_k)\phi^\top(v_k)}{1 + \|\phi(v_k)\|^2}\right)\widetilde{\theta}_k, \end{aligned} \tag{7}$$

where I_s denotes the $s \times s$ identity matrix. By the definition of v_k, it is easy to see that the stochastic regressor $\phi(v_k)$ is not an independent signal. From (7), it is clear that the property of the estimation error $\widetilde{\theta}_k$ relies essentially on the product of the random matrix sequence $\left\{I_s - \eta \frac{\phi(v_k)\phi^\top(v_k)}{1+\|\phi(v_k)\|^2}\right\}$. The reason to use the normalized algorithm is that adding the denominator makes the signal $\frac{\phi(v_k)\phi^\top(v_k)}{1+\|\phi(v_k)\|^2}$ bounded. This makes it easier in selecting an appropriate step size η to ensure the stability and convergence of the algorithm. In addition, it enables us to use the theoretical results in [23] directly, where the stability of the normalized LMS algorithm is analyzed under non-independent and non-stationary signal assumptions.

Before analyzing the property of the estimation error $\widetilde{\theta}_k$, we need the following notations. Let $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ be two symmetric matrices. $A \geq B$ means $A - B$ is positive semi-definite. $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and the smallest eigenvalues of a symmetric matrix, respectively. For any matrix $X \in \mathbb{R}^{m \times n}$, the Euclidean norm is defined as $\| X \| = \{\lambda_{\max}(XX^\top)\}^{\frac{1}{2}}$; and for any random matrix A, its L_p-norm is $\| A \|_{L_p} = \{\mathbb{E}[\| A \|^p]\}^{\frac{1}{p}}$, where $\mathbb{E}[\cdot]$ is the mathematical expectation. For theoretical analysis, we need the following definition introduced in [23].

Definition 3.1: For a sequence of $s \times s$ random matrices $A = \{A_k, k \geq 0\}$, if it belongs to the following set with $p \geq 1$,

$$\begin{aligned} S_p(\lambda) = \Bigg\{ A : \Bigg\| \prod_{j=i+1}^{k} (I_s - A_j) \Bigg\|_{L_p} \leq M\lambda^{k-i}, \\ \forall k \geq i+1, \forall i \geq 0, \text{for some } M > 0 \Bigg\}, \end{aligned} \tag{8}$$

then $\{I - A_k, k \geq 0\}$ is called L_p-exponentially stable with parameter $\lambda \in [0, 1)$.

It is well known that analysis of a random matrix product is a mathematically difficult problem. However, as demonstrated by Guo [23], for linear random equations arising from adaptive filtering algorithms, it is possible to transfer the product of the random matrices to that of a certain class of scalar sequences, and the later can be further analyzed based on some excitation conditions on the regressors. To this end, we introduce the following subclass of $S_1(\lambda)$ for a scalar sequence $a = \{a_k, k \geq 0\}$

$$\begin{aligned} S^0(\lambda) = \Bigg\{ a : a_k \in [0, 1], \mathbb{E}\Bigg[\prod_{j=i+1}^{k} (1 - a_j)\Bigg] \leq M\lambda^{k-i}, \\ \forall k \geq i+1, \forall i \geq 0, \text{for some } M > 0 \Bigg\}, \end{aligned} \tag{9}$$

where $\lambda \in [0, 1)$. This definition will be used when we transfer the product of random matrices to that of a scalar sequence. To proceed, we introduce the following excitation condition for estimating the unknown parameter θ.

Assumption 3.1: The regressor $\{\phi(v_k), \mathcal{F}_k\}$ is an adapted sequence of random vectors (i.e., $\phi(v_k)$ is $\mathcal{F}_k$-measurable, for all k, where $\{\mathcal{F}_k\}$ is a sequence of non-decreasing σ-algebras),

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

and there exists an integer $h > 0$ such that $\{\lambda_k\} \in S^0(\lambda)$ for some $\lambda \in [0, 1)$, where λ_k is defined by

$$\lambda_k \triangleq \lambda_{\min}\left\{\mathbb{E}\left[\frac{1}{h+1}\sum_{i=kh+1}^{(k+1)h}\frac{\phi(v_i)\phi^\top(v_i)}{1+\|\phi(v_i)\|^2}\Big|\mathcal{F}_{kh}\right]\right\} \quad (10)$$

where $\mathbb{E}[\cdot|\cdot]$ is the conditional expectation.

Remark 3.3: Note that no assumptions of independence, stationarity, or mixing are made on the regressor signals $\{\phi(v_k)\}$. The optimization variables in a gradient-based or other algorithms are generated to achieve optimality. Technically, it is not known if the excitation condition (i.e., *Assumption 3.1*) can be verified *a priori* without a dither. In Section V, it is shown that if no dither signal is added in the system, the errors are large since the estimator of the unknown parameters cannot converge to the true values, implying that signals do not naturally produce sufficient information for identification. Adding dithers resolves this issue and provides a general approach to integrate optimization and identification. Moreover, for sustained operation in cyber-physical systems, the concept of "persistent identifiability" is used, meaning that the integrated algorithms for identifying the unknown parameter θ must maintain their identifiability, independent of the u sequence, starting time, ending time, leading to more challenging demands on the u sequence. In this case, the introduction of dither signals is sufficient to satisfy *Assumption 3.1*.

Remark 3.4: While other stochastic signals can also be used as dithers, the uniform distribution is zero mean and bounded, and can easily fulfill *Assumption 3.1* for the estimation of θ. We should emphasize that a zero mean and bounded stochastic dither with other types of distributions can also be employed, as long as *Assumption 3.1* is satisfied. Moreover, by Section III.C, the convergence rate of the estimator of θ increases with the increase of the variance of the uniform distribution. Therefore, for a non-uniformly distributed dither signal, the convergence rate may also depend on its variance.

Theorem 3.1: Consider the estimation error equation (7). Suppose that *Assumption 3.1* is satisfied. Then for any $p \geq 1$ and $\eta \in (0, 1)$, there exists a constant $M > 0$ such that $\forall k \geq i \geq 0$,

$$\left\|\prod_{j=i+1}^{k}\left(I_s - \eta\frac{\phi(v_j)\phi^\top(v_j)}{1+\|\phi(v_j)\|^2}\right)\right\|_{L_p} \leq M\lambda^{\eta\alpha_p(k-i)} \quad (11)$$

where

$$\alpha_p = \begin{cases}\frac{1}{8h(1+h)^3}, & 1 \leq p \leq 2\\ \frac{1}{4h(1+h)^3 p}, & p > 2,\end{cases} \quad (12)$$

and λ and h are defined in *Assumption 3.1*.

Proof: By *Assumption 3.1*, we know that

$$\{\lambda_k\} \in S^0(\lambda), \qquad \lambda_k \leq \frac{h}{1+h}. \quad (13)$$

Then, by *Lemma 2.3* in [23], we have

$$\{\eta\lambda_k\} \in S^0(\lambda^{\frac{\eta}{1+h}}). \quad (14)$$

Thus, by *Theorem 2.1* in [23],

$$\left\{I_s - \eta\frac{\phi(v_k)\phi^\top(v_k)}{1+\|\phi(v_k)\|^2}, k \geq 1\right\} \in S_p(\lambda^{\eta\alpha_p})$$

holds. By the definition of S_p, it is easy to see that (11) holds. This completes the proof. ■

From *Theorem 3.1*, and by assuming that for some $p \geq 1$, $\|\widetilde{\theta}_0\|_{L_{2p}} < \infty$ holds, we have by (7) that

$$\begin{aligned}\|\widetilde{\theta}_k\|_{L_p} =& \left\|\prod_{i=0}^{k-1}\left(I_s - \eta\frac{\phi(v_i)\phi^\top(v_i)}{1+\|\phi(v_i)\|^2}\right)\widetilde{\theta}_0\right\|_{L_p}\\ \leq& \left\|\prod_{i=0}^{k-1}\left(I_s - \eta\frac{\phi(v_i)\phi^\top(v_i)}{1+\|\phi(v_i)\|^2}\right)\right\|_{L_{2p}} \cdot \|\widetilde{\theta}_0\|_{L_{2p}}\\ \leq& O(\lambda^{\eta\alpha_{2p}k}),\end{aligned} \quad (15)$$

which tends to zero exponentially as $k \to \infty$, where $\alpha_{2p} = \frac{1}{8h(1+h)^3 p}$ and $O(\cdot)$ means that there exists a constant $M > 0$ such that $\|\widetilde{\theta}_k\|_{L_p} \leq M\lambda^{\eta\alpha_{2p}k}$. Thus, as $k \to \infty$,

$$\mathbb{E}[\|\widetilde{\theta}_k\|^2] \to 0, \quad \text{exponentially fast}, \quad (16)$$

and

$$\theta_k \to \theta \text{ w.p.1 exponentially fast} \quad (17)$$

where (17) can be proved by using (16) and the Borel-Cantelli Lemma. Note also that the rate of the exponential convergence, i.e., $\lambda^{\frac{\eta}{16h(1+h)^3}}$, increases with the increase of the step size η and the decrease of the constant λ and h.

B. Convergence analysis of u_k

We are now in a position to present the first convergence result on the performance index $J_1(u, \theta)$ under the following convexity condition.

Assumption 3.2: $J_1(u, \theta)$ is differentiable and convex w.r.t u, and $\nabla_u f(u, x)$ is Lipschitz continuous. Moreover, $J_1(u, \theta)$ is level bounded, i.e., all sets of the form $\{u \in \mathbb{R}^m | J_1(u, \theta) \leq c\}$ for $c \in \mathbb{R}$, are bounded.

Remark 3.5: There are many common optimization problems that satisfy the convexity condition. For example, for separable performance indices $f(u, x) = f_1(u) + f_2(x)$ with convex $f_1(u)$ and $f_2(x)$, since the convexity is invariant under affine mappings, under the system $x = Au + B$, $J_1(u, \theta)$ will be convex w.r.t u. Also, there is a large class of general functions that satisfy this condition:

- if $f_2(x)$ is convex and non-decreasing, and $x = \phi^\top(u)\theta$ is convex w.r.t u, then $J_1(u, \theta)$ will be convex w.r.t u.
- if $f_2(x)$ is convex and non-increasing, and $x = \phi^\top(u)\theta$ is concave w.r.t u, then $J_1(u, \theta)$ will be convex w.r.t u.

