On the Fundamental Limit of Stochastic Gradient Identification Algorithm Under Non-Persistent Excitation ⋆

Senhan Yao† *, Longxu Zhang† *,****

* *State Key Laboratory of Mathematical Sciences, AMSS, Chinese Academy of Sciences, Beijing, 100190, China.*
** *School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China. (email: yaosenhan@amss.ac.cn, zhanglongxu@amss.ac.cn).*

Abstract: Stochastic gradient methods are of fundamental importance in system identification and machine learning, enabling online parameter estimation for large-scale and data-streaming processes. The stochastic gradient algorithm stands as a classical identification method that has been extensively studied for decades. Under non-persistent excitation, the best known convergence result requires the condition number of the Fisher information matrix to satisfy $\kappa(\sum_{i=1}^{n} \varphi_i \varphi_i^\top) = O((\log r_n)^\alpha)$, where $r_n = 1 + \sum_{i=1}^{n} \|\varphi_i\|^2$, with strong consistency guaranteed for $\alpha \leq 1/3$ but known to fail for $\alpha > 1$. This paper establishes that strong consistency in fact holds for the entire range $0 \leq \alpha < 1$, achieved through a novel algebraic framework that yields substantially sharper matrix norm bounds. Our result nearly resolves the four-decade-old conjecture of Chen and Guo (1986), bridging the theoretical gap from $\alpha \leq 1/3$ to nearly the entire feasible range.

Keywords: System identification, stochastic gradient algorithm, strong consistency, non-persistent excitation, fundamental limit

1. INTRODUCTION

How can we develop mathematical models for real-world physical processes from noisy observational data to support critical engineering tasks such as controller design, prediction, and fault diagnosis? This question lies at the heart of *system identification*, see Ljung (2010). In the era of artificial intelligence, characterized by large-scale data and online learning requirements, *stochastic gradient (SG)* algorithms have regained prominence. These methods support iterative, online processing of randomly perturbed data while ensuring convergent parameter estimation. They have also become the foundation for numerous algorithmic variants, including Adam, see Kingma and Ba (2015).

The SG method is closely related to *stochastic approximation (SA)*, a framework originating in the pioneering work of Robbins and Monro (1951). Subsequent milestones include the stochastic optimization approach of Kiefer and Wolfowitz (1952), the general convergence theorem of Dvoretzky (1956), asymptotic analyses by Chung (1954) and Sacks (1958), and the ODE method of Gladyshev (1965), which together established a rigorous mathematical foundation for this class of stochastic iterative algorithms.

The application of this theory to system identification emerged in the 1960s. Åström and Bohlin (1965) established a systematic *offline* identification framework using maximum likelihood estimation, introducing the pivotal concept of *persistent excitation (PE)*. Concurrently, Sakrison (1964) pioneered the application of stochastic approximation to *online* identification, shifting the focus from offline modeling to real-time estimation. During the 1970s, research evolved toward closed-loop systems. The groundbreaking work of Åström and Wittenmark (1972) initiated the theoretical analysis of stochastic adaptive control, while Ljung (1977) advanced the ODE method of Gladyshev (1965) to provide a unified tool for analyzing the asymptotic behavior of general stochastic algorithms.

Although SG algorithms possess well-established strong consistency under PE conditions, see Ljung (1977); Anderson and Taylor (1979); Chen (1981), their convergence behavior in the absence of such conditions warrants critical investigation. One motivation comes from adaptive control, see Goodwin et al. (1980), where *a decoupling between parameter estimation and optimal control is observed*: rigorously enforcing PE ensures strong consistency but may not yield an optimal controller, see Chen and Caines (1984); conversely, pursuing an optimal controller can compromise parameter consistency, see Becker et al. (1985). A second practical challenge stems from the fact that *many systems operate under insufficiently rich inputs*, as empirically observed in diverse domains such as quadrotor UAVs, see Chowdhary et al. (2012), neural networks, see Nar and Sastry (2019), and spherical parallel robots,

⋆ This research was supported by the National Natural Science Foundation of China under grant No. 12288201 and the National Key R&D Program of China under grant 2024YFC3307201.
† These authors contributed equally to this work.

see Rad et al. (2020). Without theoretical guarantees in such scenarios, inadequate excitation can lead to severely biased parameter estimates, unacceptably slow convergence, or even outright divergence of the identification algorithm.

