# ADVANCES IN SYSTEM IDENTIFICATION AND STOCHASTIC OPTIMIZATION

by
Long Wang

A dissertation submitted to The Johns Hopkins University
in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
March 2021

# Abstract

This work studies the framework of systems with subsystems, which has numerous practical applications, including system reliability estimation, sensor networks, and object detection. Consider a stochastic system composed of multiple subsystems, where the outputs are distributed according to many of the most common distributions, such as Gaussian, exponential and multinomial. In Chapter 1, we aim to identify the parameters of the system based on the structural knowledge of the system and the integration of data independently collected from multiple sources. Using the principles of maximum likelihood estimation, we provide the formal conditions for the convergence of the estimates to the true full system and subsystem parameters. The asymptotic normalities for the estimates and their connections to Fisher information matrices are also established, which are useful in providing the asymptotic or finite-sample confidence bounds.

The maximum likelihood approach is then connected to general stochastic optimization via the recursive least squares estimation in Chapter 2. For

ABSTRACT

stochastic optimization, we consider minimizing a loss function with only noisy function measurements and propose two general-purpose algorithms. In Chapter 3, the mixed simultaneous perturbation stochastic approximation (MSPSA) is introduced, which is designed for mixed variable (mixture of continuous and discrete variables) problems. The proposed MSPSA bridges the gap of dealing with mixed variables in the SPSA family, and unifies the framework of simultaneous perturbation as both the standard SPSA and discrete SPSA can now be deemed as two special cases of MSPSA. The almost sure convergence and rate of convergence of the MSPSA iterates are also derived. The convergence results reveal that the finite-sample bound of MSPSA is identical to discrete SPSA when the problem contains only discrete variables, and the asymptotic bound of MSPSA has the same order of magnitude as SPSA when the problem contains only continuous variables. In Chapter 4, the complex-step SPSA (CS-SPSA) is introduced, which utilizes the complex-valued perturbations to improve the efficiency of the standard SPSA. We prove that the CS-SPSA iterates converge almost surely to the optimum and achieve an accelerated convergence rate, which is faster than the standard convergence rate in derivative-free stochastic optimization algorithms.

**Primary Reader and Advisor:** James C. Spall

**Second Reader:** Fei Lu

*Dedication to my family.*

# Acknowledgments

I would like to thank my research advisor Prof. James Spall, for believing in me and for encouraging me throughout my study at Johns Hopkins University. It has been an unforgettable journey for me to be in this institution and I have been extremely fortunate to learn from him.

I would like to thank Prof. Daniel Naiman, Prof. Amitabh Basu, Prof. Fei Lu, Prof. Donniell Fishkind, and Prof. Avanti Athreya for their time in serving on my defense dissertation defense committee, with special thanks to Prof. Fei Lu for being the second reader of this lengthy dissertation.

I would like to thank Prof. James Spall, Prof. Daniel Naiman, Prof. Yanxun Xu, Prof. Mengyang Gu, Prof. Laurent Younes, Prof. Carey Priebe, Prof. Avanti Athreya, Prof. Donniell Fishkind, Prof. James Fill, and Prof. Daniel Robinson for their well-designed and enlightening classes in the Department of Applied Mathematics and Statistics. I also would like to give special thanks to Prof. Daniel Naiman for his kindness and support beyond classes.

Finally, I would like to thank my wife, Jingyi Zhu, who has always been

there for me. She not only helps me in academics by providing technical suggestions but also becomes an indispensable part of my life. She has made me feel loved and made me a better person.

# Contents

# List of Tables

# List of Figures

LIST OF FIGURES

# Chapter 1

# System Identification

## 1.1 Introduction

Consider a stochastic system composed of multiple subsystems, where both the full system and the subsystems have general binary or non-binary outputs. The framework of systems with subsystems is proposed and studied in a series of papers (see, e.g., Spall, 2008, 2009, 2010, 2012, 2013a,b, 2014; Maranzano and Spall, 2010a,b, 2011) with a focus on reliability estimation. In particular, previous literature considers a stochastic system composed of multiple subsystems, where each subsystem can generate binary ("0" or "1") outputs, and the full system can generate binary or non-binary outputs based on a special case of the exponential family distributions (e.g., Bernoulli or Gaussian).

One key motivation of combining full system and subsystem outputs is to

overcome the situation, where the full system outputs are difficult or infeasible to collect. Such a difficulty often arises when the full system operation requires the destruction of itself (e.g., missile launches or collision tests) or the full system is costly to operate (e.g., large-scale grid or power plant). The subsystem tests, on the other hand, are typically more feasible to obtain and much less expensive (e.g., local neighborhood tests or low-level component tests). While the full system outputs can reflect the general information of the entire system, the outputs from each subsystem can also reflect some partial information on the system. Although the full system outputs alone can provide some level of estimates for the parameter of interest at the full system level, being able to use all the data from both the full system and the subsystem can significantly improve the overall understanding and improve the accuracy of the required estimates. As shown in Spall (2014), better parameter estimations and tighter uncertainty bounds can be achieved by integrating data from different sources than by using the full system data alone. Figure 1.1 shows the conceptual illustration of the framework.

The framework of systems with subsystems appears in many fields, including system reliability estimation, sensor networks, object detection, and transportation networks. Let us now mention several applications under the framework of systems with subsystems. For system reliability assessment, Reese et al. (2011) considers a weapons-system surveillance program with multi-

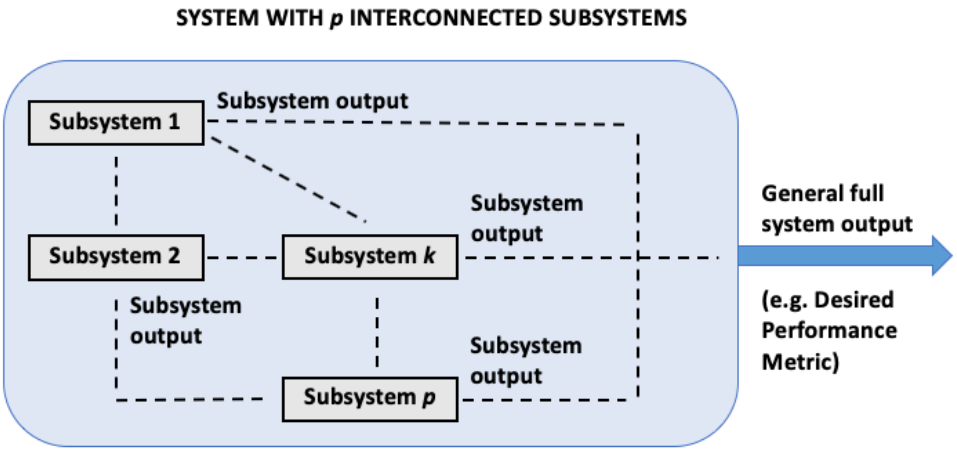



**Figure 1.1:** Conceptual illustration of the system with subsystems

ple components (i.e., subsystems) arranged in a series system, where the full system and subsystem measurements represent the lifetime data and are assumed to follow different Weibull distributions. The integration of multilevel heterogeneous data (analogous to the full system and subsystem measurements here) is investigated in Peng et al. (2013) for system reliability analysis, where different distribution assumptions are made depending on the data types, including binomial distribution for pass-fail data, Weibull distribution for lifetime data, and normal distribution for degradation data. Wilson et al. (2007) considers a case study in missile reliability that focuses on the assessment of a high fidelity launch vehicle intended to emulate a ballistic missile threat. In their work, the full system represents high-level mission events (i.e., launch, boost, booster separation, etc.) and the subsystems correspond to various parts that have to operate in concert to accomplish those high-level mis-

sions. The distribution assumptions include Dirichlet distribution and multinomial distribution. In a fault diagnosis problem, Zhou et al. (2012) constructs an Internet-based three-tank system to detect the potential leakage within the system. The end-to-end measurement of the water flow is analogous to the full system. Each subsystem output reflects whether the water level within each tank is beyond a pre-specified threshold. A simplified version of this problem is studied in Hernández and Spall (2015), where the full system output follows a normal distribution with known variance.

In transportation networks, Zhao and Spall (2016) aims to estimate the travel time in urban traffic by collecting test data from Google Maps. By identifying the transportation network from origin to destination through a specific route as a full system and the traffic link as the subsystem, Zhao and Spall (2016) assumes the full system outputs follow a log-normal distribution and the subsystem outputs follow a Bernoulli distribution.

In spatial search problems, the location of an object can be interpreted as a two-dimensional vector represented by angle and range. The full system corresponds to a direct measurement of the object. On the subsystems level, multiple sensors are installed in an area of interest. Each sensor can then actively monitor the activity with its local neighborhood and produce binary outputs to indicate if an object is present or a certain measurement is beyond a specified threshold. A problem of vehicle tracking with an autonomous interception over

a region is considered in Sharp et al. (2005). Multiple sensors spread out the entire region and each sensor can detect if the object is approaching or departing from itself. On the full system level, a pursuer could directly measure the object by installing a range sensor. The goal is to integrate the sensor data with the pursuer information for path estimation and interception of the intruder. In wireless sensor networks, Son et al. (2006) designs and implements a surveillance system for forest fires, where several sensors are installed in the forest to monitor temperature, humidity, and smoke. Those sensors are analogous to the subsystems in our framework. For the full system, the traditional infrared sensor system or satellite system can suggest potential forest fire locations on a much larger scale. Integrating the information from both sources can improve the detection of forest fires. A cooperative target tracking problem using multiple unmanned aerial vehicles (UAVs) as a mobile sensor network is studied in Sun et al. (2016), where a Doppler radar that can cover a very large area is the full system and the multiple UAVs that are flying in different orbits are the subsystems. A problem of assessing the privacy level in an information-sharing scheme is studied in Nekouei et al. (2018). In their work, the local processes, which can generate noisy information only for their own sensors, are corresponding to our subsystems, and the common process, which can be simultaneously observed by all the local sensors, is corresponding to our full system. Other attack-detection works focus on monitoring and diagnosis

of cyber-physical systems. One example includes attacks against process control in sensor networks, where the subsystems correspond to the local sensors, actuators, or control processing units (Cárdenas et al., 2011).

Other examples of "systems of systems" include airplane formation flight (Wolfe et al., 1996), flocks of systems (Brockett, 2010), and vehicle platooning (Oncu et al., 2012; Knorn and Middleton, 2013). For fault detection, Boem et al. (2017) proposes a methodology based on a distributed network, where a large-scale system is composed of many interconnected subsystems. It is shown that in the case of variables shared among more than one subsystem, the fault detectability can be improved. Such an idea is similar to the work in this paper, where the overall estimation is improved by integrating information among systems.

### 1.1.1 Our Contribution

This work aims to identify the parameters of the full system and the subsystems based on knowledge of how the system and subsystems relate to each other, as well as the integration of the data independently collected from multiple sources. The challenges of integrating data often arise from three different aspects: i) data from different sources may have different probability distributions; ii) the sample sizes for each data source could be different and there might be no samples at all for some subsystems; iii) the relationship be-

tween the full system and subsystem distribution parameters could be very complicated and it might be difficult to find an explicit function. We consider the approach of using maximum likelihood estimation (MLE), which has been studied in Spall (2014) to identify the unknown "success" probabilities of the subsystems and the mean parameter of the full system. Other approaches, such as the use of Bayesian methods to integrate the multilevel data, are discussed in Guo and Wilson (2013), Li et al. (2017) and Guo et al. (2018). Aside from issues related to how to specify weighting parameters, prior distributions, and hyperprior distributions, Bayesian methods usually require multivariate numerical integration, typically carried out with Markov chain Monte Carlo methods, which is time-consuming to complex systems. Numerical optimization, on the other hand, is often sufficient for finding the MLEs by solving the likelihood equations in most applications.

The main technical aspects that distinguish this work from standard results for the conventional MLEs come in two ways. First, each subsystem and the full system is allowed to have its own sample size. Some of the subsystems may even have no samples. The lack of an appropriate single sample size precludes the use of the most standard results on the convergence and asymptotic distributions in Spall (2005, Section 13.3) and Serfling (2009, Section 4.2). Second, different distributional assumptions are made on each subsystem and the full system. Because of the different distribution assumptions, it is possible

to achieve the convergence result and asymptotic normality only for the full system parameter, while nothing can be said for the subsystem parameters. Note that this seemingly inconsistent result holds even when all the subsystem distributions are identical. The standard theory of MLEs is applicable only in the special case where no subsystem outputs are generated and the full system outputs are independently and identically distributed (i.i.d.). In general, however, the subsystem sample sizes are not all zero and we expect to collect subsystem outputs to improve the overall estimates. Therefore, some subtle technical analysis is required to handle the different sample sizes and the different distributional assumptions. Although this work assumes the full system outputs are i.i.d., it can be further generalized to the independently and non-identically distributed (i.n.i.d.) case. Using the principles of MLE, we provide the formal conditions for the convergence of the estimates to the true full system and subsystem parameters. The asymptotic normalities for the estimates and the connections to Fisher information matrices (FIMs) are also established, which are useful in providing the asymptotic or finite-sample confidence bounds. The general framework studied in this work not only provides a more realistic model but also presents the most general case in the static settings.

The remainder of this Chapter is organized as follows: Section 1.2 discusses the generalization of the distribution assumption on the full system outputs

along with the convergence of the parameter estimate. Section 1.3 discusses the generalization of the distribution assumption on the subsystem outputs. It also covers the convergence and asymptotic distribution of the parameter estimate. A numerical study, which includes both simulation experiences and real applications, is provided in Section 1.4, and conclusions are made in Section 1.5.

## 1.2 General Full System with Binary Subsystems

In Spall (2014), the subsystem outputs are assumed to have a Bernoulli distribution and the full system outputs $y$ are distributed according to special case of the exponential family probability density (or mass) function, i.e., $p_k(y|\rho) = \exp[a_k(\rho)y + b_k(\rho) + c_k(y)]$, where $\rho$ represents the unknown mean output value of the full system, $a_k(\rho), b_k(\rho)$ and $c_k(y)$ are all real-valued functions with $a_k(\rho)$ and $c_k(\rho)$ being differentiable for all $k$. One natural extension is to consider the situation where there may be multiple parameters of interest in the full system. Further, the parameter of interest in the full system may be beyond only the mean of the output and the probability density function may include a sufficient statistic that is a function of $y$, not just $y$ itself. For example, in a one-parameter case, the full system output could follow a chi-square distribution, which is widely used in quality and reliability engineering. Note that in

this case, the sufficient statistic becomes $\log y$ and the probability density function discussed in Spall (2014) is no longer applicable. In the multi-parameter case, the full system output could follow a normal distribution with unknown mean and variance. Such a distribution is used to model the build-up of tolerances or life distribution of high-stress components. Another example includes the Weibull distribution with the shape and scale parameters, commonly used in applications such as failure time of components subjected to fatigue, scheduling inspection, or preventive maintenance activities. Our work also covers the gamma distribution with the shape and rate parameters, which is often used to model the time between maintenance actions or the failure time of the system with standby units.

### 1.2.1 Problem Formulation

Let us now consider the general formulations for the likelihood function and the score vector. Consider a system with $J \geq 1$ subsystems, which could be arranged in series, parallel, or any other form. It is assumed that the *test data* collected from each subsystem and the full system are statistically independent. Note that each subsystem is not necessarily functioning independently from each other when operating as a part of the full system.