For example, if $f_2(x) = e^x$ that is convex and nondecreasing, and $\phi(u) = e^u$ or any $|u|^p$ for $p \geq 1$ with all the elements of θ being non-negative, then $J_1(u, \theta)$ is convex. Furthermore, for common quadratic problems, the second example of Section V provides a nonlinear system that satisfies the convex assumption, i.e., if $f_2(x)$ is quadratic and $\phi(u) = e^u$, or $|u|^p$ for $p \geq 1/2$, then $J_1(u, \theta)$ is convex w.r.t u. Note that these examples also satisfy the level-bounded assumption for $J_1(u, \theta)$. Moreover, *Assumption 3.2* implies that $\nabla_u J_1(u, \theta)$ is Lipschitz continuous w.r.t both u and θ for some constant $L > 0$ since θ and θ_k are bounded, which can be derived from (15).

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

Theorem 3.2: Under *Assumptions 3.1* and *3.2*, there exists a constant $\bar{\mu}$ such that for any $0 < \mu < \bar{\mu}$, the sequence u_k generated by Algorithm 1 and (6) converges with rate

$$J_1(u_k, \theta) - J_1(u^*, \theta) \leq O(1/k) \quad \text{w.p.1.} \tag{18}$$

The proof of *Theorem 3.2* is in Appendix. Furthermore, when the function J_1 is strongly convex, we can show that the convergence of u_k is exponentially fast.

Assumption 3.3: $J_1(u, \theta)$ is differentiable and strongly convex w.r.t u, and $\nabla_u f(u, x)$ is Lipschitz continuous.

Theorem 3.3: Under *Assumptions 3.1* and *3.3*, there exists a constant $\bar{\mu}$ such that for any $0 < \mu < \bar{\mu}$, the sequence u_k generated by Algorithm 1 and (6) satisfies

$$\mathbb{E}[\|u_k - u^*\|^2] \to 0 \text{ exponentially fast,} \tag{19}$$

and

$$u_k \to u^* \quad \text{w.p.1 exponentially fast,} \tag{20}$$

as $k \to \infty$.

The proof of *Theorem 3.3* is in Appendix.

Remark 3.6: From the above analysis, it is easy to see that the rate of exponential convergence of u_k, i.e., $1 - \xi = 1 - \mu[(c - L\varepsilon) - \mu L^2(1 + \varepsilon)] > \lambda^{\frac{\eta}{4h(1+h)^3}}$ where μ and ε satisfy (65), increases with the increase of the step sizes μ and η. Note that the step size μ satisfies (65) and $0 < \eta < 1$, which cannot be too large.

Note also that the actual implemented control input to the physical system is $v_k = u_k + \delta_k$, where δ_k is independently and identically distributed with each element $\delta_k^i \sim U(-\delta, \delta)$. Thus, it is easy to see that

$$\begin{aligned}\mathbb{E}[\|v_k - u^*\|^2] =& \mathbb{E}[\|u_k + \delta_k - u^*\|^2] \\ \leq& \mathbb{E}[\|u_k - u^*\|^2] + \mathbb{E}[\|\delta_k\|^2] \to \frac{m}{3}\delta^2,\end{aligned} \tag{21}$$

as $k \to \infty$. This will create a physical disturbance to the system and deviation from the optimal solution. This deviation is small if a small δ is selected. However, the smaller the magnitude δ, the larger the exponent λ, and hence the slower the convergence of θ_k and u_k. Such a tradeoff is fundamental.

For practical implementation, one may choose a threshold value σ, e.g., 0.01, such that when the error between u_k and u_{k-1} is smaller than σ, the updating of u_k and θ_k stops. After that one may remove the dither ($\delta_k = 0$), and use u_k as the final optimal solution of the optimization problem.

C. Verification of Assumption 3.1

Assumption 3.1 is essential for convergence of the adaptive optimization algorithms. In this subsection, we will show that adding stochastic dithers δ_k can provide sufficient excitation for estimation task of θ in some common systems. From the estimation error equation (7), it is not difficult to see that $\|\widetilde{\theta}_k\|$ is uniformly bounded by $\|\widetilde{\theta}_0\|$, which implies that $\|\theta_k\|$ is uniformly bounded. Note also that for the convex case, $\|u_k\|$ is bounded by (50). As for the strongly convex case, we know from (66) that $\|u_k - u^*\|^2 \leq \|u_0 - u^*\|^2 + \frac{\omega}{\xi}\|\widetilde{\theta}_0\|^2$. Thus, $\|u_k\|$ is also bounded.

Example 1: We first consider the linear system

$$x = Au, \tag{22}$$

where $x \in \mathbb{R}^n, u \in \mathbb{R}^m$, and $A = [a^{ij}]_{n\times m}$. Denote $x = [x^1, \ldots, x^n]^\top$ and $u = [u^1, \ldots, u^m]^\top$. Then, (22) can be represented as

$$x^i = u^\top\theta^i = \phi^\top(u)\theta^i, \ i = 1, \ldots, n, \tag{23}$$

where $\theta^i = [a^{i1}, \ldots, a^{im}]^\top$ and $\phi^\top(u) = u^\top$.

It is noted that ϕ is independent of i. Consequently, we focus on one specific i for a generic verification of the identifiability. Consider the following generic regression model:

$$x = u^\top\theta = \phi^\top(u)\theta, \tag{24}$$

where x is a scalar, and $\theta \in \mathbb{R}^m$ is the true and unknown parameter. Thus, $\phi(v_k) = u_k + \delta_k$ in this case.

Proposition 3.1: For the regression model (24), if a dither signal $\delta_k \in \mathbb{R}^m$, which satisfies (5), is added in u_k, then *Assumption 3.1* is satisfied with $h = 1$ and $\lambda = 1 - \frac{\delta^2}{3[1+(\bar{u}+\delta)^2]} \in (0, 1)$, where $\bar{u} = \sup_k \|u_k\|$.

The proof of *Proposition 3.1* is in Appendix. Note that λ decreases with the increase of δ. Thus, for the selection of $\delta > 0$, there is a tradeoff between the convergence rate and the mean square error of v_k.

Example 2: Consider an affine system

$$x = Au + B, \tag{25}$$

where $A = [a^{ij}]_{n\times m}$ and $B = [b^1, \ldots, b^n]^\top$. Denote $x = [x^1, \ldots, x^n]^\top$ and $u = [u^1, \ldots, u^m]^\top$. Then, (25) can be represented as

$$x^i = [u^\top, 1]\theta^i = \phi^\top(u)\theta^i, \ i = 1, \ldots, n, \tag{26}$$

where $\theta^i = [a^{i1}, \ldots, a^{im}, b^i]^\top$ and $\phi^\top(u) = [u^\top, 1]$. Similarly, we consider a generic regression model

$$x = [u^\top, 1]\theta = \phi^\top(u)\theta, \tag{27}$$

where x is a scalar, and $\theta \in \mathbb{R}^{m+1}$ is the true and unknown parameter. Then we have

$$\phi^\top(v_k) = [u_k^\top + \delta_k^\top, 1].$$

Note that

$$\phi(v_k)\phi^\top(v_k) = \begin{bmatrix} u_k + \delta_k \\ 1 \end{bmatrix} [u_k^\top + \delta_k^\top, 1]. \tag{28}$$

Proposition 3.2: For the regression model (27), if a dither signal $\delta_k \in \mathbb{R}^m$, which satisfies (5), is added in u_k, then *Assumption 3.1* is satisfied with $h = 1$ and $\lambda = 1 - \frac{\Delta}{2+(\bar{u}+\delta)^2} \in (0, 1)$, where $\bar{u} = \sup_k \|u_k\|$ and Δ is defined in the proof.

The proof of *Proposition 3.2* is in Appendix.

Example 3: Consider general nonlinear systems

$$x = \phi^\top(u)\theta, \tag{29}$$

where $\phi^\top(u) \in \mathbb{R}^{n\times p}$ is possibly nonlinear.

Proposition 3.3: For the regression model (29), we assume that $p \leq m$, $\phi(u)$ is continuously twice differentiable, and $\nabla\phi(u)$ is full rank (rank p) uniformly over all u. If a dither signal $\delta_k \in \mathbb{R}^m$ that satisfies (5) is added in u_k, then *Assumption 3.1* is satisfied with $h = 1$ and $\lambda = 1 - \frac{\Delta}{1+(\bar{\phi}+\delta\bar{\phi}_\Delta+\delta^2\bar{\rho})^2} \in (0, 1)$, where $\Delta, \bar{\phi}, \bar{\phi}_\Delta, \bar{\rho}$ are defined in the proof.

The proof of *Proposition 3.3* is in Appendix.

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

IV. Adaptive optimization with measurement noises

When the states are measured in physical systems, they are always subject to measurement or communication noises. In this section, we treat the noisy observation relationship

$$x_k = \phi^\top(v_k)\theta + d_k, \tag{30}$$

where $d_k = [d_k^1, \ldots, d_k^n]^\top \in \mathbb{R}^n$ is the measurement noise process. In this case, if we still utilize the normalized LMS algorithm with constant step size η, the estimate θ_k may not converge to θ because of the noise process d_k. Thus, we will use the following least squares (LS) algorithm to estimate the unknown parameter θ:

$$\begin{cases} \theta_{k+1} = \theta_k + a_k P_k \phi(v_k)(x_k - \phi^\top(v_k)\theta_k), \\ P_{k+1} = P_k - a_k P_k \phi(v_k)\phi^\top(v_k)P_k, \\ a_k = (1 + \phi^\top(v_k)P_k\phi(v_k))^{-1}, \end{cases} \tag{31}$$

where the initial estimate $\theta_0 \in \mathbb{R}^p$, and the initial positive definite matrix $P_0 \in \mathbb{R}^{p\times p}$ can be chosen arbitrarily. Note that in practice P_0 is usually set as $\alpha_0 I_p$, where α_0 is a positive constant.