The seminal works of Chen (1982) and Lai and Wei (1982) independently and concurrently broke new ground by systematically weakening PE condition required for the strong consistency of LS. Chen (1982) established that strong consistency holds if the condition number of the Fisher information matrix grows as

$$\kappa\left(\sum_{i=1}^{n}\phi_i\phi_i^\top\right) = O(r_n^\delta) \quad \text{a.s.}, \quad 0 \le \delta < \frac{1}{2} \tag{1}$$

where $r_n = 1 + \sum_{i=1}^{n}\|\phi_i\|^2$. Simultaneously, Lai and Wei, working within a general stochastic regression framework, derived that strong consistency holds if

$$\kappa\left(\sum_{i=1}^{n}\phi_i\phi_i^\top\right) = o\left(\frac{r_n}{\log r_n}\right) \quad \text{a.s.}, \tag{2}$$

Building upon this perspective, the Chen-Guo approach offers a profound understanding of SG algorithm under relaxed excitation. Chen and Guo (1985b) determines the convergence speed of SG algorithm in this non-PE case. Through a remarkable *sample-path-wise* analysis that relates strong consistency to the convergence of an instrumental matrix sequence $\Phi(n, 0)$, see Chen and Guo (1985a), under mild conditions on the noise (not necessarily i.i.d. or of bounded variance), Chen and Guo (1985b); Guo (1993); Chen and Guo (1986) showed that if

$$\kappa\left(\sum_{i=1}^{n}\phi_i\phi_i^\top\right) = O((\log r_n)^\alpha) \quad \text{a.s.}, \tag{3}$$

then:

- $\alpha \le 1/3$ ensures strong consistency.
- $\alpha > 1$ precludes strong consistency.

This stark dichotomy inevitably raises a fundamental question: *What are the fundamental limits of SG identification algorithms? To what extent can we rely on the results of SG algorithms?* More broadly, it challenges us to understand how much insight into a system can be gained from finite, noisy, and poorly-structured observations in a world replete with uncertainties.

Chen and Guo (1986) conjectured that the established sufficient condition is inherently conservative. *They posited that strong consistency should, in fact, be guaranteed for the entire range $\alpha \le 1$, with $\alpha = 1$ representing the ultimate "critical excitation" boundary for the algorithm.* For decades, the validity of this conjecture—and thus the characterization of the algorithm's fundamental limits within the gap $1/3 \le \alpha \le 1$ —has remained an open problem, representing a significant challenge in the theoretical foundations of SG identification.

The main contribution of this paper is a nearly complete resolution of this long-standing conjecture. Our results bridge the theoretical gap as follows:

- We prove that for the entire range $0 \le \alpha < 1$, the condition (3) *is indeed sufficient* for the almost sure convergence of the stochastic gradient algorithm to the true parameters, thereby providing a near-complete characterization of its behavioral landscape.
- To establish this result, we develop a novel *algebraic approach* that offers a more transparent and versatile framework for obtaining the requisite matrix norm bounds, superseding previous analytical obstacles.

The remainder of this paper is organized as follows: Section 2 revisits the problem formulation and Chen-Guo approach. Section 3 details our new algebraic approach bounding matrix products. Section 4 presents detailed integral estimates. Section 5 proves the main theorem, and Section 6 offers concluding remarks.

Notation: Throughout this paper, the logarithm log denotes the natural logarithm with base e. $A_n \lesssim B_n$ implies that there exists a positive constant C such that $A_n \le CB_n$. The Landau notation $O(\cdot)$ and $\Theta(\cdot)$ follow their standard asymptotic definitions. $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices, unless specified otherwise. $\|\cdot\|_F$ means the Frobenius norm. For a nonsingular matrix A, the condition number $\kappa(A)$ is defined by $\kappa(A) = \|A\|\|A^{-1}\|$.

2. PROBLEM FORMULATION

In this section, we reformulate the problem and briefy review Chen-Guo approach. Consider the MIMO system

$$\begin{aligned} y_n + A_1 y_{n-1} + \cdots + A_p y_{n-p} \\ = B_1 u_{n-1} + \cdots + B_q u_{n-q} + \varepsilon_n, \end{aligned} \tag{4}$$

where y_n and u_n denote the d-dimensional output and the l-dimensional input respectively, $A_i, i = 1, \ldots, p, B_j, j = 1, \ldots, q$ are the unknown matrices to be identified, ε_n is the d-dimensional system noise driven by a martingale difference sequence $\{w_n\}$, that is,

$$\varepsilon_n = w_n + C_1 w_{n-1} + \cdots + C_r w_{n-r}, \tag{5}$$

and w_n satisfies

$$\mathbb{E}(w_n|\mathcal{F}_{n-1}) = 0, \quad \forall n \ge 1, \tag{6}$$

where $\{\mathcal{F}_n\}$ is a family of nondecreasing σ-algebras defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $C_k, k = 1, 2, \ldots, r$ are unknown matrices.