Each subsystem is assumed to follow a Bernoulli distribution. Let the unknown parameter vector $\theta = [\rho_1, \ldots, \rho_J]^T$, where $\rho_j$ is the success probability

for the $j$-th subsystem for $j = 1, \ldots, J$. Denote $X_j$ as the number of successes and $n_j$ as the sample size; then the likelihood function of the $j$-th subsystem is

$$p(X_j; \rho_j) = \binom{n_j}{X_j} \rho_j^{X_j} (1 - \rho_j)^{n_j - X_j}.$$

For the full system, assume it has the following exponential family probability density function with a natural parameter vector $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_q]^T, q \geq 1$

$$p(y; \boldsymbol{\eta}) = \exp[\boldsymbol{\eta} \cdot \boldsymbol{T}(y) - A(\boldsymbol{\eta}) + G(y)], \tag{1.1}$$

where $A(\boldsymbol{\eta})$ and $G(y)$ are real-valued differentiable functions and the vector $\boldsymbol{T}(y) = [T_1(y), \ldots, T_q(y)]^T$. Let $\theta^* = [\rho_1^*, \ldots, \rho_J^*]^T$ and $\boldsymbol{\eta}^* = [\eta_1^*, \ldots, \eta_q^*]^T$ be the true but unknown parameters of the subsystems and full system, respectively. Note that (1.1) is the canonical form of the exponential family and $\boldsymbol{T}(y)$ is the natural sufficient statistic of the family. Let us mention some important properties of this canonical exponential family that will be very useful in the later proof of convergence. If the support of $\boldsymbol{\eta}$ is defined to be $\mathcal{E}$, then $\mathcal{E}$ is convex and the function $A(\boldsymbol{\eta})$ is also convex. Denote $\boldsymbol{\eta}^*$ as the true parameter. We have $\mathbb{E}[\boldsymbol{T}(y)] = \boldsymbol{A}'(\boldsymbol{\eta}^*)$, where $\boldsymbol{A}'(\boldsymbol{\eta}^*) = [\partial A(\boldsymbol{\eta}^*)/\partial \eta_1, \ldots, \partial A(\boldsymbol{\eta}^*)/\partial \eta_q]^T$ (Bickel and Doksum, 2001, Theorem 1.6.3 and Corollary 1.6.1). If the natural parameter space, $\mathcal{E}$, is open, then these following arguments are equivalent: (i) $\boldsymbol{\eta}$ is identifiable, (ii) $\boldsymbol{\eta} \rightarrow \boldsymbol{A}'(\boldsymbol{\eta})$ is a one-to-one function on $\mathcal{E}$ and

(iii) $A(\boldsymbol{\eta})$ is strictly convex on $\mathcal{E}$ (Bickel and Doksum, 2001, Theorem 1.6.4). This general exponential family distribution includes several special cases. For the normal distribution with unknown mean $\mu$ and variance $\sigma^2$, the natural parameter is $\boldsymbol{\eta} = [\mu/\sigma^2, -1/(2\sigma^2)]^T$, with sufficient statistics $\boldsymbol{T}(y) = [y, y^2]^T$, $A(\boldsymbol{\eta}) = -\eta_1^2/(4\eta_2) + 1/(2\log|1/(2\eta_2)|)$, and $G(y) = -1/(2\log(2\pi))$. For gamma distribution with shape parameter $\alpha$ and rate parameter $\beta$, the natural parameter is $\boldsymbol{\eta} = [\alpha - 1, -\beta]^T$, with sufficient statistics $\boldsymbol{T}(y) = [\log y, y]^T$, $A(\boldsymbol{\eta}) = \log\Gamma(\eta_1 + 1) - (\eta_1 + 1)\log(-\eta_2)$, and $G(y) = 0$.

Let $\{Y_k\}_{k=1}^n$ represent the i.i.d. observations (the *test data*) from the full system. Assume that all the subsystems and the full system data are independent. Then, the log-likelihood function $\mathcal{L}(\theta)$ can be written as

$$\begin{aligned}\mathcal{L}(\theta) &= \mathcal{L}(\{Y_k\}_{k=1}^n, \{X_j\}_{j=1}^p; \theta) \\ &= \sum_{k=1}^{n}[\boldsymbol{\eta}\cdot\boldsymbol{T}(Y_k) - A(\boldsymbol{\eta})] + \sum_{j=1}^{J}[X_j\log\rho_j + (n_j - X_j)\log(1-\rho_j)] \\ &\quad + \text{constant}, \end{aligned} \tag{1.2}$$

where the constant term does not depend on $\theta$. Note that the structure of the system, i.e., the relationship of the subsystems to the full system, is not explicitly reflected in the general expression of (1.2). As we show below, however, the full system parameter vector $\boldsymbol{\eta}$ is uniquely determined by the subsystem parameter vector $\theta$ and that is how the structure of the system enters into

the likelihood function. Hence, one goal is to estimate the parameter vector $\boldsymbol{\eta}$ based on the available data $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^T$ and $\boldsymbol{X} = [X_1, \ldots, X_J]^T$.

We now show how to formulate the maximum likelihood optimization problem. Let $\boldsymbol{\Theta}$ be the feasible region for the parameters of interest, $\boldsymbol{\theta}$, appearing in the subsystems. The MLE can then be found by solving the following optimization problem

$$\begin{aligned} &\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta}) \\ &\text{subject to } \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \boldsymbol{0}, \end{aligned}$$

where $\boldsymbol{f} : \mathbb{R}^{p+q} \to \mathbb{R}^q$ is a function defined by the structure of the system and reflecting the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Also define $f_i$ to be the $i$-th component of $\boldsymbol{f}$, such that $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\eta}) = [f_1(\boldsymbol{\theta}, \boldsymbol{\eta}), \ldots, f_q(\boldsymbol{\theta}, \boldsymbol{\eta})]^T$. Suppose that $\boldsymbol{f}(\cdot, \cdot)$ is a continuously differentiable function with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ in some open set of $\mathbb{R}^{p+q}$. Consider a fixed point $\boldsymbol{\theta}' \in (0, 1)^p$ and a corresponding point $\boldsymbol{\eta}'$ such that $\boldsymbol{f}(\boldsymbol{\theta}', \boldsymbol{\eta}') = \boldsymbol{0}$. If the matrix $\partial \boldsymbol{f}(\boldsymbol{\theta}', \boldsymbol{\eta}')/\partial \boldsymbol{\eta}^T$ with components $[\partial \boldsymbol{f}(\boldsymbol{\theta}', \boldsymbol{\eta}')/\partial \boldsymbol{\eta}^T]_{ii'} = \partial f_i(\boldsymbol{\theta}', \boldsymbol{\eta}')/\partial \eta_{i'}$ for $i = 1, \ldots, q$ and $i' = 1, \ldots, q$ is invertible, by the implicit function theorem (Apostol, 1974, Section 13.4), there exists an open neighborhood of $\boldsymbol{\theta}'$, an open neighborhood of $\boldsymbol{\eta}'$, and a unique continuously differentiable function $\boldsymbol{h} : \mathbb{R}^p \to \mathbb{R}^q$ such that for all $\boldsymbol{\theta}$ is this neighborhood, we have $\boldsymbol{\eta} = \boldsymbol{h}(\boldsymbol{\theta})$

with components $\eta_i = h_i(\theta)$ for $i = 1, \ldots, q$. The Jacobian matrix is

$$\boldsymbol{h}'(\theta) = -\left[\frac{\partial \boldsymbol{f}(\theta, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T}\right]^{-1} \frac{\partial \boldsymbol{f}(\theta, \boldsymbol{\eta})}{\partial \theta^T}, \tag{1.3}$$

where $\partial \boldsymbol{f}(\theta, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^T$ is a $q \times q$ invertible matrix and $\partial \boldsymbol{f}(\theta, \boldsymbol{\eta})/\partial \theta^T$ is a $q \times p$ matrix. Hence, the Jacobian matrix $\boldsymbol{h}'(\theta)$ is a $q \times p$ matrix with components $[\boldsymbol{h}'(\theta)]_{ij} = h'_{ij}(\theta) = \partial \eta_i / \partial \rho_j$ for $i = 1, \ldots, q$ and $j = 1, \ldots, J$.

Since $\hat{\theta}$ is the MLE of $\theta$, by the invariance property of the MLE (Bickel and Doksum, 2001, Problem 2.2.16), we have that $\hat{\boldsymbol{\eta}} = \boldsymbol{h}(\hat{\theta})$ will also be the MLE of $\boldsymbol{h}(\theta) = \boldsymbol{\eta}$. Here, we give an example to show that there may exist an explicit form of $\boldsymbol{h}$. This example is a simplified version of the work in Zhao and Spall (2016). Assume that in a series system, each output from subsystem $j$ is Bernoulli distributed with mean $\rho_j$ for $j = 1, \ldots, p$, and each output from the full system is normally distributed with mean $\mu = \sum_{j=1}^{J} \rho_j$ and $\sigma^2 = \sum_{j=1}^{J} \rho_j(1 - \rho_j)$. In terms of the natural parameters, we have $\eta_1 = h_1(\theta) = \sum_{j=1}^{J} \rho_j / \sum_{j=1}^{J} \rho_j(1 - \rho_j)$ and $\eta_2 = h_2(\theta) = -1/[2 \sum_{j=1}^{J} \rho_j(1 - \rho_j)]$.

In general, the MLE can be found by solving $\partial \mathcal{L}(\theta)/\partial \theta = 0$ assuming the existence of the above-mentioned function $\boldsymbol{h}(\theta) = \boldsymbol{\eta}$. The score vector $\partial \mathcal{L}(\theta)/\partial \theta$

has the following general form

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{k=1}^{n} \begin{bmatrix} \sum_{i=1}^{q} \frac{\partial \eta_i}{\partial \rho_1} \cdot T_i(Y_k) - \frac{\partial A(\eta)}{\partial \rho_1} \\ \vdots \\ \sum_{i=1}^{q} \frac{\partial \eta_i}{\partial \rho_J} \cdot T_i(Y_k) - \frac{\partial A(\eta)}{\partial \rho_J} \end{bmatrix} + \begin{bmatrix} \frac{X_1}{\rho_1} - \frac{n_1 - X_1}{1-\rho_1} \\ \vdots \\ \frac{X_J}{\rho_J} - \frac{n_J - X_J}{1-\rho_J} \end{bmatrix}.$$

Now the multi-variable chain rule implies that for $j = 1, \ldots, J$

$$\frac{\partial A(\eta)}{\partial \rho_j} = \sum_{i=1}^{q} \frac{\partial \eta_i}{\partial \rho_j} \frac{\partial A(\eta)}{\partial \eta_i} = \sum_{i=1}^{q} h'_{ij}(\theta) A'_i(\eta), \tag{1.4}$$

where $h'_{ij}(\theta) = \partial \eta_i / \partial \rho_j$ and $A'_i(\eta) = \partial A(\eta)/\partial \eta_i$. Substituting the above terms into (1.4), the score vector may be written as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta} &= \sum_{k=1}^{n} \begin{bmatrix} \sum_{i=1}^{q} h'_{i1}(\theta)[T_i(Y_k) - A'_i(\eta)] \\ \vdots \\ \sum_{i=1}^{q} h'_{ip}(\theta)[T_i(Y_k) - A'_i(\eta)] \end{bmatrix} + \begin{bmatrix} \frac{X_1}{\rho_1} - \frac{n_1 - X_1}{1-\rho_1} \\ \vdots \\ \frac{X_J}{\rho_J} - \frac{n_J - X_J}{1-\rho_J} \end{bmatrix} \\ &= \boldsymbol{h}'(\theta)^T \left[ \sum_{k=1}^{n} [\boldsymbol{T}(Y_k) - \boldsymbol{A}'(\eta)] \right] + \begin{bmatrix} \frac{X_1}{\rho_1} - \frac{n_1 - X_1}{1-\rho_1} \\ \vdots \\ \frac{X_J}{\rho_J} - \frac{n_J - X_J}{1-\rho_J} \end{bmatrix}, \end{aligned} \tag{1.5}$$

where the Jacobian matrix $\boldsymbol{h}'(\theta)$ may be computed according to (1.3) when $\boldsymbol{h}(\theta)$ is not explicitly available. Note that $\eta$ is a function of $\theta$. Hence solving the score equation, $\partial \mathcal{L}(\theta)/\partial \theta = 0$, reflects a careful balancing between the full

system and the subsystems. Because of the general form for $\boldsymbol{h}(\theta)$, the solution of $\partial\mathcal{L}(\theta)/\partial\theta = 0$ is usually found numerically. Even though it is possible that solving (1.5) may lead to a local maximum, finding MLEs by solving the score equation is a standard practice in system identification and related areas. When there exists enough subsystem data, using $[X_1/n_1, \ldots, X_J/n_J]^T$ as a starting point for a standard optimization algorithm may help in finding MLE of $\theta$.

### 1.2.2 Convergence

This subsection proves the convergence result of the full system vector parameter $\eta$. As argued in Section 1.1.1, due to the different sample sizes, the classic MLE convergence theory does not apply. Nonetheless, it is shown below that under some reasonable assumptions, the MLEs for $\eta$ will converge to the true parameter value, $\eta^*$.

Let us first show the derivation of the FIM, which is usually used in determining whether or not the subsystems and full system parameters are locally identifiable. Moreover, it is also used to construct confidence regions for MLE when the relevant sample size is sufficiently large. By the definition of the FIM for a differentiable log-likelihood function, we have

$$\boldsymbol{F}_N(\theta) = \mathbb{E}\left[\frac{\partial\mathcal{L}(\theta)}{\partial\theta} \cdot \frac{\partial\mathcal{L}(\theta)}{\partial\theta^T}\right]$$

where $\partial\mathcal{L}(\theta)/\partial\theta^T = [\partial\mathcal{L}/\partial\theta]^T$. Assume the existence of the unique continuously differentiable function $\boldsymbol{h}(\theta)$. Using natural parameters and the exponential family probability density function (1.1), the FIM is $\boldsymbol{I}(\eta) = \boldsymbol{A}''(\eta)$ with $[\boldsymbol{A}''(\eta)]_{i,i'} = \partial^2 A(\eta)/\partial\eta_i\eta_{i'}$. By the classic change-of-variable technique, the FIM can be re-written in terms of the subsystems parameter vector $\theta$, i.e., $\boldsymbol{I}(\theta) = \boldsymbol{h}'(\theta)^T\boldsymbol{I}(\eta)\boldsymbol{h}'(\theta)$ (Bickel and Doksum, 2001, Section 1.6.4). Let $N = n + \sum_{j=1}^{J} n_j$ be the total sample size. Since the full system data are independent of the subsystems data, the general form of the Fisher information matrix for the entire system has the additive form:

$$\boldsymbol{F}_N(\theta) = n\boldsymbol{h}'(\theta)^T\boldsymbol{I}(\eta)\boldsymbol{h}'(\theta) + \boldsymbol{J}_N(\theta), \tag{1.6}$$

where

$$\boldsymbol{J}_N(\theta) = \begin{bmatrix} \frac{n_1}{\rho_1(1-\rho_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{n_J}{\rho_J(1-\rho_J)} \end{bmatrix}.$$

When $n_j > 0$ for $j = 1, \ldots, J$, it is clear that $\boldsymbol{F}_N(\theta)$ is positive definite and when $n \geq 0$ the first term is positive semi-definite. More importantly, it matches the intuition that more data leads to less uncertainty about of estimates.