A. Convergence analysis of θ_k

We assume that $\phi(v_k)$ is $\mathcal{F}_k$-measurable, where $\{\mathcal{F}_k\}$ is a sequence of nondecreasing σ-algebras. For theoretical analysis, we need the following standard condition on the noise processes.

Assumption 4.1: The noise sequence $\{d_k, \mathcal{F}_k\}$ is a martingale difference sequence where $\{\mathcal{F}_k\}$ is a sequence of nondecreasing σ-algebras (i.e., d_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[d_k|\mathcal{F}_{k-1}] = 0$, for all k), and there exists a constant $\beta > 2$ such that

$$\sup_{k\geq 0} \mathbb{E}[\|d_k\|^\beta | \mathcal{F}_{k-1}] < \infty \quad \text{w.p.1.} \tag{32}$$

An important example is the case where $\{d_k\}$ is a sequence of independent random variables with zero mean and satisfying (32). It is well-known that the estimation error of the above LS algorithm has the following upper bound (see [11], [24]) as $k \to \infty$.

Lemma 4.1: Let *Assumption 4.1* be satisfied, the LS estimates (31) such that

$$\begin{aligned} &\|\theta_{k+1} - \theta\|^2 \\ =&O\left(\frac{\log\left(\lambda_{\max}\{P_0^{-1}\} + \sum\limits_{j=0}^{k} \|\phi(v_j)\|^2\right)}{\lambda_{\min}\left\{P_0^{-1} + \sum\limits_{j=0}^{k} \phi(v_j)\phi^\top(v_j)\right\}} \right) \quad \text{w.p.1.} \end{aligned} \tag{33}$$

Remark 4.1: The proof of *Lemma 4.1* can be found in [11], [24]. It is easy to see that the LS estimates will converge to the true parameter if

$$\lim_{k\to\infty} \frac{\log\left(\lambda_{\max}\{P_0^{-1}\} + \sum\limits_{j=0}^{k} \|\phi(v_j)\|^2\right)}{\lambda_{\min}\left\{P_0^{-1} + \sum\limits_{j=0}^{k} \phi(v_j)\phi^\top(v_j)\right\}} = 0, \quad \text{w.p.1.} \tag{34}$$

Moreover, examples can be constructed to show that if the above limit is a nonzero constant, then the LS estimate cannot converge to the true parameter (see [24]). In this sense, one can say that the condition (34) is the weakest possible one for the convergence of the LS algorithm [24]. Note that for the linear, affine, and nonlinear systems studied in Section II.C, $\log(\lambda_{\max}\{P_0^{-1}\} + \sum_{j=0}^k \|\phi(v_j)\|^2)$ is of the order $O(\log(k + 1 + \alpha_0))$.

For the linear system case, we know that

$$\begin{aligned} &\frac{1}{k+1}\sum_{j=0}^{k} \phi(v_j)\phi^\top(v_j) = \frac{1}{k+1}\sum_{j=0}^{k} (u_j + \delta_j)(u_j + \delta_j)^\top \\ =&\frac{1}{k+1}\sum_{j=0}^{k} (u_j u_j^\top + \delta_j u_j^\top + u_j \delta_j^\top + \delta_j \delta_j^\top) \\ \geq&\frac{1}{k+1}\sum_{j=0}^{k} (\delta_j u_j^\top + u_j \delta_j^\top + \delta_j \delta_j^\top) \\ \to&\frac{1}{k+1}\sum_{j=0}^{k} \delta_j \delta_j^\top, \ k \to \infty, \end{aligned} \tag{35}$$

which means that $\lambda_{\min}\{\sum_{j=0}^k \phi(v_j)\phi^\top(v_j)\}/k+1$ tends to be larger than $\lambda_{\min}\{\mathbb{E}[\delta_0\delta_0^\top]\}$ as $k \to \infty$, which is positive since the expectation of the covariance matrix is positive definite, i.e., $\frac{\delta^2}{3} I_m$. Thus,

$$\|\theta_k - \theta\|^2 = O\left(\frac{\log(k+\alpha_0)}{k+\alpha_0}\right) \quad \text{w.p.1.} \tag{36}$$

In a similar way, for the affine and general nonlinear systems (i.e., Examples 2 and 3), (36) also holds.

B. Convergence analysis of u_k

By *Lemma 4.1*, we can obtain the following result on the convergence of u_k, whose proof is in Appendix.

Theorem 4.1: Under *Assumptions 3.3* and *4.1*, there exists a constant $\bar{\mu}$ such that for any $0 < \mu < \bar{\mu}$, the sequence u_k generated by Algorithm 1 and (31) satisfies

$$u_k \to u^* \quad \text{w.p.1 as} \quad k \to \infty, \tag{37}$$

and the convergence speed is not slower than $O\left(\frac{\log(k/2+\alpha_0)}{k/2+\alpha_0}\right)$.

V. Simulation Examples

In this section, we use some examples to illustrate the adaptive optimization algorithms. The goal is to verify the theoretical results established in the paper and to show the relationships and trade-off among updating step sizes, dither magnitudes, optimization errors, and convergence rates. Here we will focus on the noise-free case for simplicity.

In the first example, we consider an affine system:

$$x = \begin{bmatrix} x^1 \\ \vdots \\ x^{20} \end{bmatrix} = Au+b = \begin{bmatrix} a^{11} & \ldots & a^{1,20} \\ \vdots & \ddots & \vdots \\ a^{20,1} & \ldots & a^{20,20} \end{bmatrix} \begin{bmatrix} u^1 \\ \vdots \\ u^{20} \end{bmatrix} + \begin{bmatrix} b^1 \\ \vdots \\ b^{20} \end{bmatrix},$$

where $a^{ij}, b^i, i, j = 1, \ldots, 20$ are unknown parameters which need to be estimated during the optimization process.

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

The performance index is

$$J(x,u)=\frac{1}{2}\sum_{i=1}^{20}[(x^i)^2+(u^i)^2], \tag{38}$$

i.e.,

$$J_1(u,\theta)=\frac{1}{2}\sum_{i=1}^{20}\sum_{j=1}^{20}(a^{ij}u^j+b^i)^2+\frac{1}{2}\sum_{i=1}^{20}(u^i)^2.$$

Here we assume that the true values of the unknown parameters are random integers which are chosen from $[-10,10]$. For the simulation settings, the initial values of a_0^{ij} and b_0^i are all chosen to be 1. Here for $i=1,\ldots,20$, we compare the noise-free adaptive optimization algorithm with the following algorithm without estimation:

$$u_{k+1}^i=u_k^i-\mu\left[\sum_{j=1}^{20}a_0^{ji}\left(\sum_{\ell=1}^{20}a_0^{j\ell}u_k^\ell+b_0^j\right)+u_k^i\right], \tag{39}$$

where the initial values are used to calculate the optimal solution. Let $\mu=0.0001,\delta=0.1$. Then, we can obtain the error trajectories

$$v_k^i-u^{i,*},\quad i=1,\ldots,20,\quad k=0,100,\ldots,15000,$$

for the noise-free adaptive optimization algorithm with $\eta=0.3,0.9$, and

$$u_k^i-u^{i,*},\quad i=1,\ldots,20,\quad k=0,100,\ldots,15000,$$

for (39), where $u^{i,*}$ is the optimal value of u^i. To make the figures be more concise, we only plot the error trajectories for two elements, i.e., $i\in\{1,2\}$. Note that we can obtain similar results for the remaining ones.

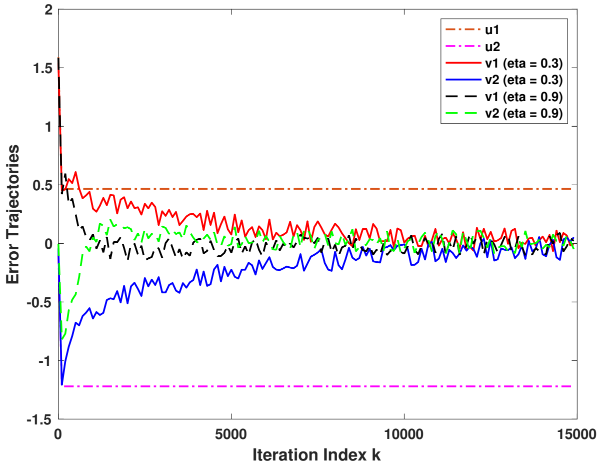


Fig. 1: Error trajectories for $i\in\{1,2\}$ of the noise-free adaptive optimization algorithm and (39)

The upper and lower lines in Fig. 1 are the error trajectories for $i\in\{1,2\}$ of (39), showing that errors are quite large because the values of the parameters used are incorrect. The other four lines in the middle of Fig. 1 are the error trajectories of the noise-free adaptive optimization algorithm with $\eta=0.3$ and 0.9, illustrating that the error trajectories tend to zero because of the estimation of the unknown parameters. The fluctuations around 0 stem from the random perturbations from the dither δ_k. Also the convergence rate increases as the step size η increases.

For different magnitude δ values, we can also obtain the error trajectories

$$v_k^i-u^{i,*},\quad i=1,\ldots,20,\quad k=0,100,\ldots,15000,$$

for the noise-free adaptive optimization algorithm in Fig. 2 with $\mu=0.0001,\eta=0.9,\delta=0,0.05,0.1$. Note that only the error trajectories for $i\in\{1,2\}$ are plotted here for simplicity.

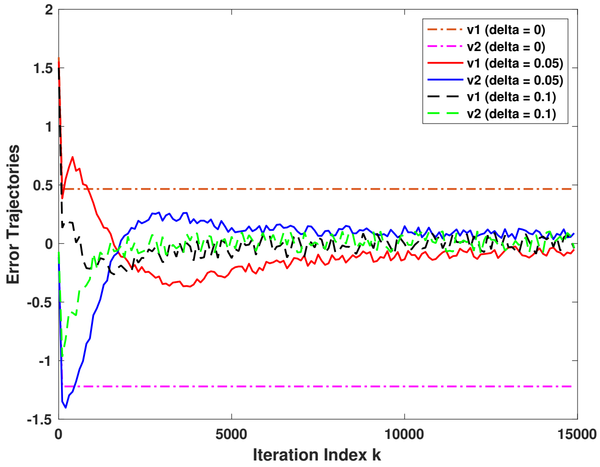


Fig. 2: Error trajectories for $i\in\{1,2\}$ of the noise-free adaptive optimization algorithm with $\delta=0,0.05,0.1$

When $\delta=0$, i.e., there is no excitation dither in the system, the errors are large since the estimation of the unknown parameters cannot converge to the true values. As δ increases, both the convergence rate and the magnitude of random fluctuations increase, which shows the fundamental tradeoff between the convergence rate and optimization error.