Let z be the shift-back operator, define

$$C(z) = I + C_1 z + \cdots C_r z^r, \tag{7}$$

and set

$$\theta^\top = [-A_1 \cdots -A_p \; B_1 \cdots B_q \; C_1 \cdots C_r], \tag{8}$$

$$\begin{aligned} \varphi_n^\top = [y_{n-1}^\top, \cdots, y_{n-p+1}^\top, \\ u_{n-1}^\top, \cdots, u_{n-q+1}^\top, \\ \hat{w}_{n-1}^\top, \cdots, \hat{w}_{n-r+1}^\top], \end{aligned} \tag{9}$$

where we use the priori estimate of w_n, which is also called the estimated residuals:

$$\hat{w}_{n-1} := y_n - \theta_{n-1}^\top \varphi_{n-1}. \tag{10}$$

It is clear that φ_n is defined recursively, which is $\mathcal{F}_n$-measurable. The estimation problem is cast as a linear regression problem:

$$y_{n+1} = \theta^\top \varphi_n + w_{n+1}. \tag{11}$$

Denote by θ_n the estimate for θ at time n. Given any determined initial values θ_0 and φ_0, we can state the SG algorithm as follows:

$$\theta_{n+1} = \theta_n + \frac{\varphi_n}{r_n}(y_{n+1}^\top - \varphi_n^\top \theta_n), \tag{12}$$

where r_n is given by

$$r_n = 1 + \sum_{i=1}^{n} \|\varphi_i\|^2, \quad r_0 = 1. \tag{13}$$

Chen and Guo introduce the following instrumental transition matrix:

$$\Phi(n+1, i) = (I - A_n)\Phi(n, i), \quad n \geq i, \tag{14}$$

$$\Phi(i, i) = I, \tag{15}$$

where we denote

$$A_n = \frac{\varphi_n \varphi_n^\top}{r_n}. \tag{16}$$

Chen and Guo leverage the *strictly positive real (SPR)* condition to bound the error introduced by the estimated residuals, thereby ensuring that the parameter error dynamics vanish as the instrumental transition matrix converges to zero.

Theorem 1. (Chen and Guo, 1985). If $r = 0$, or $r > 0$ but $C(z) - \frac{1}{2}I$ is SPR, then $\Phi(n, 0) \to 0$ implies $\theta_n \to \theta$.

Notably, under mild noise assumptions, this sufficient condition is also necessary when $r = 0$.

Assumption 1. ("Condition A").

- As $n \to \infty$, $\sum_{i=0}^{n} \frac{\varphi_i}{r_i} \varepsilon_{i+1}^\top$ tends to a finite limit S.
- There exist $c > 0$ and $\delta > 0$ which may depend on ω such that $\left\| S - \sum_{i=0}^{n-1} \frac{\varphi_i}{r_i} \varepsilon_{i+1}^\top \right\| \leq c r_n^{-\delta}$.

Remark 1. "Condition A" imposes a constraint on the cumulative noise effect, requiring that the weighted sum of noise terms converges sufficiently fast. This condition is weaker than typical independence or stationarity assumptions and allows for correlated noise sequences, making it applicable to a wide range of practical scenarios. The parameter δ quantifies the rate of this convergence.

Theorem 2. (Chen and Guo, 1985). Assume $r = 0$ and the noise sequence $\{\varepsilon_n\}$ satisfy "Condition A" at a sample $\omega \in \Omega$, then for this given sample path, for any initial value θ_0, we have $\theta_n \to \theta$ if and only if $\Phi(n, 0) \to 0$, and in this case, we have convergence speed

$$\|\theta_n - \theta\| = O(\|\Phi(n, 0)\|^{\delta/1+\delta}), \tag{17}$$

where $\delta > 0$ possibly depends on the sample ω.

This transforms the problem of establishing strong consistency for the SG algorithm into one of proving that $\Phi(n, 0) \to 0$. In essence, this is a problem of bounding the norm of a deterministic matrix product.

Assumption 2. w_n is a $\mathcal{F}_n$-measurable process satisfying

- $\mathbb{E}[w_n \mid \mathcal{F}_{n-1}] = 0$.
- $\mathbb{E}[\|w_n\|^2 \mid \mathcal{F}_{n-1}] \leq c_0 r_{n-1}^{\varepsilon}$, where $c_0 > 0$, $0 \leq \varepsilon \leq 1$.

Remark 2. The bound r_{n-1}^{ε} allows for potential growth in the noise variance, with $\varepsilon = 0$ corresponding to the bounded variance case and $\varepsilon > 0$ permitting variance that grows with the accumulated regressor norm r_n. Assumption 2 can often be used to verify that "Condition A" holds.