**Theorem 1.1** (Wang and Spall, 2017)**.** *Suppose $\boldsymbol{J}_N(\theta)$ has full rank and the implicit function theorem on $\boldsymbol{f}(\cdot,\cdot)$ holds. Assume that $n > 0$ and $n_j > 0$ for*

$j = 1, \ldots, J$. *Then* $\boldsymbol{F}_N(\theta)^{-1}$ *is given by*

$$\begin{aligned}\boldsymbol{F}_N(\theta)^{-1} = &- n\boldsymbol{J}_N(\theta)^{-1}\boldsymbol{h}'(\theta)^T[\boldsymbol{I}(\eta)^{-1} + n\boldsymbol{h}'(\theta)\boldsymbol{J}_N(\theta)^{-1}\boldsymbol{h}'(\theta)^T]^{-1}\boldsymbol{h}'(\theta)\boldsymbol{J}_N(\theta)^{-1} \\ &+ \boldsymbol{J}_N(\theta)^{-1}.\end{aligned}$$

**Remark 1.1.** *From (1.6), a sufficient condition for* $\boldsymbol{F}_N(\theta)$ *to have full rank is either* $\boldsymbol{h}'(\theta)^T\boldsymbol{I}(\eta)\boldsymbol{h}'(\theta)$ *or* $\boldsymbol{J}_N(\theta)$ *has full rank. Assume the number of parameters for the full system is no greater than the number of parameters for the subsystem, i.e.,* $q \leq J$*, which is common in practice and further assume that* $\partial \boldsymbol{f}(\theta, \theta)/\partial\theta^T$ *has full row rank. By (1.3), since* $\partial \boldsymbol{f}(\theta, \eta)/\partial\eta^T$ *is an invertible matrix, we see that* $\boldsymbol{h}'(\theta)$ *has full row rank, which implies* $\boldsymbol{h}'(\theta)^T\boldsymbol{I}(\eta)\boldsymbol{h}'(\theta)$ *has full rank as well. Now if* $n > 0$*,* $\boldsymbol{F}_N$ *is guaranteed to have full rank.*

*Proof.* Because $n > 0$ and $n_j > 0$ for $j = 1, \ldots, J$, we can decompose $\boldsymbol{J}_N(\theta)^{-1}$ into $\boldsymbol{J}_N(\theta)^{-1} = \boldsymbol{J}_N(\theta)^{-1/2}[\boldsymbol{J}_N(\theta)^{-1/2}]^T$, where

$$\boldsymbol{J}_N(\theta)^{-1/2} = \begin{bmatrix} \frac{\rho_1^{1/2}(1-\rho_1)^{1/2}}{n_1^{1/2}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\rho_J^{1/2}(1-\rho_J)^{1/2}}{n_J^{1/2}} \end{bmatrix}.$$

Hence, $\boldsymbol{h}'(\theta)\boldsymbol{J}_N(\theta)^{-1}\boldsymbol{h}'(\theta)^T$ can be written as a product of two matrices, i.e., $[\boldsymbol{h}'(\theta)\boldsymbol{J}_N(\theta)^{-1/2}][\boldsymbol{h}'(\theta)\boldsymbol{J}_N(\theta)^{-1/2}]^T$, which is clearly a positive semi-definite matrix. Also note that $\boldsymbol{I}(\eta)$ is positive definite, since it is the FIM of the canonical

exponential family (Bickel and Doksum, 2001, Example 3.4.5). Therefore, the term $\boldsymbol{I}(\boldsymbol{\eta})^{-1} + n\boldsymbol{h}'(\boldsymbol{\theta})\boldsymbol{J}_N(\boldsymbol{\theta})^{-1}\boldsymbol{h}'(\boldsymbol{\theta})^T$ is positive definite and invertible. After using the fact that $\boldsymbol{J}_N(\boldsymbol{\theta})$ is also invertible and applying the matrix inversion lemma (Spall, 2005, Appendix A), the result follows. □

Recall that $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\eta}}$ are the corresponding MLEs of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Before presenting Theorem 1.2 below on the convergence of $\hat{\boldsymbol{\eta}}$, let us define some notation that will be used in the later proof. Define the index $s \in \{1, \ldots, J\}$ such that $\limsup_{N\to\infty}(n_s/n_j) \leq 1$ for $j = 1, \ldots, J$, where note that it possible that $n \to \infty$ and $n_j \to \infty$ for $j = 1, \ldots, J$.

**Theorem 1.2** (Wang and Spall, 2017)**.** *Assume the parameter space of* $\boldsymbol{\theta}$ *is the open set* $\boldsymbol{\Theta} = (0, 1)^p$ *and the existence of the unique continuously differentiable function* $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{\eta}$ *with the bounded Jacobian matrix, i.e.,* $\|\boldsymbol{h}'(\boldsymbol{\theta})\| < \infty$*. Further assume the limits exist (including bounded and unbounded) for any ratios among the full system and subsystem sample sizes. Finally, assume* $\hat{\boldsymbol{\theta}}$ *exists, satisfies* $\partial\mathcal{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = 0$*, and is unique. Then,* $\hat{\boldsymbol{\eta}} \overset{a.s.}{\to} \boldsymbol{\eta}^*$ *in each of the following three cases: i)* $n \to \infty$ *and* $n_j = 0$ *for* $j = 1, \ldots, J$*; ii)* $n < \infty$ *and* $n_s \to \infty$*; iii)* $n \to \infty$ *and* $n_s \to \infty$*.*

*Proof.* All the limits below are as $N \to \infty$ and the convergence results are in a.s. sense.

**Case i)** When $n \to \infty$ and $n_j = 0$ for $j = 1, \ldots, J$, there is no subsystem data and the sample size of the full system goes to infinity. Under this case, $\hat{\boldsymbol{\eta}}$ is just

the MLE of the canonical exponential family defined in (1.1). Hence, $\hat{\boldsymbol{\eta}}$ exists, is unique, and $\hat{\boldsymbol{\eta}} \xrightarrow{\text{a.s.}} \boldsymbol{\eta}^*$ (Bickel and Doksum, 2001, Theorem 2.3.1 and 5.2.2).

**Case ii)** When $n < \infty$ and $n_s \to \infty$, the subsystem sample sizes go to infinity and the full system sample size is finite. Note that this contains a special case, where there is no full system data, i.e., $n = 0$. When $0 < n < \infty$ and at least one subsystem has non-zero data, the score vector in (1.5) will be a mixture of two parts, i.e., the full system and the subsystems, and the classic MLE theory cannot apply. Also, note that $n_s \to \infty$ implies $n_j > 0$ for all $j$. Using the fact that $\partial \mathcal{L}(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$ at $\hat{\boldsymbol{\theta}}$, we have for $j = 1, \ldots, J$

$$\sum_{k=1}^{n} \left[ \sum_{i=1}^{q} [T_i(Y_k) - A_i'(\hat{\boldsymbol{\eta}})] h_{ij}'(\hat{\boldsymbol{\theta}}) \right] + \frac{X_j}{\hat{\rho}_j} - \frac{n_j - X_j}{1 - \hat{\rho}_j} = 0. \tag{1.7}$$

Define $\bar{\boldsymbol{T}} = (1/n) \sum_{k=1}^{n} \boldsymbol{T}(Y_k)$ with components $\bar{T}_i$ for $i = 1, \ldots, q$. It is easy to see that only the term $T_i(Y_k)$ in (1.7) depends on $k$ and it can be written as

$$n \sum_{i=1}^{q} [\bar{T}_i(Y_k) - A_i'(\hat{\boldsymbol{\eta}})] h_{ij}'(\hat{\boldsymbol{\theta}}) + \frac{X_j}{\hat{\rho}_j} - \frac{n_j - X_j}{1 - \hat{\rho}_j} = 0.$$

Since $n > 0$ and $n_j > 0$ for all $j$, simplifying the above equation yields that for $j = 1, \ldots, J$,

$$\hat{\rho}_j - \bar{X}_j = \frac{n}{n_j} \hat{r}_j \sum_{i=1}^{q} [\bar{T}_i(Y_k) - A_i'(\hat{\boldsymbol{\eta}})] h_{ij}'(\hat{\boldsymbol{\theta}}),$$

where $\bar{X}_j = X_j/n_j$ and $\hat{r}_j = \hat{\rho}_j(1 - \hat{\rho}_j)$ for $j = 1, \ldots, J$. Applying the mean-value

expansion to $\hat{\eta}_i = h_i(\hat{\theta})$ around the vector $\bar{\boldsymbol{X}} = [\bar{X}_1, \ldots, \bar{X}_J]^T$ for $i = 1, \ldots, q$, we get

$$\hat{\eta}_i = h_i(\bar{\boldsymbol{X}}) + \boldsymbol{h}'_{i\cdot}(\tilde{\theta}^{(i)})(\hat{\theta} - \bar{\boldsymbol{X}}), \tag{1.8}$$

where $\boldsymbol{h}'_{i\cdot}(\tilde{\theta}^{(i)})$ is the $i$-th row of $\boldsymbol{h}'(\theta)$ evaluated at $\tilde{\theta}^{(i)}$ that lies on the interior of the line segment between $\hat{\theta}$ and the vector $\bar{\boldsymbol{X}}$ (Fleming, 2012, Page 86). The superscript $(i)$ on $\tilde{\theta}$ is necessary, since for each different $i$ the mean value expansion about $\hat{\eta}_i$ may require a different $\tilde{\theta}$. Substituting the term $(\hat{\theta} - \bar{\boldsymbol{X}})$ in (1.8) gives

$$\begin{aligned}
\hat{\eta}_i &= h_i(\bar{\boldsymbol{X}}) + \boldsymbol{h}'_{i\cdot}(\tilde{\theta}^{(i)}) \begin{bmatrix} \frac{n}{n_1}\hat{r}_1 \sum_{l=1}^{q}[\bar{T}_l - A'_l(\hat{\boldsymbol{\eta}})]h'_{l1}(\hat{\theta}) \\ \vdots \\ \frac{n}{n_J}\hat{r}_J \sum_{l=1}^{q}[\bar{T}_l - A'_l(\hat{\boldsymbol{\eta}})]h'_{lp}(\hat{\theta}) \end{bmatrix} \\
&= h_i(\bar{\boldsymbol{X}}) + \sum_{l=1}^{q}[\bar{T}_l - A'_l(\hat{\boldsymbol{\eta}})]\boldsymbol{h}'_{i\cdot}(\tilde{\theta}^{(i)})\boldsymbol{B}_N(\hat{\theta})\boldsymbol{h}'_{l\cdot}(\hat{\theta})^T,
\end{aligned}$$

where $\boldsymbol{B}_N(\hat{\theta}) = \text{diag}\,[n\hat{r}_1/n_1, \ldots, n\hat{r}_J/n_J]$. Combining all the expansions about $\hat{\eta}_i$ for $i = 1, \ldots, J$, it is convenient to express the above into the following matrix form

$$\boldsymbol{K}_N[\boldsymbol{A}'(\hat{\boldsymbol{\eta}}) - \bar{\boldsymbol{T}}] = \boldsymbol{h}(\bar{\boldsymbol{X}}) - \hat{\boldsymbol{\eta}}, \tag{1.9}$$

where $\boldsymbol{K}_N = \boldsymbol{h}'(\tilde{\theta})\boldsymbol{B}_N(\hat{\theta})\boldsymbol{h}'(\hat{\theta})^T$ is a $q \times q$ matrix. Note that for a finite sample, $\tilde{\theta}$ does not necessarily equal $\hat{\theta}$, which implies that $\boldsymbol{K}_N$ may not be a positive semi-definite matrix. Since $n < \infty$ and $n_s \to \infty$, we have $\boldsymbol{B}_N(\hat{\theta}) \to 0$ and

the boundedness of $\boldsymbol{h}'(\theta)$ implies $\boldsymbol{K}_N \to 0$. Therefore, the left-hand side of (1.9) converges to 0. Moreover, using the law of large numbers (LLN) on the subsystems data implies that $\bar{\boldsymbol{X}} \to \theta^*$ a.s. Again, by the continuity of $\boldsymbol{h}(\theta)$ at $\theta^*$, we have $\hat{\eta} \overset{\text{a.s.}}{\to} \eta^*$.

**Case iii)** When $n \to \infty$ and $n_s \to \infty$, the sample sizes of both the full system and subsystems go to infinity. Using the similar technique as in case ii), applying the mean-value expansion on $\boldsymbol{A}'(\hat{\eta}) = \boldsymbol{A}'(\boldsymbol{h}(\hat{\theta}))$ gives

$$[\boldsymbol{I} + \tilde{\boldsymbol{A}}'' \boldsymbol{K}_N][\boldsymbol{A}'(\hat{\eta}) - \bar{\boldsymbol{T}}] = [\boldsymbol{A}'(\boldsymbol{h}(\bar{\boldsymbol{X}})) - \boldsymbol{A}'(\eta^*)] - \tilde{\boldsymbol{A}}'' \boldsymbol{K}_N[\bar{\boldsymbol{T}} - \boldsymbol{A}'(\eta^*)], \quad (1.10)$$

where each row of $\tilde{\boldsymbol{A}}''$ is evaluated at a possibly different interior point of the line segment between $\boldsymbol{h}(\hat{\theta})$ and the vector $\boldsymbol{h}(\bar{\boldsymbol{X}})$. Note that the right-hand side of (1.10) goes to zero, since under this case, we must have $\bar{\boldsymbol{T}} \to \boldsymbol{A}'(\eta^*)$ and $\boldsymbol{h}(\bar{\boldsymbol{X}}) \to \boldsymbol{h}(\theta^*) = \eta^*$ by LLN. By the uniqueness of $\hat{\theta}$, we have $\boldsymbol{A}'(\hat{\eta}) - \boldsymbol{A}'(\eta^*) \to 0$. Since $\boldsymbol{A}'(\cdot)$ is a one-to-one function (Bickel and Doksum, 2001, Theorem 1.6.4), it is clear that $\hat{\eta} \overset{\text{a.s.}}{\to} \eta^*$. □

**Remark 1.2.** *To completely finish the discussion for $N \to \infty$ subject to $n + n_s \to \infty$, we also need to show the convergence under the case, $n \to \infty$ and $n_s < \infty$. As shown in case i) above, when there is an unbounded amount of full system data but no subsystems data, we have shown that $\hat{\eta} \to \eta^*$. By providing more subsystems data, it is intuitive to see that $\hat{\eta}$ still converges to $\eta^*$, since the full*

*system data will dominate the likelihood function. Eventually, when* $n_s \to \infty$*, where both the full system and subsystems have an unbounded amount of data, we can show that* $\hat{\boldsymbol{\eta}} \to \boldsymbol{\eta}^*$.

## 1.3 General Subsystems

In previous works of Spall (2014) and Wang and Spall (2017), the subsystem outputs are distributed according to a Bernoulli distribution. The binary assumption on the subsystems, however, remains a crucial assumption and limits the use of the framework in practice. Hence, this section considers the general exponential family distribution assumptions on both the full system and subsystem outputs, which makes the previous work in Spall (2014) and Wang and Spall (2017) special cases and allows the framework to be more broadly applicable for more practical applications, especially when subsystems are non-binary. It is also worth noting that the proof techniques in Spall (2014) and Wang and Spall (2017) are no longer directly applicable in this extension due to the non-binary assumptions on the subsystems. Hence this section provides a new method to present the formal convergence results and the asymptotic distributions of the MLEs under the most general case in static-parametric estimation. Although the discussion here assumes the system outputs i.i.d., one can consider the generalization to i.n.i.d. case by using probability density

function: $p_k(y|\boldsymbol{\eta}) = \exp[\boldsymbol{\eta} \cdot \boldsymbol{T}_k(y) - A_k(\boldsymbol{\eta}) + G_k(y)]$ with $A_k(\boldsymbol{\eta})$ and $G_k(y)$ being real-value differentiable functions and $\boldsymbol{T}_k(y)$ being the sufficient statistic.