In addition, we can obtain the following logarithm trajectories of the estimation error

$$\log(|a_k^{ij}-a^{ij}|)/k,\ k=400,410,\ldots,1000,\ i,j=1,2,$$

and

$$\log(|b_k^i-b^i|)/k,\ k=400,410,\ldots,1000,\ i=1,2,$$

for the noise-free adaptive optimization algorithm in Fig. 3 with $\mu=0.0001,\eta=0.9,\delta=0.5$, which show that the convergence rate of the estimation algorithm is exponential since the trajectories stay negative and away from zero. Note that only the error trajectories for $i,j\in\{1,2\}$ are plotted here for simplicity.

In the second example, we consider a nonlinear system

$$x=\phi^\top(u)\theta=\left[e^{u^1},\ldots,e^{u^{20}}\right]\begin{bmatrix}\theta^1\\ \vdots\\ \theta^{20}\end{bmatrix},$$

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

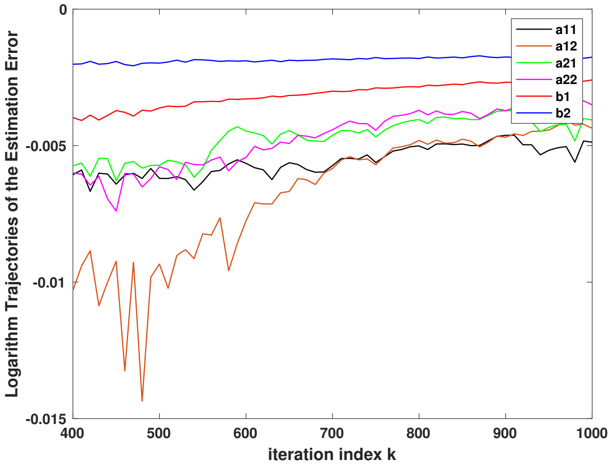


Fig. 3: Logarithm trajectories of the estimation error for $i, j \in \{1, 2\}$ of the noise-free adaptive optimization algorithm

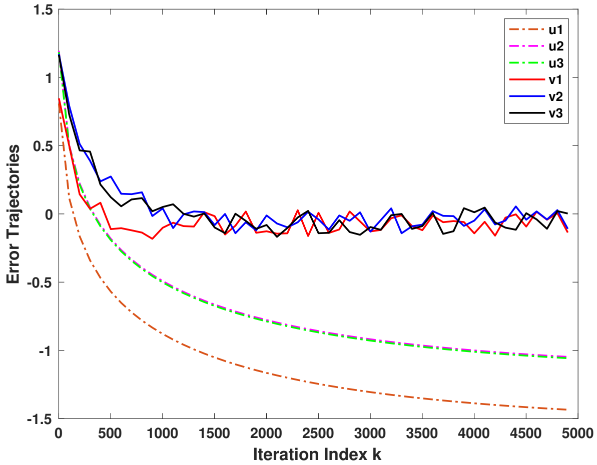


Fig. 4: Error trajectories for $i \in \{1, 2, 3\}$ of the noise-free adaptive optimization algorithm and (41)

where $\theta^i, i = 1, \ldots, 20$ are unknown parameters which need to be estimated during the optimization process. We consider the following performance index

$$J(x, u) = \frac{1}{2}x^2 + \frac{1}{2}\sum_{i=1}^{20}(u^i)^2, \tag{40}$$

and assume that the true values of the unknown parameters are randomly chosen from $[-5, 5]$. Note that the initial values of θ_0^i are chosen to be 1. Here we compare the noise-free adaptive optimization algorithm with the following algorithm without estimation:

$$u_{k+1}^i = u_k^i - \mu\left[\left(\sum_{j=1}^{20}\theta_0^j e^{u_k^j}\right)\theta_0^i e^{u_k^i} + u_k^i\right], \tag{41}$$

where the initial values are used to calculate the optimal solution. Let $\mu = 0.0001, \eta = 0.9, \delta = 0.1$. Then, we can obtain the error trajectories

$$v_k^i - u^{i,*}, \quad i = 1, \ldots, 20, \quad k = 0, 100, \ldots, 5000,$$

for the noise-free adaptive optimization algorithm, and

$$u_k^i - u^{i,*}, \quad i = 1, \ldots, 20, \quad k = 0, 100, \ldots, 5000,$$

for (41), where $u^{i,*}$ is the optimal value of u^i. To make the figures be more concise, we only plot the error trajectories for three elements, i.e., $i \in \{1, 2, 3\}$. Note that we can obtain similar results for the remaining ones.

The lower three lines in Fig. 4 are the error trajectories of (41), which are large because the values of the parameters used are not correct. The upper three lines are the error trajectories of the noise-free adaptive optimization algorithm, which fluctuate randomly around 0 because of the perturbations of δ_k.

For different δ, we also consider the error trajectories

$$v_k^i - u^{i,*}, \quad i = 1, \ldots, 20, \quad k = 0, 300, \ldots, 50000,$$

for the noise-free adaptive optimization algorithm in Fig. 5 with $\delta = 0, 0.05, 0.1$, which shows that as δ increases, the error on the optimal solution decreases, but the magnitude of random fluctuations increases.

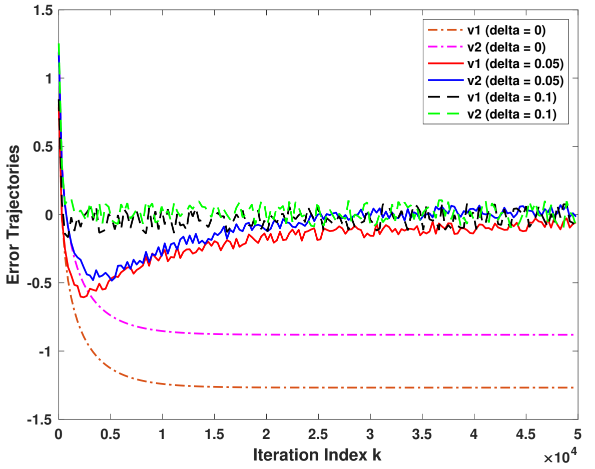


Fig. 5: Error trajectories for $i \in \{1, 2\}$ of the noise-free adaptive optimization algorithm with $\delta = 0, 0.05, 0.1$

VI. Concluding Remarks

This paper investigated constrained optimization problems with parameter uncertainties. A novel method of adding stochastic dither signals into the system has been introduced for achieving parameter identification and optimization simultaneously. Adaptive optimization algorithms have been developed, and their convergence and convergence rates have been established. The tradeoff among the step size, the stochastic dither magnitude, the error on the estimate, the error on the optimal solution, and convergence rate were presented quantitatively.

Extensions to non-smooth optimization problems and more general constraint models are possible and worth investigation. The results of this paper are generic and can be applied to different application domains. It will be highly valuable to apply

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

the results to emerging technical areas of network systems to verify their feasibility and understand their limitations.

VII. Appendix

Proof of *Theorem 3.2*: By (4a), we know that

$$u_{k+1} = u_k - \mu\nabla_u J_1(u_k,\theta_k) = u_k - \mu\nabla_u J_1(u_k,\theta) + \ell_k, \quad (42)$$

where $\ell_k = \mu[\nabla_u J_1(u_k,\theta) - \nabla_u J_1(u_k,\theta_k)]$.

By the Lipschitz continuity of $\nabla_u J_1(u,\theta)$, we know that there exists a constant $L > 0$ such that

$$\begin{aligned} & J_1(u_{k+1},\theta) - J_1(u_{k+1}-\ell_k,\theta) \\ \le & \langle \ell_k, \nabla_u J_1(u_{k+1}-\ell_k,\theta)\rangle + \frac{L}{2}\|\ell_k\|^2, \end{aligned} \quad (43)$$

and

$$\begin{aligned} & J_1(u_{k+1}-\ell_k,\theta) - J_1(u_k,\theta) \\ = & J_1(u_k - \mu\nabla_u J_1(u_k,\theta),\theta) - J_1(u_k,\theta) \\ \le & \langle -\mu\nabla_u J_1(u_k,\theta), \nabla_u J_1(u_k,\theta)\rangle + \frac{L\mu^2}{2}\|\nabla_u J_1(u_k,\theta)\|^2 \\ = & -\mu\left(1-\frac{L\mu}{2}\right)\|\nabla_u J_1(u_k,\theta)\|^2, \end{aligned} \quad (44)$$

and

$$\begin{aligned} & \|\nabla_u J_1(u_{k+1}-\ell_k,\theta) - \nabla_u J_1(u_k,\theta)\| \\ = & \|\nabla_u J_1(u_k - \mu\nabla_u J_1(u_k,\theta),\theta) - \nabla_u J_1(u_k,\theta)\| \\ \le & L\mu\|\nabla_u J_1(u_k,\theta)\|. \end{aligned} \quad (45)$$