Assumption 3. $r_n \to \infty$, and $r_n = O(r_{n-1})$.

Remark 3. The condition $r_n \to \infty$ ensures we have enough information for identification. In real-world applications, most physical systems exhibit bounded input-output behavior due to physical constraints, actuator limits, and sensor ranges. The condition $r_n = O(r_{n-1})$ naturally arises in such scenarios, as it implies that the energy injected into the system cannot grow arbitrarily fast between consecutive time steps.

Using analytical techniques, Chen and Guo were able to significantly relax the excitation requirements for strong consistency of parameter estimates and provided quantitative convergence rates.

Theorem 3. (Guo, 1993). Provided that "Condition A", Assumption 3 and condition number growth

$$\kappa\left(\sum_{i=1}^{n} \varphi_i \varphi_i^\top\right) = O((\log r_n)^{1/3}) \tag{18}$$

are satisfied at a sample $\omega \in \Omega'$, then for this given sample path, we have $\Phi(n, 0) \to 0$. Moreover, we have convergence speed

$$\|\theta_n - \theta\| = O((\log r_n)^{-\delta}), \tag{19}$$

where $\delta > 0$ possibly depends on the sample ω.

Remark 4. This theorem establishes that the logarithmic growth of the condition number is sufficient for strong consistency of the SG algorithm. The condition number bound $O((\log r_n)^{1/3})$ represents a remarkably weak excitation condition, significantly relaxing earlier requirements of persistent excitation. The convergence rate $O((\log r_n)^{-\delta})$ shows that parameter estimates converge almost surely, though the rate can be slow depending on the sample path.

Through a counterexample inspired by Lai and Wei (1982), Chen and Guo show that if the condition number grows strictly faster than $O(\log r_n)$, the SG algorithm can be made to fail.

Theorem 4. (Chen and Guo, 1986). Assume Assumption 2 is satisfied almost surely, then for any $\delta > 0$, there exists a random vector sequence $\{\varphi_n\}$ satisfies Assumption 3 and the condition number growth

$$\kappa\left(\sum_{i=1}^{n} \varphi_i \varphi_i^\top\right) = O((\log r_n)^{1+\delta}), \text{ a.s.} \tag{20}$$

but $\Phi(n, 0) \not\to 0$, a.s.

Chen and Guo posit that the established threshold for the condition number's growth rate might not be the final word, suggesting that the critical exponent in the bound could be sharpened to its ultimate limit $O(\log r_n)$.

3. MATHEMATICAL FRAMEWORK

In this section, to ensure our theoretical framework is sufficiently general to address a broad class of problems, we let $0 \leq A_n \leq I$ denote general positive semi-definite symmetric matrices, which admit a decomposition of the form

$$A_n = \phi_n \phi_n^\top. \tag{21}$$

Our goal is to estimate products involving the instrumental matrix (Theorem 5). We introduce an auxiliary sequence:

$$x_{i+1} = (I - A_i)x_i, \quad i \geq k. \tag{22}$$

From this, we obtain

$$x_i - x_k = \sum_{j=k}^{i-1} A_j x_j, \tag{23}$$

and since $A_i^2 \leq A_i$, it follows that

$$\|x_i\|^2 \leq \|x_{i-1}\|^2 - \langle A_i x_{i-1}, x_{i-1} \rangle. \tag{24}$$

Summing over j yields

$$\sum_{j=k}^{i-1} \|\phi_j^\top x_j\|^2 \leq \|x_k\|^2 - \|x_i\|^2. \tag{25}$$

Now, introduce a nonnegative real sequence $\mu_n \geq 0$ called the weights and define a weighted sum S_{ik} of the matrices A_n over the interval $[k, i]$:

$$S_{ik} = \sum_{j=k}^{i-1} \mu_j A_j. \tag{26}$$

Remark 5. The design of S_{ik} represents one of the fundamental innovations in this framework. The weights μ_j serve multiple purposes: they can compensate for non-uniform regressor magnitudes, emphasize periods of high information content, or discount older measurements in time-varying systems.