### 1.3.1 Problem Formulation

Consider again a complex system that is composed of $J \geq 1$ subsystems, which can be arranged in series, parallel or any other form, and the test data are collected statistically independently from all data sources. Assume the full system follows an exponential family distribution with a vector parameter $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_q]^T$. Given the full system data $\boldsymbol{Y}$, the log-likelihood function has the standard canonical form

$$\mathcal{L}_F(\boldsymbol{Y}; \boldsymbol{\eta}) = \boldsymbol{\eta} \cdot \boldsymbol{T}_F(\boldsymbol{Y}) - A_F(\boldsymbol{\eta}) + G_F(\boldsymbol{Y}), \tag{1.11}$$

where $A_F(\cdot)$ and $G_F(\cdot)$ are real-value differentiable functions and the components of vector $\boldsymbol{T}_F(\boldsymbol{Y})$ are the natural sufficient statistics corresponding to the full subsystem. We use the script $F$ to explicitly indicate the notations are for the full systems. Although it is a little bit unconventional, we think this is the best approach to avoid overly cumbersome notations. Analogously, for $j = 1, \ldots, J$, assume the $j$-th subsystem has an exponential family distribution with a natural parameter vector $\boldsymbol{\theta}_j = [\theta_j^{(1)}, \ldots, \theta_j^{(p_j)}]^T$. Given the subsystem

data $\boldsymbol{X}^j$, the log-likelihood function has the standard canonical form

$$\mathcal{L}_j(\boldsymbol{X}^j;\theta_j) = \theta_j \cdot \boldsymbol{T}_j(\boldsymbol{X}^j) - A_j(\theta_j) + G_j(\boldsymbol{X}^j), \tag{1.12}$$

where $A_j(\cdot)$ and $G_j(\cdot)$ are real-value differentiable functions and the vector $\boldsymbol{T}_j(\boldsymbol{X}^j)$ is the natural sufficient statistic corresponding to the $j$-th subsystem.

Denote the vector of all the system parameters as $\theta = [\theta_1^T, \ldots, \theta_J^T]^T$ and let $\theta^* = [[\theta_1^*]^T, \ldots, [\theta_J^*]^T]^T$ and $\eta^* = [\eta_1^*, \ldots, \eta_q^*]^T$ be the true but unknown parameters of the subsystems and full system, respectively. To reflect the connection between the full system and subsystem parameters, we introduce a function $\boldsymbol{f} : \mathbb{R}^{p+q} \to \mathbb{R}^q$ with $p = \sum_{j=1}^J p_j$ such that $\boldsymbol{f}(\theta, \eta) = [f_1(\theta, \eta), \ldots, f_q(\theta, \eta)]^T = 0$. Note that the function $\boldsymbol{f}(\cdot, \cdot)$ is defined by the structure of the system. Assume all the full system data $\{\boldsymbol{Y}_k\}_{k=1}^{n_F}$ and the subsystem data $\{\boldsymbol{X}_k^1\}_{k=1}^{n_1}, \ldots, \{\boldsymbol{X}_k^J\}_{k=1}^{n_J}$ are all independent from each other, the overall log-likelihood function $\mathcal{L}(\theta)$ can be expressed by combining (1.11) and (1.12) as

$$\begin{aligned}\mathcal{L}(\theta) &= \mathcal{L}(\{\boldsymbol{Y}_k\}_{k=1}^{n_F}, \{\boldsymbol{X}_k^1\}_{k=1}^{n_1}, \ldots, \{\boldsymbol{X}_k^J\}_{k=1}^{n_J}; \theta) \\ &= n_F[\eta \cdot \bar{\boldsymbol{T}}_F - A_F(\eta)] + \sum_{j=1}^{J} n_j[\theta_j \cdot \bar{\boldsymbol{T}}_j - A_j(\theta_j)] + \text{constant}, \end{aligned} \tag{1.13}$$

where $\bar{\boldsymbol{T}}_F = (1/n_F)\sum_{k=1}^{n_F} \boldsymbol{T}_F(\boldsymbol{Y}_k), \bar{\boldsymbol{T}}_j = (1/n_j)\sum_{k=1}^{n_j} \boldsymbol{T}_j(\boldsymbol{X}_k^j)$ for $j = 1, \ldots, J$, and the constant term does not depend on $\theta$. Note that we express the overall log-likelihood function $\mathcal{L}(\theta)$ in (1.13) as a function of $\theta$ since the relationship of the

subsystems to the full system is implicitly reflected in $\boldsymbol{f}(\theta, \eta)$, and we require the full system parameter $\eta$ to be uniquely determined by the subsystem parameters $\theta$. Denote $\Theta$ as the feasible region of the subsystem parameters $\theta$. Similar to Section 1.2.1, we can formulate the maximum likelihood optimization problem as

$$\begin{aligned} &\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta) \\ &\text{subject to } \boldsymbol{f}(\theta, \eta) = 0. \end{aligned} \tag{1.14}$$

Again, using the implicit function theorem, we can have a unique system structural function $\boldsymbol{h} : \mathbb{R}^p \to \mathbb{R}^q$ such that $\boldsymbol{h}(\theta) = \eta$ with its Jacobian matrix $\boldsymbol{h}'(\theta)$ following (1.3), and we denote $\hat{\eta} = \boldsymbol{h}(\hat{\theta})$.

In general, assume the existence of the function $\boldsymbol{h}(\theta) = \eta$, then the MLE $\hat{\theta}$ is often founded by solving the score equation $\boldsymbol{\mathcal{L}}'(\theta) = \partial \mathcal{L}(\theta)/\partial\theta = 0$. In our setting, the score vector $\boldsymbol{\mathcal{L}}'(\theta)$ has the following general form,

$$\boldsymbol{\mathcal{L}}'(\theta) = \sum_{k=1}^{n_F} \left[ \boldsymbol{h}'(\theta)^T \boldsymbol{T}_F(\boldsymbol{Y}_k) - \frac{\partial A_F(\eta)}{\partial \theta} \right] + \begin{bmatrix} \sum_{k=1}^{n_1} [\boldsymbol{T}_1(\boldsymbol{X}_k^1) - \boldsymbol{A}_1'(\theta_1)] \\ \cdots \\ \sum_{k=1}^{n_J} [\boldsymbol{T}_J(\boldsymbol{X}_k^J) - \boldsymbol{A}_J'(\theta_J)] \end{bmatrix}, \tag{1.15}$$

where $\partial A_F(\eta)/\partial\theta = [[\partial A_F(\eta)/\partial\theta_1]^T, \ldots, [\partial A_F(\eta)/\partial\theta_J]^T]^T$ is a $p$-dimensional vector, and $\boldsymbol{A}_j'(\theta_j) = [\partial A_j(\theta_j)/\ \partial\theta_j^{(1)}, \ldots, \partial A_j(\theta_j)/\partial\theta_j^{(p_j)}]^T$ is a $p_j$-dimensional vector for $j = 1, \ldots, J$. Using the multivariate chain rule, the $l$-th component of

$\partial A_F(\boldsymbol{\eta})/\partial\boldsymbol{\theta}$ is

$$\left[\frac{\partial A_F(\boldsymbol{\eta})}{\partial\boldsymbol{\theta}}\right]_l = \sum_{i=1}^{q}\frac{\partial A_F(\boldsymbol{\eta})}{\partial\eta_i}\frac{\partial\eta_i}{\partial[\boldsymbol{\theta}]_l} = \sum_{i=1}^{q}[\boldsymbol{h}'(\boldsymbol{\theta})]_{i,l}[\boldsymbol{A}'_F(\boldsymbol{\eta})]_i, \tag{1.16}$$

where $\boldsymbol{A}'_F(\boldsymbol{\eta}) = [\partial A_F(\boldsymbol{\eta})/\partial\eta_1, \ldots, \partial A_F(\boldsymbol{\eta})/\partial\eta_q]^T$ is a $q$-dimensional vector, and we use $[\boldsymbol{A}'_F(\boldsymbol{\eta})]_i$ to denote its $i$-th component. Substituting (1.16) into (1.15), the $p$-dimensional score vector $\boldsymbol{\mathcal{L}}'(\boldsymbol{\theta})$ can be written as

$$\boldsymbol{\mathcal{L}}'(\boldsymbol{\theta}) = n_F\boldsymbol{h}'(\boldsymbol{\theta})^T[\bar{\boldsymbol{T}}_F - \boldsymbol{A}'_F(\boldsymbol{\eta})] + \begin{bmatrix} n_1\boldsymbol{I}_{p_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & n_J\boldsymbol{I}_{p_J} \end{bmatrix}[\bar{\boldsymbol{T}}_S - \boldsymbol{A}'_S(\boldsymbol{\theta})], \tag{1.17}$$

where $\bar{\boldsymbol{T}}_S = [\bar{\boldsymbol{T}}_1^T, \ldots, \bar{\boldsymbol{T}}_J^T]^T$ is a $p$-dimensional vector and $\boldsymbol{A}'_S(\boldsymbol{\theta}) = [\boldsymbol{A}'_1(\boldsymbol{\theta})^T, \ldots, \boldsymbol{A}'_J(\boldsymbol{\theta})^T]^T$ is another $p$-dimensional vector. Because $\boldsymbol{\eta}$ is a function of $\boldsymbol{\theta}$, solving the score equation $\boldsymbol{\mathcal{L}}'(\boldsymbol{\theta}) = \boldsymbol{0}$ requires a careful balance between the first and second term corresponding to the full system and the subsystems, respectively. Because of the general expression of $\boldsymbol{h}(\boldsymbol{\theta})$, the solution of $\boldsymbol{\mathcal{L}}'(\boldsymbol{\theta}) = \boldsymbol{0}$ is usually found numerically. Even though it is possible (perhaps likely) for the solution to be only a local maximum, finding MLEs by solving the score equation is a standard practice in system identification and related areas. When there exists enough subsystem data, using the solution of $\boldsymbol{A}'_j(\boldsymbol{\theta}_j) = \bar{\boldsymbol{T}}_j$ for $j = 1, \ldots, J$, as the starting point in standard optimization algorithm can help to find the

overall solution of $\boldsymbol{\mathcal{L}}'(\theta) = \mathbf{0}$.

### 1.3.2 Convergence

In this subsection, we provide the formal convergence proof in terms of the estimate of full system and subsystem parameters in Theorem 1.3 below. Let $N = n_F + \sum_{j=1}^{J} n_j$ be the total number of sample size. Because of the different samples sizes among the subsystems, we need to identify the slowest increasing subsystem sample size in order to prove the convergence as well as the asymptotic normality. Let the index $s \in \{1, \ldots, J\}$ be such that $\lim_{N\to\infty} n_s/n_j \leq 1$ for all $j$ such that $n_j > 0$. The index $s$ may not be unique, as arises, for example, when the subsystem sample sizes are the same (so $s$ may be any value in $\{1, \ldots, J\}$). Note that either $n_s < \infty$ or $n_s \to \infty$ may occur when $N \to \infty$, depending on the system under study.

**Assumption 1.1** (Subsystems)**.** *For $j = 1, \ldots, J$, denote $\boldsymbol{\Theta}_j$ as the parameter space of the natural parameter $\theta_j$. Assume the output of the $j$-th subsystem follows the canonical exponential family distribution described in (1.12), where the natural parameter space $\boldsymbol{\Theta}_j \subseteq \mathbb{R}^{p_j}$ is open and the family is of rank $p_j$. Further assume for any observed data vector $\boldsymbol{x}^j \in \mathbb{R}^{p_x^j}$ and any $p_j$-dimensional vector $\boldsymbol{c}_j \neq \mathbf{0}$, $\mathbb{P}(\boldsymbol{c}_j \cdot \boldsymbol{T}_j(\boldsymbol{X}^j) > \boldsymbol{c}_j \cdot \boldsymbol{T}_j(\boldsymbol{x}^j)) > 0$. Denote $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \times \cdots \times \boldsymbol{\Theta}_J \subseteq \mathbb{R}^p$ with $p = \sum_{j=1}^{J} p_j$ as the parameter space of all the subsystem parameters $\theta$.*

**Assumption 1.2** (Full System)**.** *Assume there exists a system structure function*

*$\boldsymbol{h}(\cdot)$ and denote $\mathcal{E} = \{\boldsymbol{\eta} = \boldsymbol{h}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ such that $\mathcal{E}$ is open and $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{\eta}$ is continuously differentiable. Assume the output of the full system follows the canonical exponential family distribution described in (1.11), where the natural parameter space $\mathcal{E} \subseteq \mathbb{R}^q$ is open and the family is of rank $q$. Further assume for any observed data vector $\boldsymbol{y}$ and the $q$-dimensional vector $\boldsymbol{c}_F \neq \boldsymbol{0}$, $\mathbb{P}(\boldsymbol{c}_F \cdot \boldsymbol{T}_F(\boldsymbol{Y}) > \boldsymbol{c}_F \cdot \boldsymbol{T}_F(\boldsymbol{y})) > 0$.*

**Remark 1.3.** *All the assumptions are necessary to guarantee the existence of MLE (Bickel and Doksum, 2001, Theorem 2.3.1), and are similar to those in Spall (2014). Recall that despite of the classical MLE properties for exponential family (Bickel and Doksum, 2001, Proposition 2.3.1 and Theorem 5.2.2), the log-likelihood function in (1.13) is not a special case of exponential family when both the full system and subsystems have non-zero data sizes. Hence, one focus here is to prove that the estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\eta}}$ are still consistent and asymptotically normally distributed, even though the classical theorems of convergence in MLE are not directly applicable.*

**Theorem 1.3.** *Let Assumptions 1.1 and 1.2 hold. Further assume that all the data are collected independently with the overall log-likelihood function. Finally, assume the limits exist (including bounded and unbounded) for any ratios among the full system and subsystem sample sizes. When $N \to \infty$ subject to $n_F + n_s \to \infty$, we have*

$$\hat{\boldsymbol{\eta}} \xrightarrow{\mathbb{P}} \boldsymbol{\eta}^*.$$

*Proof.* The condition $n_F + n_s \to \infty$ holds if and only if one of the following two cases holds: i) $\lim_{N\to\infty} n_F/n_s < \infty$ or ii) $\lim_{N\to\infty} n_F/n_s = \infty$. All the limits below are as $N \to \infty$. Before analyzing each case, denote $D_j(\theta_j) = \mathbb{E}[\mathcal{L}_j(\boldsymbol{X}^j, \theta_j)]$ for any $\theta_j \in \boldsymbol{\Theta}_j$ and $j = 1, \ldots, J$, and denote $D_F(\boldsymbol{\eta}) = \mathbb{E}[\mathcal{L}_F(\boldsymbol{Y}, \boldsymbol{\eta})]$ for any $\boldsymbol{\eta} \in \mathcal{E}$. It is easy to see that the law of large numbers (LLN) guarantees that, for any $\varepsilon > 0$,

$$\lim_{n_j\to\infty} \mathbb{P}\left(\sup_{\theta_j\in\boldsymbol{\Theta}_j} |\frac{1}{n_j}\sum_{k=1}^{n_j} \mathcal{L}_j(\boldsymbol{X}_k^j, \theta_j) - D_j(\theta_j)| \geq \varepsilon\right) = 0 \text{ for } j = 1, \ldots, J, \tag{1.18}$$

and

$$\lim_{n_F\to\infty} \mathbb{P}\left(\sup_{\boldsymbol{\eta}\in\mathcal{E}} |\frac{1}{n_F}\sum_{k=1}^{n_F} \mathcal{L}_F(\boldsymbol{Y}_k, \boldsymbol{\eta}) - D_F(\boldsymbol{\eta})| \geq \varepsilon\right) = 0 \tag{1.19}$$

Since $\theta_j^*$ for $j = 1, \ldots, J$ and $\boldsymbol{\eta}^*$ are the true parameter values, they are also corresponding to the unique maximizer of $D_j(\theta)$ for $j = 1, \ldots, J$ and $D_F(\boldsymbol{\eta})$, respectively (Bickel and Doksum, 2001, Theorem 2.3.1). Hence, we have for any $\varepsilon > 0$,

$$\sup_{\|\theta_j^* - \theta_j\| \geq \varepsilon} D_j(\theta_j) < D_j(\theta_j^*) \text{ for } j = 1, \ldots, J, \tag{1.20}$$

and

$$\sup_{\|\boldsymbol{\eta}^* - \boldsymbol{\eta}\| \geq \varepsilon} D_F(\boldsymbol{\eta}) < D_F(\boldsymbol{\eta}^*). \tag{1.21}$$

Now, let us discuss the convergence of $\hat{\theta}$ under each case.