Thus, if $0 < \mu \le \frac{1}{L}$, we have by (43)-(45) that

$$\begin{aligned} & J_1(u_{k+1},\theta) - J_1(u_k,\theta) \\ = & J_1(u_{k+1},\theta) - J_1(u_{k+1}-\ell_k,\theta) \\ & + J_1(u_{k+1}-\ell_k,\theta) - J_1(u_k,\theta) \\ \le & \langle \ell_k, \nabla_u J_1(u_{k+1}-\ell_k,\theta)\rangle + \frac{L}{2}\|\ell_k\|^2 \\ & - \mu\left(1-\frac{L\mu}{2}\right)\|\nabla_u J_1(u_k,\theta)\|^2 \\ \le & (1+L\mu)\|\ell_k\|\cdot\|\nabla_u J_1(u_k,\theta)\| + \frac{L}{2}\|\ell_k\|^2 \\ & - \mu\left(1-\frac{L\mu}{2}\right)\|\nabla_u J_1(u_k,\theta)\|^2 \\ \le & 2\|\ell_k\|\cdot\|\nabla_u J_1(u_k,\theta)\| + \frac{L}{2}\|\ell_k\|^2 - \frac{\mu}{2}\|\nabla_u J_1(u_k,\theta)\|^2 \\ \le & \left(\frac{2}{\mu}\|\ell_k\|^2 + \frac{\mu}{2}\|\nabla_u J_1(u_k,\theta)\|^2\right) \\ & + \frac{L}{2}\|\ell_k\|^2 - \frac{\mu}{2}\|\nabla_u J_1(u_k,\theta)\|^2 \\ = & \left(\frac{2}{\mu}+\frac{L}{2}\right)\|\ell_k\|^2, \end{aligned} \quad (46)$$

which implies that

$$J_1(u_k,\theta) \le J_1(u_0,\theta) + \left(\frac{2}{\mu}+\frac{L}{2}\right)\sum_{i=0}^{k-1}\|\ell_i\|^2. \quad (47)$$

Since the function $\nabla_u J_1(u,\theta)$ is Lipschitz continuous w.r.t some constant $L > 0$, then by (17),

$$\|\ell_k\| \le \mu L\|\theta_k - \theta\| \le \mu L M\gamma^k \ \text{ w.p.1}, \quad (48)$$

where $\gamma \in (0,1)$ and $M > 0$ are two constants. Thus, we know that

$$\sum_{i=0}^{k}\|\ell_i\| \le \mu L M\sum_{i=0}^{k}\gamma^i = \frac{\mu L M(1-\gamma^{k+1})}{1-\gamma} \ \text{ w.p.1.}$$

Therefore, there exists a constant $M_1 > 0$ such that $\sum_{i=0}^{k}\|\ell_i\| \le M_1$, and by (47), we know that

$$J_1(u_k,\theta) \le J_1(u_0,\theta) + \left(\frac{2}{\mu}+\frac{L}{2}\right)M_1. \quad (49)$$

which is bounded. Then since J_1 is level bounded, there exists a constant c_0 such that

$$\|u_k - u^*\| \le c_0, \quad \forall k \ge 0, \quad (50)$$

By the convexity and smoothness of $J_1(u,\theta)$, we know that

$$\begin{aligned} & J_1(u_k,\theta) - J_1(u^*,\theta) + \frac{1}{2L}\|\nabla_u J_1(u_k,\theta)\|^2 \\ \le & \langle u_k - u^*, \nabla_u J_1(u_k,\theta)\rangle \\ = & \frac{1}{2\mu}(\|u_k-u^*\|^2 + \mu^2\|\nabla_u J_1(u_k,\theta)\|^2 \\ & - \|u_k - u^* - \mu\nabla_u J_1(u_k,\theta)\|^2) \\ = & \frac{1}{2\mu}(\|u_k-u^*\|^2 - \|u_{k+1}-u^*-\ell_k\|^2) + \frac{\mu}{2}\|\nabla_u J_1(u_k,\theta)\|^2 \\ \le & \frac{1}{2\mu}(\|u_k-u^*\|^2 - \|u_{k+1}-u^*\|^2 + 2\|\ell_k\|\cdot\|u_{k+1}-u^*\|) \\ & + \frac{\mu}{2}\|\nabla_u J_1(u_k,\theta)\|^2, \end{aligned} \quad (51)$$

which implies that when $0 < \mu \le \frac{1}{L}$,

$$\begin{aligned} & J_1(u_k,\theta) - J_1(u^*,\theta) \\ \le & \frac{1}{2\mu}(\|u_k-u^*\|^2 - \|u_{k+1}-u^*\|^2 + 2\|\ell_k\|\cdot\|u_{k+1}-u^*\|) \\ & + \frac{1}{2}\left(\mu - \frac{1}{L}\right)\|\nabla_u J_1(u_k,\theta)\|^2 \\ \le & \frac{1}{2\mu}(\|u_k-u^*\|^2 - \|u_{k+1}-u^*\|^2) + \frac{1}{\mu}\|\ell_k\|\cdot\|u_{k+1}-u^*\| \\ \le & \frac{1}{2\mu}(\|u_k-u^*\|^2 - \|u_{k+1}-u^*\|^2) + \frac{c_0}{\mu}\|\ell_k\|. \end{aligned} \quad (52)$$

Therefore, we have

$$\begin{aligned} & \sum_{i=0}^{k}(J_1(u_i,\theta) - J_1(u^*,\theta)) \\ \le & \frac{1}{2\mu}\|u_0-u^*\|^2 + \sum_{i=0}^{k}\frac{c_0}{\mu}\|\ell_i\|. \end{aligned} \quad (53)$$

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

Also, by (46), we know that

$$\begin{aligned}&(k+1)J_1(u_k,\theta)-\sum_{i=0}^{k}J_1(u_i,\theta)\\ =&\sum_{i=0}^{k}(J_1(u_k,\theta)-J_1(u_i,\theta))\\ \le&\left(\frac{2}{\mu}+\frac{L}{2}\right)\sum_{i=0}^{k-1}\sum_{j=i}^{k-1}\|\ell_j\|^2.\end{aligned}\tag{54}$$

Thus, we have by (53) and (54) that

$$\begin{aligned}&J_1(u_k,\theta)-J_1(u^*,\theta)\\ =&\frac{1}{k+1}\sum_{i=0}^{k}(J_1(u_i,\theta)-J_1(u^*,\theta))\\ &+\frac{1}{k+1}\sum_{i=0}^{k}(J_1(u_k,\theta)-J_1(u_i,\theta))\\ \le&\frac{1}{k+1}\left(\frac{1}{2\mu}\|u_0-u^*\|^2+\sum_{i=0}^{k}\frac{c_0}{\mu}\|\ell_i\|\right)\\ &+\frac{1}{k+1}\left(\frac{2}{\mu}+\frac{L}{2}\right)\sum_{i=0}^{k-1}\sum_{j=i}^{k-1}\|\ell_j\|^2.\end{aligned}\tag{55}$$

and

$$\begin{aligned}\sum_{i=0}^{k-1}\sum_{j=i}^{k-1}\|\ell_j\|^2\le&\mu^2L^2M^2\sum_{i=0}^{k-1}\sum_{j=i}^{k-1}\gamma^{2j}\\ =&\mu^2L^2M^2\frac{1+k\gamma^{2(k+1)}-(k+1)\gamma^{2k}}{(1-\gamma^2)^2}\quad \text{w.p.1.}\end{aligned}$$

Combining these with (55), it is easy to see that

$$\begin{aligned}&J_1(u_k,\theta)-J_1(u^*,\theta)\\ \le&\frac{\|u_0-u^*\|^2}{2\mu(k+1)}+\frac{c_0LM}{1-\gamma}\cdot\frac{(1-\gamma^{k+1})}{k+1}\\ &+\frac{(4+\mu L)\mu L^2M^2}{(1-\gamma^2)^2}\cdot\frac{1+k\gamma^{2(k+1)}-(k+1)\gamma^{2k}}{k+1}\\ \le&O(1/k)\quad \text{w.p.1,}\end{aligned}\tag{56}$$

which completes the proof.

Proof of *Theorem 3.3*: Let $e_k=\|\theta_k-\theta\|$ and $r_k=\|u_k-u^*\|$. To prove convergence of u_k, define a Lyapunov candidate:

$$V_k=r_k^2.\tag{57}$$

Hence, we can obtain that

$$\begin{aligned}&r_{k+1}^2-r_k^2\\ =&\langle u_{k+1}-u^*,u_{k+1}-u^*\rangle-\langle u_k-u^*,u_k-u^*\rangle\\ =&\langle u_{k+1}+u_k-2u^*,u_{k+1}-u_k\rangle\\ =&\langle 2(u_k-u^*)-\mu\nabla_uJ_1(u_k,\theta_k),-\mu\nabla_uJ_1(u_k,\theta_k)\rangle\\ =&-2\mu\langle u_k-u^*,\nabla_uJ_1(u_k,\theta_k)\rangle\\ &+\mu^2\|\nabla_uJ_1(u_k,\theta_k)\|^2.\end{aligned}\tag{58}$$

Since the function $\nabla_uJ_1(u,\theta)$ is Lipschitz continuous w.r.t some constant $L>0$, we have

$$\|\nabla_uJ_1(u_k,\theta_k)-\nabla_uJ_1(u_k,\theta)\|\le Le_k,\tag{59}$$

and

$$\|\nabla_uJ_1(u_k,\theta)-\nabla_uJ_1(u^*,\theta)\|\le Lr_k.\tag{60}$$

Since $J_1(u,\theta)$ is strongly convex in u, then there exists a constant $c>0$ such that

$$\begin{aligned}0\ge&J_1(u^*,\theta)-J_1(u_k,\theta)\\ \ge&\langle u^*-u_k,\nabla_uJ_1(u_k,\theta)\rangle+\frac{c}{2}\|u^*-u_k\|^2.\end{aligned}\tag{61}$$

Thus, we have

$$\begin{aligned}&-2\mu\langle u_k-u^*,\nabla_uJ_1(u_k,\theta_k)\rangle\\ =&-2\mu\langle u_k-u^*,\nabla_uJ_1(u_k,\theta)\rangle\\ &-2\mu\langle u_k-u^*,\nabla_uJ_1(u_k,\theta_k)-\nabla_uJ_1(u_k,\theta)\rangle\\ \le&-2\mu\left(J_1(u_k,\theta)-J_1(u^*,\theta)+\frac{c}{2}r_k^2\right)+2\mu Le_kr_k\\ \le&-c\mu r_k^2+2\mu Le_kr_k.\end{aligned}\tag{62}$$

Moreover, we know that

$$\begin{aligned}&\mu^2\|\nabla_uJ_1(u_k,\theta_k)\|^2\\ \le&\mu^2(\|\nabla_uJ_1(u_k,\theta_k)-\nabla_uJ_1(u_k,\theta)\|\\ &+\|\nabla_uJ_1(u_k,\theta)-\nabla_uJ_1(u^*,\theta)\|)^2\\ \le&\mu^2L^2(e_k+r_k)^2.\end{aligned}\tag{63}$$