We estimate the quadratic form:

$$\begin{aligned} x_k^\top S_{ik} x_k &= x_k^\top \left(\sum_{j=k}^{i-1} \mu_j A_j \right) x_k = \sum_{j=k}^{i-1} \mu_j x_k^\top A_j x_k \\ &= \sum_{j=k}^{i-1} \mu_j x_k^\top \phi_j \phi_j^\top x_k = \sum_{j=k}^{i-1} \mu_j \|\phi_j^\top x_k\|^2. \end{aligned} \tag{27}$$

To gain better insight into this quantity, define the vectors

$$v = \begin{bmatrix} \ldots & \phi_j^\top x_k & \ldots \end{bmatrix}^\top, \tag{28}$$

$$u = \begin{bmatrix} \ldots & \phi_j^\top x_j & \ldots \end{bmatrix}^\top. \tag{29}$$

Note that

$$u - v = \begin{bmatrix} \ldots & \phi_j^\top (x_j - x_k) & \ldots \end{bmatrix}^\top. \tag{30}$$

Multiplying both sides of (23) by $\phi_i^\top$ gives

$$(u - v)_j = \phi_j^\top x_j - \phi_j^\top x_k = \sum_{l=k}^{j-1} (\phi_j^\top \phi_l)(\phi_l^\top x_l). \tag{31}$$

Since $\phi_l^\top x_l = u_l$, we define a strictly lower-triangular matrix C by

$$C_{jl} = \phi_j^\top \phi_l, \quad k \leq l < j. \tag{32}$$

By construction, we have the matrix identity:

$$v = (I - C)u \tag{33}$$

Remark 6. The matrix C encodes the *inter-temporal correlation structure* of the regressor sequence, i.e. it quantifies how much information each new regressor ϕ_i shares with previous regressors $\phi_l, l < j$. The strictly lower-triangular structure reflects the causal nature of time: future regressors cannot influence past ones. The matrix identity reveals that the initial projection error v_j equals the concurrent projection error u_j minus a correction term that accounts for how much the state has evolved due to previous updates.

Set $\Lambda = \mathrm{diag}(\ldots, \sqrt{\mu_j}, \ldots)$, then

$$\begin{aligned} \lambda_{\min}(S_{ik})\|x_k\|^2 &\leq x_k^\top S_{ik} x_k = \|\Lambda v\|^2 \\ &= \|\Lambda(I - C)u\|^2 \leq \|\Lambda(I - C)\|^2 \|u\|^2 \\ &= \|\Lambda(I - C)\|^2 (\|x_k\|^2 - \|x_i\|^2). \end{aligned} \tag{34}$$

It is a standard fact that the operator norm of a matrix can be bounded by its Frobenius norm:

$$\frac{1}{\sqrt{\mathrm{rank}(A)}} \|A\|_F \leq \|A\| \leq \|A\|_F. \tag{35}$$

This yields a good estimate of the matrix norm:

$$\begin{aligned} \|\Lambda(I - C)\| &\leq \|\Lambda\| + \|\Lambda C\| \leq \|\Lambda\| + \|\Lambda C\|_F \\ &= \sqrt{\max_{k \leq j < i} \mu_j} + \sqrt{\sum_{j=k}^{i-1} \mu_j \sum_{l=k}^{j-1} (\phi_j^\top \phi_l)^2}. \end{aligned} \tag{36}$$

Remark 7. The separate estimation of Λ and ΛC is strategically important, because it allows us to distinguish between two fundamentally different sources of "complexity" in the system: the *amplitude* of weights (controlled by $\|\Lambda\|$) and the *temporal correlation structure* (captured by $\|\Lambda C\|$). In applications, this means we can independently control system behavior through weight selection (choice of μ_j) and through regressor design (affecting correlation structure C).

For convenience, we denote

$$B_{jk} = \sum_{l=k}^{j-1} (\phi_j^\top \phi_l)^2. \tag{37}$$

Thus, we obtain the following inequality relating $\|x_k\|$ and $\|x_i\|$:

$$\begin{aligned} &\lambda_{\min}(S_{ik})\|x_k\|^2 \\ &\leq \left[\sqrt{\max_{k \leq j < i} \mu_j} + \sqrt{\sum_{j=k}^{i-1} \mu_j B_{jk}} \right] (\|x_k\|^2 - \|x_i\|^2). \end{aligned} \tag{38}$$

Since x_k is arbitrary, we may bound the norm of the state transition matrix as follows.

Theorem 5. Under the preceding notations, the following inequality holds:

$$\|\Phi(N, k)\|^2 \leq 1 - \frac{\lambda_{\min}(S_{Nk})}{\left(\sqrt{\max_{k \leq j < N} \mu_j} + \sqrt{\sum_{j=k}^{N-1} \mu_j B_{jk}} \right)^2}. \tag{39}$$

A key result in mathematical analysis is that, let $0 \leq a_k < 1$, then

$$\prod_{i=1}^{\infty} (1 - a_i) = 0 \Leftrightarrow \sum_{i=1}^{\infty} a_i = \infty. \tag{40}$$

Remark 8. Inspired by this fact, Guo set out to express the convergence of high-dimensional matrices in terms of divergence of one-dimensional series, which became the focus of his early work.