**Case i)** $\lim_{N\to\infty} n_F/n_s < \infty$: Under this case, we must have $n_j > 0$ and

$n_j \to \infty$ for $j = 1, \ldots, J$. When $n_F = 0$, there is no full system data and (1.13) becomes a separable function in terms of the subsystem parameters $\theta_1, \ldots, \theta_J$. Hence, for $j = 1, \ldots, J$, the estimate $\hat{\theta}_j$ is simply the standard MLE of the log-likelihood function defined in (1.12). By Bickel and Doksum (2001, Theorem 2.3.1 and 5.2.2), $\hat{\theta}_j$ exists, is unique, and $\hat{\theta}_j \xrightarrow{\mathbb{P}} \theta_j^*$ for $j = 1, \ldots, J$ and consequently $\hat{\eta} \xrightarrow{\mathbb{P}} \eta^*$.

Now assume $n_F > 0$. Since $n_s$ represent the sample size of the slowest increasing subsystem, we proceed by considering the following two subcases: a) $1 \leq \lim_{N\to\infty} n_j/n_s < \infty$ for $j = 1, \ldots, J$ and b) there exists some $j$ such that $\lim_{N\to\infty} n_j/n_s = \infty$.

**Subcase i-a)** $\lim_{N\to\infty} n_F/n_s < \infty$ subject to $1 \leq \lim_{N\to\infty} n_j/n_s < \infty$ for $j = 1, \ldots, J$: Denote

$$D(\theta) = n_F D_F(\eta) + \sum_{j=1}^{J} n_j D_j(\theta_j). \tag{1.22}$$

From the definition of $\mathcal{L}(\theta)$ in (1.13), we have that

$$\begin{aligned}
\sup_{\theta\in\Theta} |\mathcal{L}(\theta) - D(\theta)| \leq \sup_{\theta\in\Theta} \Bigg\{ & n_F \left| \frac{1}{n_F} \sum_{k=1}^{n_F} \mathcal{L}_F(\boldsymbol{Y}_k, \eta) - D_F(\eta) \right| \\
& + \sum_{j=1}^{J} n_j \left| \frac{1}{n_j} \sum_{k=1}^{n_j} \mathcal{L}_j(\boldsymbol{X}_k^j, \theta_j) - D_j(\theta_j) \right| \Bigg\} \\
\leq n_F \sup_{\eta\in\mathcal{E}} & \left\{ \left| \frac{1}{n_F} \sum_{k=1}^{n_F} \mathcal{L}_F(\boldsymbol{Y}_k, \eta) - D_F(\eta) \right| \right\} \\
& + \sum_{j=1}^{J} n_j \sup_{\theta_j\in\Theta_j} \left\{ \left| \frac{1}{n_j} \sum_{k=1}^{n_j} \mathcal{L}_j(\boldsymbol{X}_k^j, \theta_j) - D_j(\theta_j) \right| \right\}.
\end{aligned} \tag{1.23}$$

Given that $\lim_{N\to\infty} n_F/n_s = 0$ and $\lim_{N\to\infty} n_j/n_s < \infty$ for $j = 1, \ldots, J$, it is easy to see that (1.23), together with (1.18) and (1.19), implies for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n_s}|\mathcal{L}(\boldsymbol{\theta}) - D(\boldsymbol{\theta})| \geq \varepsilon\right) \to 0. \tag{1.24}$$

Recall that $\hat{\boldsymbol{\theta}}$ is the maximizer of $\mathcal{L}(\boldsymbol{\theta})$ by the definition in (1.14). Therefore, given any $\varepsilon > 0$ and the fact that $\boldsymbol{\mathcal{L}}(\hat{\boldsymbol{\theta}}) \geq \boldsymbol{\mathcal{L}}(\boldsymbol{\theta}^*)$, we must have that the event $\hat{\boldsymbol{\theta}} \in \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \geq \varepsilon\}$ implies the event $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon}(1/n_s)[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] \geq 0$. Hence,

$$\mathbb{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \geq \varepsilon) \leq \mathbb{P}\left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \frac{1}{n_s}[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] \geq 0\right). \tag{1.25}$$

Moreover, for any $\delta > 0$, we have

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \frac{1}{n_s}[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] - \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \frac{1}{n_s}[D(\boldsymbol{\theta}) - D(\boldsymbol{\theta}^*)] > \delta\right) \to 0, \tag{1.26}$$

since the event within the probability operation on the left-hand side of (1.26) implies the event $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\ (1/n_s)[\mathcal{L}(\boldsymbol{\theta}) - D(\boldsymbol{\theta})] > \delta/2$, which has the probability $\mathbb{P}(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}(1/n_s)[\mathcal{L}(\boldsymbol{\theta}) - D(\boldsymbol{\theta})] > \delta/2) \to 0$ from (1.24). Using (1.20) and (1.21), we see that

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} D(\boldsymbol{\theta}) \leq n_F \sup_{\boldsymbol{\eta}\in\mathcal{E}} D_F(\boldsymbol{\eta}) + \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \sum_{j=1}^{J} n_j D_j(\boldsymbol{\theta}_j) < D(\boldsymbol{\theta}^*),$$

where $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \geq \varepsilon$ guarantees that there exists some $j$ such that $\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*\| \geq \xi$

for some $\xi > 0$ and $\sup_{\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*\| \geq \xi} D_j(\boldsymbol{\theta}_j) < D_j(\boldsymbol{\theta}_j^*)$. Hence, by choosing

$$2\delta = -\lim_{N\to\infty} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \frac{1}{n_s}[D(\boldsymbol{\theta}) - D(\boldsymbol{\theta}^*)] > 0,$$

we have (1.26) becomes

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \frac{1}{n_s}[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] > -\delta\right) \to 0,$$

which, combining with (1.25), leads to

$$\begin{aligned}\mathbb{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \geq \varepsilon) &\leq \mathbb{P}\left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \frac{1}{n_s}[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] \geq 0\right) \\ &\leq \mathbb{P}\left(\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \frac{1}{n_s}[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] > -\delta\right) \to 0. \qquad (1.27)\end{aligned}$$

Hence, we conclude $\hat{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}^*$ and consequently $\hat{\boldsymbol{\eta}} \xrightarrow{\mathbb{P}} \boldsymbol{\eta}^*$.

**Subcase i-b)** $\lim_{N\to\infty} n_F/n_s < \infty$ subject to there exists some $j$ such that $\lim_{N\to\infty} n_j/n_s = \infty$: Denote the index set $\mathcal{J}_1 = \{j : \lim_{N\to\infty} n_j/n_s < \infty\}$ and $\mathcal{J}_2 = \{j : \lim_{N\to\infty} n_j/n_s = \infty\}$. Rewrite the overall log-likelihood function as $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}^{(1)}(\boldsymbol{\theta}) + \sum_{j\in\mathcal{J}_2} \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j)$ with

$$\begin{aligned}\mathcal{L}^{(1)}(\boldsymbol{\theta}) &= \sum_{k=1}^{n_F} \mathcal{L}_F(\boldsymbol{Y}_k, \boldsymbol{\eta}) + \sum_{j\in\mathcal{J}_1}\sum_{k=1}^{n_j} \mathcal{L}_j(\boldsymbol{X}_k^j, \boldsymbol{\theta}_j) + \sum_{j\in\mathcal{J}_2}\sum_{k=1}^{n_s} \mathcal{L}_j(\boldsymbol{X}_k^j, \boldsymbol{\theta}_j), \\ \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j) &= \sum_{k=n_s+1}^{n_j} \mathcal{L}_j(\boldsymbol{X}_k^j, \boldsymbol{\theta}_j).\end{aligned}$$

Since $\mathcal{L}(\boldsymbol{\theta}^*) = \mathcal{L}^{(1)}(\boldsymbol{\theta}^*) + \sum_{j\in\mathcal{J}_2} \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j^*)$ and the event relationships

$$\begin{aligned}
&\left\{ \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \mathcal{L}^{(1)}(\boldsymbol{\theta}) + \sum_{j\in\mathcal{J}_2} \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j) \geq \mathcal{L}^{(1)}(\boldsymbol{\theta}^*) + \sum_{j\in\mathcal{J}_2} \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j^*) \right\} \\
&\subseteq \left\{ \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\xi} \mathcal{L}^{(1)}(\boldsymbol{\theta}) \geq \mathcal{L}^{(1)}(\boldsymbol{\theta}^*) \right\} \bigcup \left\{ \bigcup_{j\in\mathcal{J}_2} \left\{ \sup_{\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_j^*\|\geq\xi} \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j) \geq \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j^*) \right\} \right\}
\end{aligned} \tag{1.28}$$

for some $\xi > 0$ depending on $\varepsilon$ such that $\xi \to 0$ as $\varepsilon \to 0$, we have (1.28) implies

$$\begin{aligned}
&\mathbb{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \geq \varepsilon) \\
&\leq \mathbb{P}\left( \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) \geq 0 \right) \\
&= \mathbb{P}\left( \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\varepsilon} \mathcal{L}^{(1)}(\boldsymbol{\theta}) + \sum_{j\in\mathcal{J}_2} \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j) - \mathcal{L}^{(1)}(\boldsymbol{\theta}^*) - \sum_{j\in\mathcal{J}_2} \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j^*) \geq 0 \right) \\
&\leq \mathbb{P}\left( \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\xi} \mathcal{L}^{(1)}(\boldsymbol{\theta}) - \mathcal{L}^{(1)}(\boldsymbol{\theta}^*) \geq 0 \right) + \sum_{j\in\mathcal{J}_2} \mathbb{P}\left( \sup_{\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_j^*\|\geq\xi} \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j) - \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j^*) \geq 0 \right) \\
&= \mathbb{P}\left( \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\geq\xi} \frac{1}{n_s}[\mathcal{L}^{(1)}(\boldsymbol{\theta}) - \mathcal{L}^{(1)}(\boldsymbol{\theta}^*)] \geq 0 \right) \\
&\quad + \sum_{j\in\mathcal{J}_2} \mathbb{P}\left( \sup_{\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_j^*\|\geq\xi} \frac{1}{n_j - n_s}[\mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j) - \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j^*)] \geq 0 \right).
\end{aligned} \tag{1.29}$$

From (1.27), we have that the first probability on the right-hand side of (1.29) goes to 0 since $\lim_{N\to\infty} n_j/n_s < \infty$ for any $j \in \mathcal{J}_1$. Moreover, for any $j \in \mathcal{J}_2$, standard MLE theory implies for any $\xi > 0$,

$$\mathbb{P}\left( \left\| \arg\max_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j} \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j) - \boldsymbol{\theta}_j^* \right\| \geq \xi \right) \to 0$$

and using the fact that $\lim_{N\to\infty} n_j/(n_j - n_s) = 1$, we must have

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*\| \geq \xi} \frac{1}{n_j - n_s}[\mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j) - \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j^*)] \geq 0\right) \to 0$$

as well. Therefore, all the probabilities on the right-hand side of (1.29) go to $0$ implying $\mathbb{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \geq \varepsilon) \to 0$, i.e., $\hat{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}^*$ and consequently $\hat{\boldsymbol{\eta}} \xrightarrow{\mathbb{P}} \boldsymbol{\eta}^*$.

**Case ii)** $\lim_{N\to\infty} n_F/n_s = \infty$: Under this case, we must have $n_F > 0$ and $n_F \to \infty$. When $n_j = 0$ for $j = 1, \ldots, J$, there is no subsystem data and (1.13) becomes a function in terms of the full system parameter $\boldsymbol{\eta}$. Hence, the estimate $\hat{\boldsymbol{\eta}}$ is simply the standard MLE of the log-likelihodd function defined in (1.11). By Bickel and Doksum (2001, Theorem 2.3.1 and 5.2.2), $\hat{\boldsymbol{\eta}}$ exists, is unique, and $\hat{\boldsymbol{\eta}} \xrightarrow{\mathbb{P}} \boldsymbol{\eta}^*$. Now assume $n_j > 0$ for some $j$ and let us again proceed by considering the following two subcases: a) $\lim_{N\to\infty} n_F/n_j > 0$ for $j = 1, \ldots, J$ and b) there exists some $j$ such that $\lim_{N\to\infty} n_F/n_j = 0$.

**Subcase ii-a)** $\lim_{N\to\infty} n_F/n_s = \infty$ subject to $\lim_{N\to\infty} n_F/n_j > 0$ for $j = 1, \ldots, J$: Given $\lim_{N\to\infty} n_F/n_j > 0$ or equivalently $\lim_{N\to\infty} n_j/n_F < \infty$ for $j = 1, \ldots, J$, similar to how (1.24) is derived, for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n_F}|\mathcal{L}(\boldsymbol{\theta}) - D(\boldsymbol{\theta})| \geq \varepsilon\right) \to 0. \tag{1.30}$$

Note also that $\{\boldsymbol{\theta} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\} = \{\boldsymbol{\theta} : \boldsymbol{\eta} \in \mathcal{E}\}$ by Assumption 1.2 and we consider below the set $\{\boldsymbol{\theta} : \|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \geq \varepsilon\}$ since it is possible that $\lim_{N\to\infty} n_j/n_F = 0$ for

some or all $j$. Given $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta})$, for any $\varepsilon > 0$, the event $\hat{\boldsymbol{\theta}} \in \{\boldsymbol{\theta} : \|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \geq \varepsilon\}$ implies $\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon}(1/n_F)[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] \geq 0$ and hence

$$\mathbb{P}(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\| \geq \varepsilon) \leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon} \frac{1}{n_F}[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] \geq 0\right). \tag{1.31}$$

Note that the set $\{\boldsymbol{\theta} : \|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \geq \varepsilon\}$ is different from the set $\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \geq \varepsilon\}$ used in case i). Similar to the proof in case i), for any $\delta > 0$, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon} \frac{1}{n_F}[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] - \sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon} \frac{1}{n_F}[D(\boldsymbol{\theta}) - D(\boldsymbol{\theta}^*)] > \delta\right) \to 0, \tag{1.32}$$

since the event within the probability operation on the left-hand side of (1.32) implies $\sup_{\boldsymbol{\theta}:\boldsymbol{\eta}\in\mathcal{E}} (1/n_F)|\mathcal{L}(\boldsymbol{\theta}) - D(\boldsymbol{\theta})| > \delta/2$ with $\mathbb{P}(\sup_{\boldsymbol{\theta}:\boldsymbol{\eta}\in\mathcal{E}}(1/n_F)|\mathcal{L}(\boldsymbol{\theta}) - D(\boldsymbol{\theta})| > \delta/2) \to 0$ from (1.30). Using (1.20) and (1.21), we see that

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon} D(\boldsymbol{\theta}) \leq n_F \sup_{\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon} D_F(\boldsymbol{\eta}) + \sum_{j=1}^{J} n_j \sup_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j} D_j(\boldsymbol{\theta}_j) < D(\boldsymbol{\theta}^*).$$

Hence, by choosing $2\delta = -\lim_{N\to\infty} \sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon}(1/n_F)[D(\boldsymbol{\theta}) - D(\boldsymbol{\theta}^*)] > 0$, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon} \frac{1}{n_F}[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)] > -\delta\right) \to 0,$$

which, combining with (1.31), leads to

$$\begin{aligned}\mathbb{P}(\|\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}^*\|\geq\varepsilon) &\leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon}\frac{1}{n_F}[\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}^*)]\geq 0\right)\\ &\leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon}\frac{1}{n_F}[\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}^*)]>-\delta\right)\to 0, \qquad (1.33)\end{aligned}$$

i.e., $\hat{\boldsymbol{\eta}}\xrightarrow{\mathbb{P}}\boldsymbol{\eta}^*$.