Thus, we have

$$\begin{aligned}r_{k+1}^2-r_k^2\le&-c\mu r_k^2+2\mu Le_kr_k+\mu^2L^2(e_k+r_k)^2\\ =&-\mu(c-\mu L^2)r_k^2\\ &+\mu L(2r_k+2\mu Lr_k+\mu Le_k)e_k\\ =&-\mu(c-\mu L^2)r_k^2+\mu^2L^2e_k^2\\ &+2\mu L(1+\mu L)r_ke_k\\ \le&-\mu(c-\mu L^2)r_k^2+\mu^2L^2e_k^2\\ &+\mu L(1+\mu L)(\varepsilon r_k^2+\frac{1}{\varepsilon}e_k^2)\\ =&-\mu[c-\mu L^2-L(1+\mu L)\varepsilon]r_k^2\\ &+\left[\mu^2L^2+\frac{\mu L(1+\mu L)}{\varepsilon}\right]e_k^2\\ =&-\mu[(c-L\varepsilon)-\mu L^2(1+\varepsilon)]r_k^2\\ &+\mu L\left(\mu L+\frac{1+\mu L}{\varepsilon}\right)e_k^2,\end{aligned}\tag{64}$$

where $\varepsilon>0$ can be any constant. Here we should choose $\varepsilon>0$ and $\mu>0$ such that

$$\begin{cases}c-L\varepsilon>0,\\ \mu L^2(1+\varepsilon)<c-L\varepsilon,\\ \mu[(c-L\varepsilon)-\mu L^2(1+\varepsilon)]<1.\end{cases}$$

Thus, if we choose

$$\begin{cases}0<\varepsilon<\frac{c}{L},\\ 0<\mu<\frac{c-L\varepsilon}{L^2(1+\varepsilon)},\quad \text{if}\ \ (c-L\varepsilon)^2<4L^2(1+\varepsilon),\\ 0<\mu<\frac{(c-L\varepsilon)-\sqrt{(c-L\varepsilon)^2-4L^2(1+\varepsilon)}}{2L^2(1+\varepsilon)},\\ \qquad \text{if}\ \ (c-L\varepsilon)^2\ge 4L^2(1+\varepsilon).\end{cases}\tag{65}$$

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

then there exists a constant $0 < \xi = \mu[(c-L\varepsilon)-\mu L^2(1+\varepsilon)] < 1$ such that

$$\begin{aligned} r_{k+1}^2 \leq&(1-\xi)r_k^2 + \omega e_k^2 \\ \leq&(1-\xi)^{k+1}r_0^2 + \omega\sum_{j=0}^{k}(1-\xi)^j e_{k-j}^2, \end{aligned} \tag{66}$$

where $\omega = \mu L\big(\mu L + \frac{1+\mu L}{\varepsilon}\big)$. Thus, by $\mathbb{E}[e_k^2] \leq M\gamma^k$ where $\gamma = \lambda^{2\eta\alpha_2} \in (0,1)$, we have

$$\begin{aligned} \mathbb{E}[r_{k+1}^2] \leq&(1-\xi)^{k+1}\mathbb{E}[r_0^2] + \omega\sum_{j=0}^{k}(1-\xi)^j\mathbb{E}[e_{k-j}^2] \\ \leq&(1-\xi)^{k+1}\mathbb{E}[r_0^2] + \omega M\sum_{j=0}^{k}(1-\xi)^j\gamma^{k-j}, \end{aligned}$$

which implies that

$$\frac{\mathbb{E}[r_{k+1}^2]}{(1-\xi)^{k+1}} \leq\mathbb{E}[r_0^2] + \frac{\omega M}{1-\xi}\sum_{j=0}^{k}\frac{\gamma^{k-j}}{(1-\xi)^{k-j}}. \tag{67}$$

Choosing μ such that $1-\xi > \gamma$. Thus, we know that

$$\sum_{j=0}^{k}\frac{\gamma^{k-j}}{(1-\xi)^{k-j}} \leq \frac{1-\xi}{1-\xi-\gamma}.$$

Combining with (67), it is easy to see that

$$\mathbb{E}[r_{k+1}^2] \leq\bigg(\mathbb{E}[r_0^2] + \frac{\omega M}{1-\xi-\gamma}\bigg)(1-\xi)^{k+1} \to 0, \tag{68}$$

as $k \to \infty$ holds, where $1-\xi \in (0,1)$ is the exponential convergence rate. In a similar way,

$$u_k \to u^* \quad \text{w.p.1 exponentially fast.} \tag{69}$$

Proof of *Proposition 3.1*: When $h = 1$, by the property of δ_k, we have

$$\begin{aligned} &\mathbb{E}\bigg[\frac{\phi(v_{k+1})\phi^\top(v_{k+1})}{1+\|\phi(v_{k+1})\|^2}|\mathcal{F}_k\bigg] \\ =&\mathbb{E}\bigg[\frac{(u_{k+1}+\delta_{k+1})(u_{k+1}+\delta_{k+1})^\top}{1+\|u_{k+1}+\delta_{k+1}\|^2}|\mathcal{F}_k\bigg] \\ =&\mathbb{E}\bigg[\frac{u_{k+1}u_{k+1}^\top+\delta_{k+1}u_{k+1}^\top+u_{k+1}\delta_{k+1}^\top+\delta_{k+1}\delta_{k+1}^\top}{1+\|u_{k+1}+\delta_{k+1}\|^2}|\mathcal{F}_k\bigg] \\ =&\mathbb{E}\bigg[\frac{u_{k+1}u_{k+1}^\top+\delta_{k+1}\delta_{k+1}^\top}{1+\|u_{k+1}+\delta_{k+1}\|^2}|\mathcal{F}_k\bigg] \\ \geq&\mathbb{E}\bigg[\frac{\delta_{k+1}\delta_{k+1}^\top}{1+\|u_{k+1}+\delta_{k+1}\|^2}|\mathcal{F}_k\bigg] \\ \geq&\mathbb{E}\bigg[\frac{\delta_{k+1}\delta_{k+1}^\top}{1+\|u_{k+1}\|^2+2\|u_{k+1}\|\cdot\|\delta_{k+1}\|+\|\delta_{k+1}\|^2}|\mathcal{F}_k\bigg] \\ \geq&\mathbb{E}\bigg[\frac{\delta_{k+1}\delta_{k+1}^\top}{1+\bar{u}^2+2\bar{u}\delta+\delta^2}|\mathcal{F}_k\bigg] \\ =&a\mathbb{E}[\delta_{k+1}\delta_{k+1}^\top] \\ =&\frac{a\delta^2}{3}I_m, \end{aligned} \tag{70}$$

where $a = \frac{1}{1+(\bar{u}+\delta)^2}$, and $\bar{u} = \sup_k\|u_k\|$ is bounded by (50). Then we know that *Assumption 3.1* is satisfied with $h = 1$, and $\lambda = 1 - \frac{\delta^2}{3[1+(\bar{u}+\delta)^2]} \in (0,1)$.

Proof of *Proposition 3.2*: When $h = 1$, we have

$$\begin{aligned} &\mathbb{E}\bigg[\frac{\phi(v_{k+1})\phi^\top(v_{k+1})}{1+\|\phi(v_{k+1})\|^2}|\mathcal{F}_k\bigg] \\ =&\mathbb{E}\Bigg[\frac{\begin{bmatrix}u_{k+1}+\delta_{k+1}\\1\end{bmatrix}[u_{k+1}^\top+\delta_{k+1}^\top,1]}{1+\|\begin{bmatrix}u_{k+1}+\delta_{k+1}\\1\end{bmatrix}\|^2}|\mathcal{F}_k\Bigg] \\ \geq&\mathbb{E}\Bigg[\frac{\begin{bmatrix}(u_{k+1}+\delta_{k+1})(u_{k+1}^\top+\delta_{k+1}^\top) & u_{k+1}+\delta_{k+1}\\ u_{k+1}^\top+\delta_{k+1}^\top & 1\end{bmatrix}}{2+\|u_{k+1}+\delta_{k+1}\|^2}|\mathcal{F}_k\Bigg] \\ \geq&b\mathbb{E}\bigg[\begin{bmatrix}u_{k+1}u_{k+1}^\top+\delta_{k+1}\delta_{k+1}^\top & u_{k+1}\\ u_{k+1}^\top & 1\end{bmatrix}|\mathcal{F}_k\bigg] \\ =&b\mathbb{E}\bigg[\begin{bmatrix}u_{k+1}u_{k+1}^\top+\frac{\delta^2}{3}I_m & u_{k+1}\\ u_{k+1}^\top & 1\end{bmatrix}|\mathcal{F}_k\bigg] \\ \overset{\triangle}{=}&b\mathbb{E}[M_{k+1}|\mathcal{F}_k], \end{aligned} \tag{71}$$

where $b = \frac{1}{2+(\bar{u}+\delta)^2}$.