In our new framework, this basic fact yields an important corollary.

Corollary 1. Under the preceding notations, we have $\Phi(n, 0) \to 0$ if

$$\sum_{k=1}^{\infty} \frac{\lambda_{\min}(S_{t_k t_{k-1}})}{\left(\sqrt{\max_{t_{k-1} \leq j < t_k} \mu_j} + \sqrt{\sum_{j=t_{k-1}}^{t_k - 1} \mu_j B_{j t_{k-1}}} \right)^2} = \infty. \tag{41}$$

where t_k is a strictly increasing natural number sequence which tends to infinity, i.e. $t_k \to \infty$. In particular, the SG algorithm is strongly consistent.

Remark 9. This result relies on a "time-scale rescaling" technique, which is not only natural but essential. The core idea is to partition the time axis into intervals $[t_{k-1}, t_k)$ and analyze the system over these aggregated blocks. This approach is necessary because if we considered each matrix individually (i.e., interval length 1), the minimum eigenvalue $\lambda_{\min}(S_{t_k t_{k-1}})$ would be zero due to rank deficiency, causing the criterion to fail.

The weight sequence μ_j can be chosen as a *streaming statistic* of the regressors ϕ_k's, meaning each μ_j is computed from ϕ_j and μ_{j-1} alone. This property is particularly useful for online computation.

4. CONVERGENCE ANALYSIS

In this section, we substitute ϕ_j with $\varphi_j/\sqrt{r_j}$, yielding the following estimate by Cauchy-Schwarz inequality:

$$\begin{aligned} B_{jk} &= \sum_{l=k}^{j-1} (\phi_j^\top \phi_l)^2 = \sum_{l=k}^{j-1} \frac{(\varphi_j^\top \varphi_l)^2}{r_j r_l} \\ &\leq \sum_{l=k}^{j-1} \frac{\|\varphi_j\|^2}{r_j} \frac{\|\varphi_l\|^2}{r_l} = \frac{\|\varphi_j\|^2}{r_j} \sum_{l=k}^{j-1} \frac{r_l - r_{l-1}}{r_l} \\ &\leq \frac{\|\varphi_j\|^2}{r_j} \int_{r_{k-1}}^{r_{j-1}} \frac{dx}{x} = \frac{r_j - r_{j-1}}{r_j} \log \frac{r_{j-1}}{r_{k-1}}. \end{aligned} \tag{42}$$

Substituting this into the denominator of (41), and choosing $\mu_j = r_j$, we have

$$\begin{aligned} &\sum_{j=t_{k-1}}^{t_k-1} \mu_j B_{j t_{k-1}} \leq \sum_{j=t_{k-1}}^{t_k-1} (r_j - r_{j-1}) \log \frac{r_{j-1}}{r_{t_{k-1}-1}} \\ &\leq \int_{r_{t_{k-1}-1}}^{r_{t_k-1}} \log x dx - (r_{t_k-1} - r_{t_{k-1}-1}) \log r_{t_{k-1}-1} \\ &= r_{t_k-1} \log r_{t_k-1} - r_{t_k-1} - r_{t_{k-1}-1} \log r_{t_{k-1}-1} \\ &\quad + r_{t_{k-1}-1} - (r_{t_k-1} - r_{t_{k-1}-1}) \log r_{t_{k-1}-1} \\ &= r_{t_k-1} \left(\log r_{t_k-1} - \log r_{t_{k-1}-1} \right) - (r_{t_k-1} - r_{t_{k-1}-1}). \end{aligned} \tag{43}$$

This yields a applicable sufficient condition for the strong consistency.

Corollary 2. Under the preceding notations, we have $\Phi(n, 0) \to 0$ if

$$\sum_{k=1}^{\infty} \frac{\lambda_{\min}(S_{t_k t_{k-1}})}{D_k} = \infty. \tag{44}$$

where t_k is defined similarly, $S_{t_k t_{k-1}} = \sum_{i=t_{k-1}}^{t_k-1} \varphi_i \varphi_i^\top$, and $D_k = r_{t_k-1} \left(\log r_{t_k-1} - \log r_{t_{k-1}-1} \right) - (r_{t_k-1} - r_{t_{k-1}-1})$.

This lemma serves a function analogous to the 'time inverse function $m(t)$" introduced by Chen and Guo, but admits a formulation that facilitates the subsequent analysis.

Lemma 1. Assume Assumption 3 is satisfied, then there exists strictly increasing natural number sequence $\{t_k\}$, such that

$$\frac{k}{l} < \frac{r_{t_k}}{r_{t_{k-1}}} < lk, \tag{45}$$

i.e., $r_{t_k} = \Theta(k r_{t_{k-1}})$.