**Subcase ii-b)** $\lim_{N\to\infty} n_F/n_s=\infty$ subject to there exists some $j$ such that $\lim_{N\to\infty} n_F/n_j=0$: Denote the index set $\mathcal{J}_3=\{j:\lim_{N\to\infty} n_F/n_j>0\}$ and $\mathcal{J}_4=\{j:\lim_{N\to\infty} n_F/n_j=0\}$. Similar to the proof in case i), rewrite the overall likelihood function as $\mathcal{L}(\boldsymbol{\theta})=\mathcal{L}^{(3)}(\boldsymbol{\theta})+\sum_{j\in\mathcal{J}_4}\mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j)$ with

$$\begin{aligned}\mathcal{L}^{(3)}(\boldsymbol{\theta}) &= \sum_{k=1}^{n_F}\mathcal{L}_F(\boldsymbol{Y}_k,\boldsymbol{\eta})+\sum_{j\in\mathcal{J}_3}\sum_{k=1}^{n_j}\mathcal{L}_j(\boldsymbol{X}_k^j,\boldsymbol{\theta}_j)+\sum_{j\in\mathcal{J}_4}\sum_{k=1}^{n_F}\mathcal{L}_j(\boldsymbol{X}_k^j,\boldsymbol{\theta}_j),\\ \mathcal{L}^{(4)}(\boldsymbol{\theta}_j) &= \sum_{k=n_F+1}^{n_j}\mathcal{L}_j(\boldsymbol{X}_k^j,\boldsymbol{\theta}_j).\end{aligned}$$

Since $\mathcal{L}(\boldsymbol{\theta}^*)=\mathcal{L}^{(3)}(\boldsymbol{\theta}^*)+\sum_{j\in\mathcal{J}_4}\mathcal{L}^{(4)}(\boldsymbol{\theta}_j^*)$ and the event relationships

$$\begin{aligned}&\left\{\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon}\mathcal{L}^{(3)}(\boldsymbol{\theta})+\sum_{j\in\mathcal{J}_4}\mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j)\geq\mathcal{L}^{(3)}(\boldsymbol{\theta}^*)+\sum_{j\in\mathcal{J}_4}\mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j^*)\right\}\\ &\subseteq\left\{\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon}\mathcal{L}^{(3)}(\boldsymbol{\theta})\geq\mathcal{L}^{(3)}(\boldsymbol{\theta}^*)\right\}\bigcup\left\{\bigcup_{j\in\mathcal{J}_4}\left\{\sup_{\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_j^*\|\geq\xi}\mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j)\geq\mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j^*)\right\}\right\}\end{aligned}$$

for some $\xi > 0$ depending on $\varepsilon$ such that $\xi \to 0$ as $\varepsilon \to 0$, we have

$$\begin{aligned}
&\mathbb{P}(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\| \geq \varepsilon) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon} \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) \geq 0\right) \\
&= \mathbb{P}\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon} \mathcal{L}^{(3)}(\boldsymbol{\theta}) + \sum_{j\in\mathcal{J}_4} \mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j) - \mathcal{L}^{(3)}(\boldsymbol{\theta}^*) - \sum_{j\in\mathcal{J}_4} \mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j^*) \geq 0\right) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\varepsilon} \mathcal{L}^{(3)}(\boldsymbol{\theta}) - \mathcal{L}^{(3)}(\boldsymbol{\theta}^*) \geq 0\right) \\
&\quad + \sum_{j\in\mathcal{J}_4} \mathbb{P}\left(\sup_{\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_j^*\|\geq\xi} \mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j) - \mathcal{L}_j^{(2)}(\boldsymbol{\theta}_j^*) \geq 0\right) \\
&= \mathbb{P}\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\|\geq\xi} \frac{1}{n_F}[\mathcal{L}^{(3)}(\boldsymbol{\theta}) - \mathcal{L}^{(3)}(\boldsymbol{\theta}^*)]\right) \\
&\quad + \sum_{j\in\mathcal{J}_4} \mathbb{P}\left(\sup_{\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_j^*\|\geq\xi} \frac{1}{n_j - n}[\mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j) - \mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j^*)] \geq 0\right), \qquad (1.34)
\end{aligned}$$

where we have the first probability on the right-hand side of (1.34) goes to 0 since $\lim_{N\to\infty} n_j/n_F < \infty$ for $j \in \mathcal{J}_3$. Moreover, for $j \in \mathcal{J}_4$, standard MLE theory implies for any $\xi > 0$,

$$\mathbb{P}\left(\left\|\arg\max_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j} \mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j) - \boldsymbol{\theta}_j^*\right\| \geq \xi\right) \to 0$$

and using the fact that $\lim_{N\to\infty} n_j/(n_j - n_F) = 1$, we must have

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\theta}_j-\boldsymbol{\theta}_j^*\|\geq\xi} \frac{1}{n_j - n_F}[\mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j) - \mathcal{L}_j^{(4)}(\boldsymbol{\theta}_j^*)] \geq 0\right) \to 0$$

as well. Therefore, all the probabilities on the right-hand side of (1.29) go to $0$ implying $\mathbb{P}(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\| \geq \varepsilon) \to 0$, i.e., $\hat{\boldsymbol{\eta}} \xrightarrow{\mathbb{P}} \boldsymbol{\eta}^*$. □

### 1.3.3 Asymptotic Distribution

This subsection provides the derivation of the asymptotic normality of $\hat{\boldsymbol{\eta}}$. Such an asymptotic distribution is useful in computing uncertainty bounds as well as the rate of convergence. Since the variance-covariance matrix in the asymptotic distribution is often closely related to the Fisher information matrix (FIM), we first derive the FIM of the overall log-likelihood function in (1.13) and then show the asymptotic distribution based on FIM.

Denote $\boldsymbol{\mathcal{L}}'_F(\boldsymbol{Y}_k; \boldsymbol{\eta}) = \partial \mathcal{L}_F(\boldsymbol{Y}_k; \boldsymbol{\eta})/\partial \boldsymbol{\eta}$ and $\boldsymbol{\mathcal{L}}'_j(\boldsymbol{X}_k^j; \boldsymbol{\theta}_j) = \partial \mathcal{L}_j(\boldsymbol{X}_k^j; \boldsymbol{\theta}_j)/\partial \boldsymbol{\theta}_j$ for $j = 1, \ldots, J$. Further denote $\boldsymbol{F}_F(\boldsymbol{\theta})$ and $\boldsymbol{F}_j(\boldsymbol{\theta}_j)$ for $j = 1, \ldots, J$ as the FIM for the full system and the $j$-th subsystem, respectively. The standard MLE theory (Bickel and Doksum, 2001, Theorem 5.3.5) on canonical form of exponential family distributions in (1.11) and (1.12) implies that

$$\boldsymbol{F}_F(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \mathcal{L}_F(\boldsymbol{Y}; \boldsymbol{\eta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_F(\boldsymbol{Y}; \boldsymbol{\eta})}{\partial \boldsymbol{\theta}^T}\right] = \boldsymbol{h}'(\boldsymbol{\theta})^T \boldsymbol{A}''_F(\boldsymbol{\eta}) \boldsymbol{h}'(\boldsymbol{\theta}), \tag{1.35}$$

$$\boldsymbol{F}_j(\boldsymbol{\theta}_j) = \mathbb{E}\left[\boldsymbol{\mathcal{L}}'_j(\boldsymbol{X}^j; \boldsymbol{\theta}_j) \boldsymbol{\mathcal{L}}'_j(\boldsymbol{X}^j; \boldsymbol{\theta}_j)^T\right] = \boldsymbol{A}''_j(\boldsymbol{\theta}_j) \text{ for } j = 1, \ldots, J, \tag{1.36}$$

where (1.35) is due to the change-of-variable technique with $\boldsymbol{A}''_F(\boldsymbol{\eta})$ being the FIM of the full system under the natural parameter $\boldsymbol{\eta}$ such that $[\boldsymbol{A}''_F(\boldsymbol{\eta})]_{a,b} =$

$\partial^2 \boldsymbol{A}_F(\boldsymbol{\eta})/\partial\eta_a\partial\eta_b$ for $a, b = 1, \ldots, q$, and (1.36) is with $[\boldsymbol{A}''_j(\boldsymbol{\theta}_j)]_{a,b} = \partial^2 A_j(\boldsymbol{\theta}_j)$ $/\partial\theta_j^{(a)}\partial\theta_j^{(b)}$ for $a, b = 1, \ldots, p_j$. Also note that $\boldsymbol{A}''_F(\boldsymbol{\eta}) = \mathbb{E}\left[\boldsymbol{\mathcal{L}}'_F(\boldsymbol{Y};\boldsymbol{\eta})\boldsymbol{\mathcal{L}}'_F(\boldsymbol{Y};\boldsymbol{\eta})^T\right]$. Furthermore, denote

$$\begin{aligned}
\boldsymbol{\mathcal{L}}''_F(\boldsymbol{Y}_k;\boldsymbol{\eta}) &= \frac{\partial^2 \mathcal{L}_F(\boldsymbol{Y}_k;\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}, \\
\boldsymbol{\mathcal{L}}''_j(\boldsymbol{X}_k^j;\boldsymbol{\theta}_j) &= \frac{\partial^2 \mathcal{L}_j(\boldsymbol{X}_k^j;\boldsymbol{\theta}_j)}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^T} \text{ for } j = 1, \ldots, J.
\end{aligned}$$

Then, we have the following alternative expressions of (1.35) and (1.36) as

$$\boldsymbol{F}_F(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}_F(\boldsymbol{Y};\boldsymbol{\eta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right], \tag{1.37}$$

$$\boldsymbol{F}_j(\boldsymbol{\theta}_j) = -\mathbb{E}\left[\boldsymbol{\mathcal{L}}''_j(\boldsymbol{X}^j;\boldsymbol{\theta}_j)\right] \text{ for } j = 1, \ldots, J. \tag{1.38}$$

Since the overall log-likelihood function in (1.13) is the sum of the log-likelihood function of full system data and the log-likelihood function of the subsystem data, it is easy to see that the $\boldsymbol{F}(\boldsymbol{\theta})$, the FIM of the overall system, has the form

$$\boldsymbol{F}(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\mathcal{L}}'(\boldsymbol{\theta})\boldsymbol{\mathcal{L}}'(\boldsymbol{\theta})^T] = n_F\boldsymbol{F}_F(\boldsymbol{\theta}) + \begin{bmatrix} n_1\boldsymbol{F}_1(\boldsymbol{\theta}_1) & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & n_J\boldsymbol{F}_J(\boldsymbol{\theta}_J) \end{bmatrix}, \tag{1.39}$$

where $\boldsymbol{F}(\boldsymbol{\theta})$ is positive definite whenever $n_j > 0$ for $j = 1, \ldots, J$, since both

$\boldsymbol{A}_F''(\boldsymbol{\eta})$ and $\boldsymbol{A}_j''(\boldsymbol{\theta}_j)$ for $j = 1, \ldots, J$ are positive definite matrices.

Now, we are ready to present the formal asymptotic normality of $\hat{\boldsymbol{\eta}}$ as below.

**Assumption 1.3** (Score Vector)**.** *Assume that the estimate* $\hat{\boldsymbol{\theta}}$ *defined in (1.14) satisfies* $\boldsymbol{\mathcal{L}}'(\hat{\boldsymbol{\theta}}) = \mathbf{0}$.

**Theorem 1.4.** *Let Assumptions 1.1–1.3 hold and let* $N \to \infty$ *subject to* $n_F + n_s \to \infty$. *Then i) if* $0 \leq \lim_{N\to\infty} n_F/n_s < \infty$

$$\sqrt{n_F + n_s}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{h}'(\boldsymbol{\theta}^*)^T \left[\lim_{N\to\infty} \frac{\boldsymbol{F}(\boldsymbol{\theta}^*)}{n_F + n_s}\right]^{-1} \boldsymbol{h}'(\boldsymbol{\theta}^*)\right);$$

*ii) if* $\lim_{N\to\infty} n_F/n_s = \infty$ *subject to* $n_j \to \infty$ *for* $j = 1, \ldots, J$ *and if*

$$\lambda_i\left(\frac{n_F}{n_s}\sum_{j=1}^{J} \boldsymbol{h}_j'(\boldsymbol{\theta}^*)[\boldsymbol{A}_j''(\boldsymbol{\theta}_j^*)]^{-1}\boldsymbol{h}_j'(\boldsymbol{\theta}^*)^T\right) = O\left(\frac{n_F}{n_s}\right),$$

*for* $i = 1, \ldots, q$, *where* $\lambda_i(\cdot)$ *denotes the* $i$*-th eigenvalue of the argument, then*

$$\sqrt{n_F}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{h}'(\boldsymbol{\theta}^*)^T \left[\lim_{N\to\infty} \frac{\boldsymbol{F}(\boldsymbol{\theta}^*)}{n_F + n_s}\right]^{-1} \boldsymbol{h}'(\boldsymbol{\theta}^*)\right);$$

*and iii) if* $\lim_{N\to\infty} n_F/n_s = \infty$ *subject to* $n_j < \infty$ *for* $j = 1, \ldots, J$, *then*

$$\sqrt{n_F}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{A}_F''(\boldsymbol{\eta}^*)^{-1}\right).$$

*Proof.* Similar to the proof of Theorem 1.3, all the limits below are as $N \to \infty$.

Similar to the proof of Theorem 1.3, we proceed by considering the following two cases: i) $\lim_{N\to\infty} n_F/n_s < \infty$ and ii) $\lim_{N\to\infty} n_F/n_s = \infty$ subject to $n_j \to \infty$ for $j = 1, \ldots, J$; and iii) $\lim_{N\to\infty} n_F/n_s = \infty$ subject to $n_j < \infty$ for $j = 1, \ldots, J$.

**Case i)** $\lim_{N\to\infty} n_F/n_s < \infty$: Under this case, we have $n_j \to \infty$ for all $j = 1, \ldots, J$. To discuss the behavior of $n_F$, we consider two subcase: a) $n_F \to \infty$ and b) $n_F < \infty$.