Now we need to study the smallest eigenvalue of the above matrix M_{k+1}. From the unique structure of the block matrix [25], and by letting

$$\begin{aligned} &|sI_{m+1} - M_{k+1}| \\ =&\bigg|sI_{m+1} - \begin{bmatrix}u_{k+1}u_{k+1}^\top+\frac{\delta^2}{3}I_m & u_{k+1}\\ u_{k+1}^\top & 1\end{bmatrix}\bigg| \\ =&\bigg|\begin{bmatrix}sI_m-(u_{k+1}u_{k+1}^\top+\frac{\delta^2}{3}I_m) & -u_{k+1}\\ -u_{k+1}^\top & s-1\end{bmatrix}\bigg| \\ =&|(s-1)(sI_m-(u_{k+1}u_{k+1}^\top+\frac{\delta^2}{3}I_m))-u_{k+1}u_{k+1}^\top| \\ =&|(s-1)(s-\frac{\delta^2}{3})I_m - su_{k+1}u_{k+1}^\top| = 0, \end{aligned} \tag{72}$$

we can study the eigenvalues of M_{k+1}. Since the eigenvalues of $u_{k+1}u_{k+1}^\top$ are 0 and $\lambda_{k+1}^u = u_{k+1}^\top u_{k+1} \geq 0$, the smallest eigenvalue of M_{k+1} is the minimum solution of

$$(s-1)(s-\frac{\delta^2}{3}) = s\lambda_{k+1}^u,$$

which can be explicitly solved from the following polynomial equation:

$$s^2 - (1+\frac{\delta^2}{3}+\lambda_{k+1}^u)s + \frac{\delta^2}{3} = 0.$$

Note that if

$$(1+\frac{\delta^2}{3}+\lambda_{k+1}^u)^2 - \frac{4\delta^2}{3} \geq 0,$$

i.e.,

$$1+\frac{\delta^2}{3}+\lambda_{k+1}^u \geq \frac{2\delta}{\sqrt{3}}, \tag{73}$$

the above equation has real solutions. Since

$$(\frac{\delta}{\sqrt{3}}-1)^2 + \lambda_{k+1}^u \geq 0,$$

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

we know that (73) holds. Thus, the minimum solution is

$$\begin{aligned}
&s_{\min} \\
=&\frac{1+\frac{\delta^2}{3}+\lambda^u_{k+1}-\sqrt{(1+\frac{\delta^2}{3}+\lambda^u_{k+1})^2-\frac{4\delta^2}{3}}}{2} \\
=&\frac{1+\frac{\delta^2}{3}+\lambda^u_{k+1}-\sqrt{(1-\frac{\delta^2}{3})^2+2(1+\frac{\delta^2}{3})\lambda^u_{k+1}+(\lambda^u_{k+1})^2}}{2},
\end{aligned}$$

which is monotone decreasing when λ^u_{k+1} increases. Since λ^u_{k+1} is uniformly bounded by $\bar{u}^2$, and denote

$$\Delta \triangleq \frac{1+\frac{\delta^2}{3}+\bar{u}^2-\sqrt{(1-\frac{\delta^2}{3})^2+2(1+\frac{\delta^2}{3})\bar{u}^2+\bar{u}^4}}{2}, \quad (74)$$

then we know that the minimum eigenvalue satisfies $s_{\min} \geq \Delta > 0$. Hence,

$$\min_{k,u_k} \lambda_{\min}(M_k) \geq s_{\min} \geq \Delta > 0 \quad (75)$$

holds. Then we know that *Assumption 3.1* is satisfied with $h = 1$, and $\lambda = 1 - \frac{\Delta}{2+(\bar{u}+\delta)^2} \in (0, 1)$, which is monotonically decreasing when δ decreases.

Proof of *Proposition 3.3*: Since $\phi(u)$ is continuously twice differentiable and $\nabla\phi(u)$ is full rank (rank p) uniformly over all u, we have

$$x_k = \phi^\top(v_k)\theta = \phi^\top(u_k+\delta_k)\theta = \phi^\top(u_k+\delta\beta_k)\theta, \quad (76)$$

where $\delta > 0$ is a small scaling factor can be selected later, and $\beta_k \in \mathbb{R}^m$ is independently and identically distributed with each element $\beta^i_k \sim U(-1, 1)$.

For a given u_k and small δ, one may expand $\phi(v_k) = \phi(u_k + \delta_k)$ into the following form:

$$\begin{aligned}
\phi(v_k) =&\phi(u_k+\delta_k) = \phi(u_k+\delta\beta_k) \\
=&\phi(u_k) + \delta\nabla\phi(u_k)\beta_k + \delta^2\rho_k. \quad (77)
\end{aligned}$$

Since u_k is uniformly bounded and the second derivative of $\phi(u)$ is continuous, $\phi(u_k), \nabla\phi(u_k)$, and ρ_k are also uniformly bounded, i.e., $\|\phi(u_k)\| \leq \bar{\phi} < \infty$, $\|\nabla\phi(u_k)\| \leq \bar{\phi}_\nabla < \infty$, and $\|\rho_k\| \leq \bar{\rho} < \infty$ hold for some constants $\bar{\phi}, \bar{\phi}_\nabla$, and $\bar{\rho}$. When $h = 1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\phi(v_{k+1})\phi^\top(v_{k+1})}{1+\|\phi(v_{k+1})\|^2}|\mathcal{F}_k\right] \\
=&\mathbb{E}\left[\frac{\phi(u_{k+1}+\delta\beta_{k+1})\phi^\top(u_{k+1}+\delta\beta_{k+1})}{1+\|\phi(u_{k+1}+\delta\beta_{k+1})\|^2}|\mathcal{F}_k\right] \\
\geq&\mathbb{E}\left[\frac{\phi(u_{k+1}+\delta\beta_{k+1})\phi^\top(u_{k+1}+\delta\beta_{k+1})}{1+(\bar{\phi}+\delta\bar{\phi}_\Delta+\delta^2\bar{\rho})^2}|\mathcal{F}_k\right] \\
=&c\mathbb{E}[(\phi(u_{k+1})+\delta\nabla\phi(u_{k+1})\beta_{k+1}+\delta^2\rho_{k+1}) \\
&\cdot(\phi(u_{k+1})+\delta\nabla\phi(u_{k+1})\beta_{k+1}+\delta^2\rho_{k+1})^\top|\mathcal{F}_k] \\
=&c\mathbb{E}[(\phi(u_{k+1})+\delta^2\rho_{k+1})(\phi(u_{k+1})+\delta^2\rho_{k+1})^\top \\
&+\delta[(\phi(u_{k+1})+\delta^2\rho_{k+1})(\nabla\phi(u_{k+1})\beta_{k+1})^\top \\
&+(\nabla\phi(u_{k+1})\beta_{k+1})(\phi(u_{k+1})+\delta^2\rho_{k+1})^\top] \\
&+\delta^2\nabla\phi(u_{k+1})\beta_{k+1}\beta^\top_{k+1}(\nabla\phi(u_{k+1}))^\top|\mathcal{F}_k] \\
\geq&c\mathbb{E}[\delta(\phi(u_{k+1})(\nabla\phi(u_{k+1})\beta_{k+1})^\top \\
&+\nabla\phi(u_{k+1})\beta_{k+1}\phi^\top(u_{k+1})) \\
&+\delta^3(\rho_{k+1}(\nabla\phi(u_{k+1})\beta_{k+1})^\top+\nabla\phi(u_{k+1})\beta_{k+1}\rho^\top_{k+1}) \\
&+\delta^2\nabla\phi(u_{k+1})\beta_{k+1}\beta^\top_{k+1}(\nabla\phi(u_{k+1}))^\top|\mathcal{F}_k] \\
=&c\mathbb{E}[\delta^3(\rho_{k+1}(\nabla\phi(u_{k+1})\beta_{k+1})^\top+\nabla\phi(u_{k+1})\beta_{k+1}\rho^\top_{k+1}) \\
&+\delta^2\nabla\phi(u_{k+1})\beta_{k+1}\beta^\top_{k+1}(\nabla\phi(u_{k+1}))^\top|\mathcal{F}_k] \\
\geq&c\mathbb{E}[-\delta^3(\rho_{k+1}\rho^\top_{k+1}+\nabla\phi(u_{k+1})\beta_{k+1}\beta^\top_{k+1}(\nabla\phi(u_{k+1}))^\top) \\
&+\delta^2\nabla\phi(u_{k+1})\beta_{k+1}\beta^\top_{k+1}(\nabla\phi(u_{k+1}))^\top|\mathcal{F}_k] \\
=&c\mathbb{E}[-\delta^3\rho_{k+1}\rho^\top_{k+1} \\
&+\frac{\delta^2-\delta^3}{3}\nabla\phi(u_{k+1})(\nabla\phi(u_{k+1}))^\top|\mathcal{F}_k] \\
=&c\mathbb{E}[N_{k+1}|\mathcal{F}_k], \quad (78)
\end{aligned}$$

where $c = \frac{1}{1+(\bar{\phi}+\delta\bar{\phi}_\Delta+\delta^2\bar{\rho})^2}$, and $N_{k+1} = -\delta^3\rho_{k+1}\rho^\top_{k+1} + \frac{\delta^2-\delta^3}{3}\nabla\phi(u_{k+1})(\nabla\phi(u_{k+1}))^\top$.

Since $\nabla\phi(u_{k+1})(\nabla\phi(u_{k+1}))^\top$ is full rank uniformly over u, we know that there exists a constant $\alpha > 0$ such that

$$\min_{u_k} \lambda_{\min}\{\nabla\phi(u_{k+1})(\nabla\phi(u_{k+1}))^\top\} = \alpha.$$

This, together with $\bar{\rho}$, implies that

$$\begin{aligned}
\min_{k,u_k} \lambda_{\min}(N_k) \geq& -\delta^3\bar{\rho} + \frac{(\delta^2-\delta^3)\alpha}{3} \\
=&\frac{\alpha}{3}\delta^2 - \left(\frac{\alpha}{3}+\bar{\rho}\right)\delta^3.
\end{aligned}$$

Thus, there exists $0 < \delta \leq \frac{\alpha}{2\alpha+6\bar{\rho}} < 1$ such that

$$\min_{k,u_k} \lambda_{\min}(N_k) \geq \frac{\alpha}{6}\delta^2 \triangleq \Delta > 0.$$

Then we know that *Assumption 3.1* is satisfied with $h = 1$, and $\lambda = 1 - \frac{\Delta}{1+(\bar{\phi}+\delta\bar{\phi}_\Delta+\delta^2\bar{\rho})^2} \in (0, 1)$, which is monotonically decreasing when δ decreases.