Proof. Define $t_k = \min\{j : r_j \geq k!\}$. By definition and Assumption 3,

$$k! \leq r_{t_k} \leq l r_{t_k - 1} < lk! \tag{46}$$

Substituting k by $k-1$ gives

$$(k-1)! \leq r_{t_{k-1}} < l(k-1)! \tag{47}$$

Combining these yields the desired inequality.

5. MAIN THEOREM

Now we are fully prepared to prove the main theorem.

Theorem 6. Assume Assumption 3 is satisfied, and condition number growth

$$\kappa \left(\sum_{i=1}^{n} \varphi_i \varphi_i^\top \right) = O((\log r_n)^\alpha), \tag{48}$$

where $\alpha < 1$, then we have $\Phi(n, 0) \to 0$.

Proof. For convenience, set

$$S_n = \sum_{i=1}^{n} \varphi_i \varphi_i^\top. \tag{49}$$

The matrix in (44) , $S_{t_k t_{k-1}}$, can be expressed as the difference of two cumulative matrices:

$$S_{t_k t_{k-1}} = S_{t_k-1} - S_{t_{k-1}-1}. \tag{50}$$

We use Weyl's inequality for the eigenvalues of a sum of Hermitian matrices, which states that

$$\lambda_{\min}(S_{t_k t_{k-1}}) \geq \lambda_{\min}(S_{t_k-1}) - \lambda_{\max}(S_{t_{k-1}-1}). \tag{51}$$

Now, we find a lower bound for the first term $\lambda_{\min}(S_{t_k-1})$ and an upper bound for the second term $\lambda_{\max}(S_{t_{k-1}-1})$. A lower bound is derived as follows:

$$\begin{aligned} \lambda_{\min}(S_{t_k-1}) &\geq \frac{\lambda_{\max}(S_{t_k-1})}{M(\log r_{t_k-1})^\alpha} \\ &\geq \frac{1}{d} \frac{\mathrm{tr}(S_{t_k-1})}{M(\log r_{t_k-1})^\alpha} = \frac{1}{d} \frac{r_{t_k-1} - 1}{M(\log r_{t_k-1})^\alpha}. \end{aligned} \tag{52}$$

The upper bound comes from monotonicity of r_n:

$$\lambda_{\max}(S_{t_{k-1}-1}) \leq \mathrm{tr}(S_{t_{k-1}-1}) = r_{t_{k-1}-1} - 1. \tag{53}$$

Substituting the results back into the Weyl's inequality gives the final lower bound:

$$\lambda_{\min}(S_{t_k t_{k-1}}) \geq \frac{1}{d} \frac{r_{t_k-1} - 1}{M(\log r_{t_k-1})^\alpha} - (r_{t_{k-1}-1} - 1). \tag{54}$$

Choose t_k as in Lemma 1. By Stirling's formula, we have

$$\log r_{t_k-1} \leq \log(lk!) \lesssim k \log k. \tag{55}$$

This offers a lower bound of $\lambda_{\min}(S_{t_k t_{k-1}})$ in terms of k:

$$\begin{aligned} \lambda_{\min}(S_{t_k t_{k-1}}) &\gtrsim \frac{k!}{(k \log k)^\alpha} - (k-1)! \\ &= (k-1)! \left(\frac{k^{1-\alpha}}{(\log k)^\alpha} - 1 \right) \gtrsim \frac{k!}{(k \log k)^\alpha}. \end{aligned} \tag{56}$$

Besides, note that

$$\begin{aligned} D_k &\leq r_{t_k-1} \left(\log r_{t_k-1} - \log r_{t_{k-1}-1} \right) \\ &\lesssim r_{t_k} \log \frac{r_{t_k}}{r_{t_{k-1}}} < r_{t_k} \log lk \lesssim k! \log k. \end{aligned} \tag{57}$$

To estimate the summand in (44), we obtain

$$\frac{\lambda_{\min}(S_{t_k t_{k-1}})}{D_k} \gtrsim \frac{k!}{(k \log k)^\alpha} \frac{1}{k! \log k} = \frac{1}{k^\alpha (\log k)^{1+\alpha}}. \tag{58}$$

Since the series on the right-hand side diverges, it follows that $\Phi(n,0) \to 0$.