**Subcase i-a)** $\lim_{N\to\infty} n_F/n_s < \infty$ subject to $n_F \to \infty$: Recall that $\hat{\theta}$ is a maximizer of $\mathcal{L}(\theta)$ satisfying $\boldsymbol{\mathcal{L}}'(\hat{\theta}) = \mathbf{0}$. First denote $\boldsymbol{\mathcal{L}}''(\theta) = \partial^2\mathcal{L}(\theta)/\partial\theta\partial\theta^T$ as the second derivative of the overall log-likelihood function with the expression

$$\boldsymbol{\mathcal{L}}''(\theta) = \sum_{k=1}^{n_F} \frac{\partial^2 \mathcal{L}_F(\boldsymbol{Y}_k; \boldsymbol{\eta})}{\partial\theta\partial\theta^T} + \begin{bmatrix} \sum_{k=1}^{n_1} \boldsymbol{\mathcal{L}}_1''(\boldsymbol{X}_k^1; \theta_1) & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sum_{k=1}^{n_J} \boldsymbol{\mathcal{L}}_J''(\boldsymbol{X}_k^J; \theta_J) \end{bmatrix}.$$

Then, using the score vector in (1.17) and applying the mean value theorem on $\boldsymbol{\mathcal{L}}'(\hat{\theta})$ around $\theta^*$, we have

$$\mathbf{0} = \boldsymbol{\mathcal{L}}'(\hat{\theta}) = \boldsymbol{\mathcal{L}}'(\theta^*) + \tilde{\boldsymbol{\mathcal{L}}}'' \cdot (\hat{\theta} - \theta^*), \tag{1.40}$$

where $\tilde{\boldsymbol{\mathcal{L}}}''$ is a $p \times p$ matrix with its $i$-th row being the $i$-th row of $\boldsymbol{\mathcal{L}}''(\theta)$ evaluated at some (possibly different) $\tilde{\theta}^{(i)}$ that lies on the line segment between $\hat{\theta}$ and $\theta^*$. To get the asymptotic normality of $\hat{\theta}$, dividing both sides of (1.40) by $\sqrt{n_F + n_s}$,

we get

$$\left[-\frac{1}{n_F+n_s}\tilde{\boldsymbol{\mathcal{L}}}''\right]\sqrt{n_F+n_s}(\hat{\theta}-\theta^*)=\frac{1}{\sqrt{n_F+n_s}}\boldsymbol{\mathcal{L}}'(\theta^*). \tag{1.41}$$

For the term in the bracket on the left-hand side of (1.41), we can express $\tilde{\boldsymbol{\mathcal{L}}}''$ as

$$\tilde{\boldsymbol{\mathcal{L}}}''=\tilde{\boldsymbol{\mathcal{L}}}_F''+\begin{bmatrix}\tilde{\boldsymbol{\mathcal{L}}}_1'' & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & \tilde{\boldsymbol{\mathcal{L}}}_J''\end{bmatrix}$$

and using $\hat{\theta}\xrightarrow{\mathbb{P}}\theta^*$ from Theorem 1.3 gives

$$\frac{1}{n_F}\tilde{\boldsymbol{\mathcal{L}}}_F''\xrightarrow{\mathbb{P}}\lim_{n_F\to\infty}\frac{1}{n_F}\sum_{k=1}^{n_F}\frac{\partial^2\mathcal{L}_F(\boldsymbol{Y}_k;\boldsymbol{\eta}^*)}{\partial\theta\partial\theta^T}, \tag{1.42}$$

$$\frac{1}{n_j}\tilde{\boldsymbol{\mathcal{L}}}_j''\xrightarrow{\mathbb{P}}\lim_{n_j\to\infty}\frac{1}{n_j}\sum_{k=1}^{n_j}\boldsymbol{\mathcal{L}}_j''(\boldsymbol{X}_k^j;\theta_j^*)\text{ for }j=1,\dots,J. \tag{1.43}$$

Since $n_F\to\infty, n_j\to\infty$ for $j=1,\dots,J$, LLN further implies that

$$\frac{1}{n_F}\sum_{k=1}^{n_F}\frac{\partial^2\mathcal{L}_F(\boldsymbol{Y}_k;\boldsymbol{\eta}^*)}{\partial\theta\partial\theta^T}\xrightarrow{\mathbb{P}}\mathbb{E}\left[\frac{\partial^2\mathcal{L}_F(\boldsymbol{Y};\boldsymbol{\eta}^*)}{\partial\theta\partial\theta^T}\right], \tag{1.44}$$

$$\frac{1}{n_j}\sum_{k=1}^{n_j}\boldsymbol{\mathcal{L}}_j''(\boldsymbol{X}_k^j;\theta_j^*)\xrightarrow{\mathbb{P}}\mathbb{E}\left[\boldsymbol{\mathcal{L}}_j''(\boldsymbol{X}^j;\theta_j^*)\right]\text{ for }j=1,\dots,J. \tag{1.45}$$

Hence, combining (1.42)–(1.45) and using (1.37) and (1.38), we have

$$\frac{1}{n_F}\tilde{\boldsymbol{\mathcal{L}}}_F''\xrightarrow{\mathbb{P}}\boldsymbol{F}_F(\theta^*)\text{ and }\frac{1}{n_j}\tilde{\boldsymbol{\mathcal{L}}}_j''\xrightarrow{\mathbb{P}}\boldsymbol{F}_j(\theta_j^*)\text{ for }j=1,\dots,J.$$

Now, since $\lim_{N\to\infty} n_F/n_s < \infty$ implies $\lim_{N\to\infty} n_j/(n_F+n_s) > 0$ for $j = 1, \ldots, J$, we see that $\lim_{N\to\infty} \boldsymbol{F}(\boldsymbol{\theta}^*)/(n_F+n_s)$ is positive definite and hence non-singular. Therefore, multiplying the term in the bracket on the left-hand side of (1.41) by $[\lim_{N\to\infty} \boldsymbol{F}(\boldsymbol{\theta}^*)/(n_F+n_s)]^{-1}$, we have it converge to the identity matrix since

$$\begin{aligned}
&\left[\lim_{N\to\infty} \frac{\boldsymbol{F}(\boldsymbol{\theta}^*)}{n_F+n_s}\right]^{-1} \left[-\frac{1}{n_F+n_s}\tilde{\boldsymbol{\mathcal{L}}}''\right] \\
&\xrightarrow{\mathbb{P}} \left[\lim_{N\to\infty} \frac{\boldsymbol{F}(\boldsymbol{\theta}^*)}{n_F+n_s}\right]^{-1} \lim_{N\to\infty} \left[\frac{n_F}{n_F+n_s}\boldsymbol{F}_F(\boldsymbol{\theta}^*) + \begin{bmatrix} \frac{n_1}{n_F+n_s}\boldsymbol{F}_1(\boldsymbol{\theta}_1^*) & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_J}{n_F+n_s}\boldsymbol{F}_J(\boldsymbol{\theta}_J^*) \end{bmatrix}\right] \\
&= \left[\lim_{N\to\infty} \frac{\boldsymbol{F}(\boldsymbol{\theta}^*)}{n_F+n_s}\right]^{-1} \left[\lim_{N\to\infty} \frac{\boldsymbol{F}(\boldsymbol{\theta}^*)}{n_F+n_s}\right] = \boldsymbol{I}_p. \qquad (1.46)
\end{aligned}$$

For the term on the right-hand side of (1.41), the standard asymptotic normality theorem for exponential family distribution (Bickel and Doksum, 2001, Theorem 5.4.3) gives

$$\begin{aligned}
&\frac{1}{\sqrt{n_F}} \sum_{k=1}^{n_F} \frac{\partial \mathcal{L}_F(\boldsymbol{Y}_k; \boldsymbol{\eta}^*)}{\partial \boldsymbol{\eta}} = \sqrt{n_F}[\bar{\boldsymbol{T}}_F - \boldsymbol{A}_F'(\boldsymbol{\eta}^*)] \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}_F''(\boldsymbol{\eta}^*)), \\
&\frac{1}{\sqrt{n_j}} \sum_{k=1}^{n_j} \boldsymbol{\mathcal{L}}_j'(\boldsymbol{X}_k^j; \boldsymbol{\theta}_j^*) = \sqrt{n_j}[\bar{\boldsymbol{T}}_j - \boldsymbol{A}_j'(\boldsymbol{\theta}_j^*)] \xrightarrow{\mathcal{D}} N(\boldsymbol{0}, \boldsymbol{A}_j''(\boldsymbol{\theta}_j^*)) \text{ for } j = 1, \ldots, J,
\end{aligned}$$

where the change-of-variable technique (Bickel and Doksum, 2001, Theorem

5.4.2) further implies

$$\frac{1}{\sqrt{n_F}}\sum_{k=1}^{n_F}\frac{\partial \mathcal{L}_F(\boldsymbol{Y}_k;\theta^*)}{\partial \theta} = \sqrt{n_F}\boldsymbol{h}'(\theta^*)^T[\bar{\boldsymbol{T}}_F - \boldsymbol{A}'_F(\boldsymbol{\eta}^*)] \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{h}'(\theta^*)^T\boldsymbol{A}''_F(\boldsymbol{\eta}^*)\boldsymbol{h}'(\theta^*)).$$

Recall that $\boldsymbol{F}_F(\theta^*) = \boldsymbol{h}'(\theta^*)^T\boldsymbol{A}''_F(\boldsymbol{\eta}^*)\boldsymbol{h}'(\theta^*)$ and $\boldsymbol{F}_j(\theta_j^*) = \boldsymbol{A}''_j(\theta_j^*)$ from (1.35) and (1.36), respectively. Using the fact that all the data are independent, we see that

$$\left[\lim_{N\to\infty}\frac{\boldsymbol{F}(\theta^*)}{n_F+n_s}\right]^{-1}\frac{1}{\sqrt{n_F+n_s}}\boldsymbol{\mathcal{L}}'(\theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}) \tag{1.47}$$

with the variance-covariance matrix $\boldsymbol{\Sigma}$ being

$$\begin{aligned}\boldsymbol{\Sigma} = &\left[\lim_{N\to\infty}\frac{\boldsymbol{F}(\theta^*)}{n_F+n_s}\right]^{-1}\lim_{N\to\infty}\left[\frac{n_F}{n_F+n_s}\boldsymbol{F}_F(\theta^*) + \begin{bmatrix}\frac{n_1}{n_F+n_s}\boldsymbol{F}_1(\theta_1^*) & \cdots & \boldsymbol{0}\\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_J}{n_F+n_s}\boldsymbol{F}_J(\theta_J^*)\end{bmatrix}\right]\\
&\times\left[\lim_{N\to\infty}\frac{\boldsymbol{F}(\theta^*)}{n_F+n_s}\right]^{-1}\\
=&\left[\lim_{N\to\infty}\frac{\boldsymbol{F}(\theta^*)}{n_F+n_s}\right]^{-1}\left[\lim_{N\to\infty}\frac{\boldsymbol{F}(\theta^*)}{n_F+n_s}\right]\left[\lim_{N\to\infty}\frac{\boldsymbol{F}(\theta^*)}{n_F+n_s}\right]^{-1}\\
=&\left[\lim_{N\to\infty}\frac{\boldsymbol{F}(\theta^*)}{n_F+n_s}\right]^{-1}.\end{aligned}$$

Hence, multiplying both sides of (1.41) by $[\lim_{N\to\infty}\boldsymbol{F}(\theta^*)/(n_F+n_s)]^{-1}$, the results in (1.46) and (1.47) imply

$$\sqrt{n_F+n_s}(\hat{\theta}-\theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\boldsymbol{0}, \left[\lim_{N\to\infty}\frac{\boldsymbol{F}(\theta^*)}{n_F+n_s}\right]^{-1}\right).$$

Finally, since $\boldsymbol{h}'(\theta)$ exists and is non-zero, applying the delta method on $\boldsymbol{\eta} = \boldsymbol{h}(\theta)$ gives

$$\sqrt{n_F + n_s}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{h}'(\theta^*)^T \left[\lim_{N\to\infty} \frac{\boldsymbol{F}(\theta^*)}{n_F + n_s}\right]^{-1} \boldsymbol{h}'(\theta^*)\right).$$

**Subcase i-b)** $\lim_{N\to\infty} n_F/n_s < \infty$ subject to $n_F < \infty$: Note that this subcase is equivalent to $\lim_{N\to\infty} n_F/n_s = 0$ subject to $n_F < \infty$. Using the condition that $\partial\mathcal{L}(\hat{\theta})/\partial\theta = \mathbf{0}$, the score vector from (1.17) implies that

$$n_F \boldsymbol{h}'(\hat{\theta})^T[\bar{\boldsymbol{T}}_F - \boldsymbol{A}'_F(\hat{\boldsymbol{\eta}})] = \begin{bmatrix} n_1\boldsymbol{I}_{p_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & n_J\boldsymbol{I}_{p_J} \end{bmatrix} [\boldsymbol{A}'_S(\hat{\theta}) - \bar{\boldsymbol{T}}_S]. \tag{1.48}$$

Since $n_j > 0$ for $j = 1, \ldots, J$, multiplying both sides of (1.48) by $\text{diag}(\boldsymbol{I}_{p_1}/\sqrt{n_1}, \ldots, \boldsymbol{I}_{p_J}/\sqrt{n_J})$ gives

$$\begin{bmatrix} \frac{n_F}{\sqrt{n_1}}\boldsymbol{I}_{p_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{n_F}{\sqrt{n_J}}\boldsymbol{I}_{p_J} \end{bmatrix} \boldsymbol{h}'(\hat{\theta})^T[\bar{\boldsymbol{T}}_F - \boldsymbol{A}'_F(\hat{\boldsymbol{\eta}})] = \begin{bmatrix} \sqrt{n_1}\boldsymbol{I}_{p_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sqrt{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} [\boldsymbol{A}'_S(\hat{\theta}) - \bar{\boldsymbol{T}}_S]. \tag{1.49}$$

Adding and subtracting the right-hand side of (1.49) by $[\sqrt{n_1}\boldsymbol{A}'_1(\theta_1^*)^T, \ldots, \sqrt{n_J}$

$\boldsymbol{A}'_J(\theta^*_J)^T]^T$ and rearranging terms, we get

$$\begin{aligned}
&\begin{bmatrix} \sqrt{n_1}\boldsymbol{I}_{p_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sqrt{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} [\boldsymbol{A}'_S(\hat{\theta}) - \boldsymbol{A}'_S(\theta^*)] \\
&= \begin{bmatrix} \sqrt{n_1}\boldsymbol{I}_{p_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sqrt{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} [\bar{\boldsymbol{T}}_S - \boldsymbol{A}'_S(\theta^*)] + \begin{bmatrix} \frac{n_F}{\sqrt{n_1}}\boldsymbol{I}_{p_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{n_F}{\sqrt{n_J}}\boldsymbol{I}_{p_J} \end{bmatrix} \\
&\quad \times \boldsymbol{h}'(\hat{\theta})^T[\bar{\boldsymbol{T}}_F - \boldsymbol{A}'_F(\hat{\eta})]. \qquad (1.50)
\end{aligned}$$

Given that $n_j \to \infty$ for $j = 1, ..., J$, the standard CLT for exponential family distribution gives

$$\sqrt{n_j}[\bar{\boldsymbol{T}}_j - \boldsymbol{A}'_j(\theta^*_j)] \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{A}''(\theta^*_j)\right) \text{ for } j = 1, \ldots, J. \qquad (1.51)$$

Furthermore. since $n_F < \infty$ and $n_j \to \infty$ implies $\lim_{N\to\infty} n_F/\sqrt{n_j} = 0$ for $j = 1, \ldots, J$ and $\bar{\boldsymbol{T}}_1, \ldots, \bar{\boldsymbol{T}}_j$ are mutually independent, plugging (1.51) into (1.50) gives

$$\begin{bmatrix} \sqrt{n_1}[\boldsymbol{A}'_1(\hat{\theta}_1) - \boldsymbol{A}'_1(\theta^*_1)] \\ \vdots \\ \sqrt{n_J}[\boldsymbol{A}'_J(\hat{\theta}_J) - \boldsymbol{A}'_J(\theta^*_J)] \end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{A}''_1(\theta^*_1) & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{A}''_J(\theta^*_J) \end{bmatrix}\right).$$