Proof of *Theorem 4.1*: By (36) and denote

$$M_k = \frac{\log(k+\alpha_0)}{k+\alpha_0}, \ k \geq 0,$$

we know that $e^2_k \leq CM_k$ w.p.1, where $C > 0$ is a constant and $e_k = \|\theta_k - \theta\|$. By the property of the sequence M_k, there exists a constant C_0 such that $M_k \leq C_0M_0$. Thus, by *Theorem 3.3*, there exist constants $0 < \xi < 1$ and $\omega > 0$ such that

$$\begin{aligned}
r^2_{k+1} \leq&(1-\xi)r^2_k + \omega e^2_k \\
\leq&(1-\xi)r^2_k + \omega CM_k \\
\leq&(1-\xi)r^2_k + \omega CC_0M_0, \quad \text{w.p.1}, \quad (79)
\end{aligned}$$

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control

where $r_k = \|u_k - u^*\|$. Then,

$$\begin{aligned} r_{k+1}^2 \leq & (1-\xi)^{k+1} r_0^2 + \omega C C_0 M_0 \sum_{i=0}^{k} (1-\xi)^{k-i} \\ \leq & (1-\xi)^{k+1} r_0^2 + \frac{\omega C C_0}{\xi} M_0 \quad \text{w.p.1.} \end{aligned} \tag{80}$$

Note that

$$\begin{aligned} & (1-\xi)^{-\ell} r_{k+\ell}^2 - r_k^2 \\ = & \sum_{i=0}^{\ell-1} [(1-\xi)^{-(i+1)} r_{k+i+1}^2 - (1-\xi)^{-i} r_{k+i}^2] \\ = & \sum_{i=0}^{\ell-1} (1-\xi)^{-(i+1)} [r_{k+i+1}^2 - (1-\xi) r_{k+i}^2] \\ \leq & \omega C \sum_{i=0}^{\ell-1} (1-\xi)^{-(i+1)} M_{k+i} \quad \text{w.p.1.} \end{aligned}$$

From this, we have by (80) that

$$\begin{aligned} r_{k+\ell}^2 \leq & (1-\xi)^{\ell} r_k^2 + \omega C \sum_{i=0}^{\ell-1} (1-\xi)^{\ell-(i+1)} M_{k+i} \\ \leq & (1-\xi)^{k+\ell} r_0^2 + \frac{\omega C C_0}{\xi} M_0 (1-\xi)^{\ell} \\ & + \omega C M_k \sum_{i=0}^{\ell-1} (1-\xi)^{\ell-(i+1)} \\ \leq & (1-\xi)^{k+\ell} r_0^2 + \frac{\omega C C_0}{\xi} M_0 (1-\xi)^{\ell} + \frac{\omega C}{\xi} M_k \quad \text{w.p.1.} \end{aligned}$$

Note that for any $\varepsilon > 0$, there exists a constant k_0 such that

$$(1-\xi)^{k+\varepsilon k} r_0^2 + \frac{\omega C C_0}{\xi} M_0 (1-\xi)^{\varepsilon k} \leq \frac{\omega C}{\xi} M_k, \quad \forall k \geq k_0.$$

Therefore, for any $\varepsilon > 0, k \geq k_0(\varepsilon)$, we have

$$\begin{aligned} r_{k+\varepsilon k}^2 \leq & (1-\xi)^{k+\varepsilon k} r_0^2 + \frac{\omega C C_0}{\xi} M_0 (1-\xi)^{\varepsilon k} + \frac{\omega C}{\xi} M_k \\ \leq & \frac{2\omega C}{\xi} M_k \quad \text{w.p.1,} \end{aligned}$$

which completes the proof.

REFERENCES

[1] A. P. Ruszczyński, *Nonlinear Optimization*. New Jersey: Princeton University Press, 2006.

[2] D. P. Bertsekas, *Convex Optimization Algorithms*. New Hampshire: Athena Scientific Belmont, 2015.

[3] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[4] Y. Zhang, Z. Deng, and Y. Hong, "Distributed optimal coordination for multiple heterogeneous Euler–Lagrangian systems," *Automatica*, vol. 79, no. 5, pp. 207–213, 2017.

[5] S. Liang, L. Wang, and G. Yin, "Distributed smooth convex optimization with coupled constraints," *IEEE Transactions on Automatic Control*, vol. 65, no. 1, pp. 347 – 353, 2020.

[6] G. Zames, "Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses," *IEEE Transactions on Automatic Control*, vol. 26, no. 2, pp. 301–320, 1981.

[7] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, "State-space solutions to standard h_2 and h_∞ control problems," in *1988 American Control Conference*. IEEE, 1988, pp. 1691–1696.

[8] L. Y. Wang, G. G. Yin, J.-F. Zhang, and Y. Zhao, *System Identification with Quantized Observations*. Boston: Birkhäuser, 2010.

[9] H. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, ser. Stochastic Modelling and Applied Probability. New York: Springer-Verlag, 2003, vol. 35.

[10] K. J. Åström and B. Wittenmark, "On self tuning regulators," *Automatica*, vol. 9, no. 2, pp. 185–199, 1973.

[11] L. Guo, "Convergence and logarithm laws of self-tuning regulators," *Automatica*, vol. 31, no. 3, pp. 435–450, 1995.

[12] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Mathematical Programming*, vol. 146, no. 1-2, pp. 37–75, 2014.

[13] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and applications of robust optimization," *SIAM Review*, vol. 53, no. 3, pp. 464–501, 2011.

[14] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, vol. 59, no. 2, pp. 341–357, 2013.

[15] E. B. Kosmatopoulos, "Adaptive control design based on adaptive optimization principles," *IEEE Transactions on Automatic Control*, vol. 53, no. 11, pp. 2680–2685, 2008.

[16] ——, "An adaptive optimization scheme with satisfactory transient performance," *Automatica*, vol. 45, no. 3, pp. 716–723, 2009.

[17] E. B. Kosmatopoulos and A. Kouvelas, "Large scale nonlinear control system fine-tuning through learning," *IEEE Transactions on Neural Networks*, vol. 20, no. 6, pp. 1009–1023, 2009.

[18] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. 7, pp. 2121–2159, 2011.

[19] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Prentice-Hall Englewood Cliffs NJ, 1985.

[20] S. Haykin, *Adaptive filter theory*. Prentice-Hall, Inc., 1996.

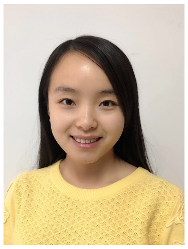
[21] V. Solo and X. Kong, *Adaptive signal processing algorithms*. Prentice-Hall, Inc., 1995.

[22] O. Macchi, *Adaptive processing: the least mean squares approach with applications in transmission*. New York: Wiley, 1995.

[23] L. Guo, "Stability of recursive stochastic tracking algorithms," *SIAM Journal on Control and Optimization*, vol. 32, no. 5, pp. 1195–1225, 1994.

[24] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *The Annals of Statistics*, vol. 10, no. 1, pp. 154–166, 1982.

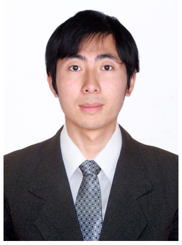
[25] J. R. Silvester, "Determinants of block matrices," *Mathematical Gazette, The Mathematical Association*, vol. 84, no. 501, pp. 460–467, 2000.



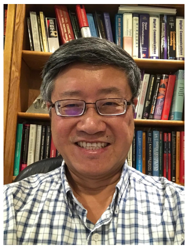
Siyu Xie received the B.S. degree in information and computing science (systems control) from Beijing University of Aeronautics and Astronautics in 2013, and Ph.D. degree in control theory from Academy of Mathematics and Systems Science, Chinese Academy of Sciences in 2018. She is currently a postdoctoral fellow at the Department of Electrical and Computer Engineering, Wayne State University, USA. She received the IEEE CSS Beijing Chapter Young Author Prize in 2015, China National Scholarship for graduate students in 2017, and the President Scholarship of Chinese Academy of Sciences in 2018. Her research interests include networked systems, distributed optimization problems for power systems, distributed adaptive filters, machine learning, compressive sensing and distributed control.

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3050438, IEEE Transactions on Automatic Control



Shu Liang received the B.E. degree in automatic control and the Ph.D. degree in engineering from the University of Science and Technology of China, Hefei, China, in 2010 and 2015, respectively. He was a postdoctoral fellow at Academy of Mathematics and Systems Science, Chinese Academy of Sciences from 2015 to 2017, and was a visiting scholar at Wayne State University from 2017 to 2018. He is currently an Associate Professor in the School of Automation and Electrical Engineering, University of Science and Technology Beijing, China. His research interests include distributed optimization, game theory and fractional-order system.



Le Yi Wang (S'85-M'89-SM'01-F'12) received the Ph.D. degree in electrical engineering from McGill University, Montreal, Canada, in 1990. Since 1990, he has been with Wayne State University, Detroit, Michigan, where he is currently a Professor in the Department of Electrical and Computer Engineering. His research interests are in the areas of complexity and information, system identification, robust control, information processing and learning, as well as medical, automotive, communications, power systems, and computer applications of control methodologies. He was a plenary speaker in many international conferences. He serves on the IFAC Technical Committee on Modeling, Identification and Signal Processing. He was an Associate Editor of the IEEE Transactions on Automatic Control and several other journals.

George Yin (S'87-M'87-SM'96-F'02) received the B.S. degree in mathematics from the University of Delaware in 1983, and the M.S. degree in electrical engineering and the Ph.D. degree in applied mathematics from Brown University in 1987. He joined Wayne State University in 1987, became Professor in 1996, and University Distinguished Professor in 2017. He moved to the University of Connecticut in 2020. His research interests include stochastic processes, stochastic systems theory, and applications. He served as Co-chair for a number of conferences, was on the Board of Directors of the American Automatic Control Council, and was Chair of the SIAM Activity Group on Control and Systems Theory. He is Editor-in-Chief of SIAM Journal on Control and Optimization. He was an Associate Editor of Automatica 1995-2011, IEEE Transactions on Automatic Control 1994-1998, and Senior Editor of IEEE Control Systems Letters 2017-2019. He is a Fellow of IFAC and a Fellow of SIAM.

Wen Chen (SM09) received the Ph.D. degree from Simon Fraser University, British Columbia, Canada in 2004. He was a postdoctoral researcher, from 2005 to 2007, at University of Louisiana at Lafayette, Louisiana, USA. In 2007, he became a Control Systems Engineer at Paton Controls and was hired by Triconex, Houston, Texas, USA in 2008. He joined Division of Engineering Technology, Wayne State University, USA, in 2009, as an Assistant Professor, where he is currently an Associate Professor. His research interests have been in the area of control and diagnosis of industrial systems.

0018-9286 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Wayne State University. Downloaded on March 10,2021 at 03:20:57 UTC from IEEE Xplore. Restrictions apply.