6. CONCLUSIONS

This study has resolved a long-standing conjecture by Chen and Guo, demonstrating that the stochastic gradient algorithm achieves strong consistency even under non-persistent excitation, provided the condition number of the Fisher information matrix grows as $O((\log r_n)^\alpha)$ for any $0 \le \alpha < 1$. Our work has significantly broadened the known sufficient condition from $\alpha \le 1/3$ and has established $\alpha = 1$ as the critical threshold. By introducing a novel algebraic framework, we have provided sharper matrix bounds and a more transparent proof. This advancement not only deepens the theoretical understanding of stochastic gradient optimization methods but also challenges researchers to finally settle the boundary case of $\alpha = 1$. Future work should extend this framework to other step-size rules and validate its implications in practical deep learning and adaptive control scenarios.

ACKNOWLEDGEMENTS

The authors are deeply grateful to Professor *Lei Guo (IFAC Fellow, IEEE Fellow, AMSS)*, the first author's supervisor, for his invaluable guidance and for posing the fundamental research question that motivated this work. His insightful comments and continuous support have greatly contributed to this study. The authors are also indebted to Yujing Liu, Cheng Zhao and Xin Zheng for their valuable feedback and stimulating discussions.

REFERENCES

Anderson, T.W. and Taylor, J.B. (1979). Strong consistency of least squares estimates in dynamic models. *The Annals of Statistics*, 7(3), 484–489.

Åström, K.J. and Bohlin, T. (1965). Numerical identification of linear dynamic systems from normal operating records. In *IFAC Symposium on Identification and Process Parameter Estimation*. Prague, Czech Republic.

Åström, K.J. and Wittenmark, B. (1972). On self tuning regulators. Technical Report 7209, Lund Institute of Technology, Division of Automatic Control.

Becker, A.H., Kumar, P.R., and Wei, C.Z. (1985). Adaptive control with the stochastic approximation algorithm: Geometry and convergence. *IEEE Transactions on Automatic Control*, 30(4), 330–338.

Chen, H. (1981). Strong consistency of recursive identification under correlated noise. *Journal of Systems Science and Mathematical Sciences*, 1(1), 34–52. In Chinese with English abstract.

Chen, H. (1982). Strong consistence and convergence rate of least squares identification. *Scientia Sinica (Series A)*, XXV(7), 772–784. Received June 10, 1981.

Chen, H. and Guo, L. (1985a). Strong consistency of parameter estimates for discrete-time stochastic systems. *Journal of Systems Science and Mathematical Sciences*, 5(2), 81–93.

Chen, H. and Guo, L. (1985b). Strong consistency of recursive identification by no use of persistent excitation condition. *Acta Mathematicae Applicatae Sinica*, 2(2), 133–145.

Chen, H. and Guo, L. (1986). The limit of stochastic gradient algorithm for identifying systems not persistently excited. *Kexue Tongbao*, 31(19), 1803–1808.

Chen, H. and Caines, P. (1984). The strong consistency of the stochastic gradient algorithm of adaptive control. *Preprint*. Presented at the Conference on Decision and Control.

Chowdhary, G., Wu, T., Cutler, M., Ure, N.K., and How, J.P. (2012). Experimental results of concurrent learning adaptive controllers. In *AIAA Guidance, Navigation, and Control Conference*. American Institute of Aeronautics and Astronautics.

Chung, K.L. (1954). On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3), 463–483.

Dvoretzky, A. (1956). On stochastic approximation. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 39–55.

Gladyshev, E.G. (1965). On stochastic approximation. *Theory of Probability and Its Applications*, 10(2), 275–278. Translated by A. R. Krainan.

Goodwin, G.C., Ramadge, P.J., and Caines, P.E. (1980). Discrete-time multivariable adaptive control. *IEEE Transactions on Automatic Control*, 25(3), 449–456.

Guo, L. (1993). *Time-Varying Stochastic Systems: Stability, Estimation and Control*. Jilin Science and Technology Press, Changchun. In Chinese.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3), 462–466.

Kingma, D.P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lai, T.L. and Wei, C.Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1), 154–166.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4), 551–575.

Ljung, L. (2010). Perspectives on system identification. *Annual Reviews in Control*, 34(1), 1–12.

Nar, K. and Sastry, S.S. (2019). Persistency of excitation for robustness of neural networks.

Rad, S.A., Tamizi, M.G., Azmoun, M., Tale Masouleh, M., and Kalhor, A. (2020). Experimental study on robust adaptive control with insufficient excitation of a 3-dof spherical parallel robot for stabilization purposes. *Mechanism and Machine Theory*, 153, 104026.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3), 400–407.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 28(2), 373–405. Received June 3, 1957; revised December 2, 1957.

Sakrison, D.J. (1964). Application of stochastic approximation methods to system optimization. Technical Report 391, Massachusetts Institute of Technology, Research Laboratory of Electronics, Cambridge, Massachusetts.