Let $\boldsymbol{t}_j = \boldsymbol{A}_j'(\theta_j)$ for $j = 1, \ldots, J$. Since $\boldsymbol{A}_j'(\cdot)$ is a one-to-one function, there exists an inverse function $[\boldsymbol{A}_j']^{-1}(\cdot)$ such that $[\boldsymbol{A}_j']^{-1}(\boldsymbol{t}_j) = \theta_j$ with a Jacobian $\partial[\boldsymbol{A}_j']^{-1}(\boldsymbol{t}_j)/\partial \boldsymbol{t}_j = [\boldsymbol{A}_j''(\theta_j)]^{-1}$ (Bickel and Doksum, 2001, Theorem 5.3.5). Hence, the delta method implies

$$\begin{bmatrix} \sqrt{n_1}(\hat{\theta}_1 - \theta_1^*) & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \sqrt{n_J}(\hat{\theta}_J - \theta_J^*) \end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left( \boldsymbol{0}, \begin{bmatrix} [\boldsymbol{A}_1''(\theta_1^*)]^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & [\boldsymbol{A}_J''(\theta_J^*)]^{-1} \end{bmatrix} \right)$$

given $\theta_j = [\boldsymbol{A}_j']^{-1}(\boldsymbol{A}_j'(\theta_j))$ for $j = 1, \ldots, J$. Further, using the delta method one more time on $\eta = \boldsymbol{h}(\theta)$, we have

$$\begin{aligned} &\sqrt{n_F + n_s}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \\ &\xrightarrow{\mathcal{D}} \mathcal{N}\left( \boldsymbol{0}, \boldsymbol{h}'(\theta^*)^T \begin{bmatrix} \lim_{N\to\infty} \frac{n_1}{n_F+n_s} \boldsymbol{F}_1(\theta_1^*) & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \lim_{N\to\infty} \frac{n_J}{n_F+n_s} \boldsymbol{F}_J(\theta_J^*) \end{bmatrix}^{-1} \boldsymbol{h}'(\theta^*) \right). \end{aligned} \tag{1.52}$$

Recall that we also have $\lim_{N\to\infty} n_F/(n_F + n_s) = 0$ in this case. Hence it is easy to see that (1.52) is equivalent to

$$\sqrt{n_s}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left( \boldsymbol{0}, \boldsymbol{h}'(\theta^*)^T \left[ \lim_{N\to\infty} \frac{\boldsymbol{F}(\theta^*)}{n_F + n_s} \right]^{-1} \boldsymbol{h}'(\theta^*) \right).$$

**Case ii)** $\lim_{N\to\infty} n_F/n_s = \infty$ subject to $n_j \to \infty$ for $j = 1, \ldots, J$: Under this case, we also have $n_F \to \infty$. First note that the standard MLE theory on exponential family distribution implies that there exists a unique $\bar{\boldsymbol{\eta}} \in \mathcal{E}$ such that $\bar{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}\in\mathcal{E}} \sum_{k=1}^{n_F} \mathcal{L}_F(\boldsymbol{Y}_k; \boldsymbol{\eta})$ and $\boldsymbol{A}_F'(\bar{\boldsymbol{\eta}}) = \bar{\boldsymbol{T}}_F$, and a unique $\bar{\boldsymbol{\theta}}_j \in \boldsymbol{\Theta}_j$ such that $\bar{\boldsymbol{\theta}}_j = \arg\max_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_j} \sum_{k=1}^{n_j} \mathcal{L}_j(\boldsymbol{X}_k^j; \boldsymbol{\theta}_j)$ and $\boldsymbol{A}_j'(\bar{\boldsymbol{\theta}}_j) = \bar{\boldsymbol{T}}_j$ for $j = 1, \ldots, J$. Note that it is not necessarily true that $\bar{\boldsymbol{\eta}} = \boldsymbol{h}(\bar{\boldsymbol{\theta}})$. Then, given $n_j > 0$ for $j = 1, \ldots, J$, we can rewrite the score equation in (1.48) as

$$\begin{bmatrix} \frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} \boldsymbol{h}'(\hat{\boldsymbol{\theta}})^T [\boldsymbol{A}_F'(\bar{\boldsymbol{\eta}}) - \boldsymbol{A}_F'(\hat{\boldsymbol{\eta}})] = \boldsymbol{A}_S'(\hat{\boldsymbol{\theta}}) - \boldsymbol{A}_S'(\bar{\boldsymbol{\theta}}). \tag{1.53}$$

Since there exists a differentiable function $\boldsymbol{h}(\cdot)$ such that $\boldsymbol{\eta} = \boldsymbol{h}(\boldsymbol{\theta})$, and both $\boldsymbol{A}_F'(\cdot)$ and $\boldsymbol{A}_j'(\cdot)$ for $j = 1, \ldots, J$ are differentiable and one-to-one functions, there must exist a differentiable function $\boldsymbol{g}(\cdot)$ such that $\boldsymbol{A}_F'(\boldsymbol{\eta}) = \boldsymbol{g}(\boldsymbol{A}_S'(\boldsymbol{\theta}))$. Denote the Jacobian of $\boldsymbol{g}(\cdot)$ as $\boldsymbol{G}(\cdot)$. The mean-value theorem implies that

$$\boldsymbol{A}_F'(\hat{\boldsymbol{\eta}}) = \boldsymbol{g}(\boldsymbol{A}_S'(\hat{\boldsymbol{\theta}})) = \boldsymbol{g}(\boldsymbol{A}_S'(\bar{\boldsymbol{\theta}})) + \tilde{\boldsymbol{G}} \cdot [\boldsymbol{A}_S'(\bar{\boldsymbol{\theta}}) - \boldsymbol{A}_S'(\hat{\boldsymbol{\theta}})], \tag{1.54}$$

where $\tilde{\boldsymbol{G}}$ is a $q \times p$ matrix with its $i$-th row being the $i$-th row of $\boldsymbol{G}(\boldsymbol{A}_S(\boldsymbol{\theta}))$ evaluated at some $\boldsymbol{A}_S(\tilde{\boldsymbol{\theta}})^{(i)}$ that lies on the line segment between $\boldsymbol{A}_S(\bar{\boldsymbol{\theta}})$ and

$\boldsymbol{A}'_S(\hat{\theta})$ for $i = 1, \ldots, q$. Plugging (1.53) into (1.54), we get

$$\boldsymbol{A}'_F(\hat{\boldsymbol{\eta}}) = \boldsymbol{g}(\boldsymbol{A}'_S(\bar{\theta})) + \tilde{\boldsymbol{G}} \cdot \begin{bmatrix} \frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} \boldsymbol{h}'(\hat{\theta})^T [\boldsymbol{A}'_F(\bar{\boldsymbol{\eta}}) - \boldsymbol{A}'_F(\hat{\boldsymbol{\eta}})]. \quad (1.55)$$

After subtracting both sides of (1.55) by $\boldsymbol{g}(\boldsymbol{A}'_S(\theta^*))$ and rearranging the terms, it is convenient to express (1.55) as

$$\begin{aligned} &\left[ \boldsymbol{I}_q + \tilde{\boldsymbol{G}} \cdot \begin{bmatrix} \frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} \boldsymbol{h}'(\hat{\theta})^T \right] [\boldsymbol{A}'_F(\hat{\boldsymbol{\eta}}) - \boldsymbol{A}'_F(\boldsymbol{\eta}^*)] \\ &= \tilde{\boldsymbol{G}} \cdot \begin{bmatrix} \frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} \boldsymbol{h}'(\hat{\theta})^T [\boldsymbol{A}'_F(\bar{\boldsymbol{\eta}}) - \boldsymbol{A}'_F(\boldsymbol{\eta}^*)] + [\boldsymbol{g}(\boldsymbol{A}'_S(\bar{\theta})) - \boldsymbol{g}(\boldsymbol{A}'_S(\theta^*))], \end{aligned} \quad (1.56)$$

where we use that $\boldsymbol{A}'_F(\boldsymbol{\eta}^*) = \boldsymbol{g}(\boldsymbol{A}'_S(\theta^*))$. Furthermore, applying the mean-value theorem on $\boldsymbol{A}'_F(\cdot)$, we have $\boldsymbol{A}'_F(\hat{\boldsymbol{\eta}}) - \boldsymbol{A}'_F(\boldsymbol{\eta}^*) = \hat{\boldsymbol{A}}''_F \cdot (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)$, where $\hat{\boldsymbol{A}}''_F$ is a $q \times q$ matrix with its $i$-th row being the $i$-th row of $\boldsymbol{A}''_F(\boldsymbol{\eta})$ evaluated at some $\hat{\boldsymbol{\eta}}^{(i)}$ that lies on the line segment between $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}^*$ for $i = 1, \ldots, q$. Similarly, we have $\boldsymbol{A}'_F(\bar{\boldsymbol{\eta}}) - \boldsymbol{A}'_F(\boldsymbol{\eta}^*) = \bar{\boldsymbol{A}}''_F \cdot (\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)$, where $\bar{\boldsymbol{A}}''_F$ is another $q \times q$ matrix with its $i$-th row being the $i$-th row of $\boldsymbol{A}''_F(\boldsymbol{\eta})$ evaluated at some $\bar{\boldsymbol{\eta}}^{(i)}$ that lies on the line

segment between $\bar{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}^*$. Hence, we have (1.56) implies

$$\begin{aligned}
&\left[\hat{\boldsymbol{A}}_F'' + \tilde{\boldsymbol{G}} \cdot \begin{bmatrix} \frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} \boldsymbol{h}'(\hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{A}}_F''\right](\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \\
&= \tilde{\boldsymbol{G}} \cdot \begin{bmatrix} \frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} \boldsymbol{h}'(\hat{\boldsymbol{\theta}})^T \bar{\boldsymbol{A}}_F'' \cdot (\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) + [\boldsymbol{g}(\boldsymbol{A}_S'(\bar{\boldsymbol{\theta}})) - \boldsymbol{g}(\boldsymbol{A}_S'(\boldsymbol{\theta}^*))].
\end{aligned} \tag{1.57}$$

Since $n_F \to \infty$ and $n_j \to \infty$ for $j = 1, \ldots, J$, the standard MLE theory on exponential family implies $\bar{\boldsymbol{\eta}} \xrightarrow{\mathbb{P}} \boldsymbol{\eta}^*$ and $\bar{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}^*$. Furthermore, from the proof of Theorem 1.3, we also have $\hat{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\eta}} \xrightarrow{\mathbb{P}} \boldsymbol{\eta}^*$. Hence, we must have $\tilde{\boldsymbol{G}} \xrightarrow{\mathbb{P}} \boldsymbol{G}(\boldsymbol{A}'(\boldsymbol{\theta}^*)), \boldsymbol{h}'(\hat{\boldsymbol{\theta}}) \xrightarrow{\mathbb{P}} \boldsymbol{h}'(\boldsymbol{\theta}^*), \bar{\boldsymbol{A}}_F'' \xrightarrow{\mathbb{P}} \boldsymbol{A}_F''(\boldsymbol{\eta}^*)$ and $\hat{\boldsymbol{A}}_F'' \xrightarrow{\mathbb{P}} \boldsymbol{A}_F''(\boldsymbol{\eta}^*)$. Given the definition of $\boldsymbol{g}(\cdot)$, we can write out $\boldsymbol{g}(\cdot)$ and $\boldsymbol{G}(\cdot)$ explicitly using function compositions as

$$\begin{aligned}
\boldsymbol{g}(\boldsymbol{A}_S'(\boldsymbol{\theta})) &= (\boldsymbol{A}_F' \circ \boldsymbol{h} \circ [\boldsymbol{A}_S']^{-1})(\boldsymbol{A}_S'(\boldsymbol{\theta})), \\
\boldsymbol{G}(\boldsymbol{A}_S'(\boldsymbol{\theta})) &= \boldsymbol{A}_F''(\boldsymbol{\eta})\boldsymbol{h}'(\boldsymbol{\theta})[\boldsymbol{A}_S''(\boldsymbol{\theta})]^{-1},
\end{aligned}$$

where the symbol $\circ$ represent function composition such that $f \circ g(\cdot) = f(g(\cdot))$.

Hence, with $\boldsymbol{G}(\boldsymbol{A}_S'(\theta^*))$ and $\boldsymbol{h}'(\theta^*)$, denote $\boldsymbol{D}_N(\theta^*)$ as

$$
\begin{aligned}
&\boldsymbol{D}_N(\theta^*) \\
&= \boldsymbol{A}_F''(\eta^*) + \boldsymbol{G}(\boldsymbol{A}_S'(\theta^*)) \begin{bmatrix} \frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J} \end{bmatrix} \boldsymbol{h}'(\theta^*)^T \boldsymbol{A}_F''(\eta^*) \\
&= \boldsymbol{A}_F''(\eta^*) + \boldsymbol{A}_F''(\eta^*)\boldsymbol{h}'(\theta^*) \begin{bmatrix} \frac{n_F}{n_1}[\boldsymbol{A}_1''(\theta_1^*)]^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \frac{n_F}{n_J}[\boldsymbol{A}_J''(\theta_J^*)]^{-1} \end{bmatrix} \boldsymbol{h}'(\theta^*)^T \boldsymbol{A}_F''(\eta^*) \\
&= \boldsymbol{A}_F''(\eta^*) + \boldsymbol{A}_F''(\eta^*) \left[ \sum_{j=1}^{J} \frac{n_F}{n_j} \boldsymbol{h}_j'(\theta^*)[\boldsymbol{A}_j''(\theta_j^*)]^{-1}\boldsymbol{h}_j'(\theta^*)^T \right] \boldsymbol{A}_F''(\eta^*) \\
&= \boldsymbol{A}_F''(\eta^*) \left[ [\boldsymbol{A}_F''(\eta^*)]^{-1} + \sum_{j=1}^{J} \frac{n_F}{n_j} \boldsymbol{h}_j'(\theta^*)[\boldsymbol{A}_j''(\theta_j^*)]^{-1}\boldsymbol{h}_j'(\theta^*)^T \right] \boldsymbol{A}_F''(\eta^*), \qquad (1.58)
\end{aligned}
$$

where $\boldsymbol{h}_j'(\theta)$ is a $q \times p_j$ matrix with components $[\boldsymbol{h}_j'(\theta)]_{a,b} = \partial\eta_a/\partial\theta_j^{(b)}$ for $j = 1, \ldots, J$. Note that $\boldsymbol{h}'(\theta) = [\boldsymbol{h}_1'(\theta); \ldots; \boldsymbol{h}_J'(\theta)]$. Since $\boldsymbol{A}_F''(\eta^*)$ is positive definite and $\boldsymbol{h}_j'(\theta^*)[\boldsymbol{A}_j'']^{-1}(\theta_j^*)\boldsymbol{h}_j'(\theta^*)^T$ is positive semi-definite for $j = 1, \ldots, J$, we see $\boldsymbol{D}_N(\theta^*)$ must also be positive definite and hence non-singular by Sylvester's law of inertia (Horn and Johnson, 2012, Theorem 4.5.8). Given the assumption that for $i = 1, \ldots, q$

$$
\lambda_i \left( \frac{n_F}{n_s} \sum_{j=1}^{J} \boldsymbol{h}_j'(\theta^*)[\boldsymbol{A}_j''(\theta_j^*)]^{-1}\boldsymbol{h}_j'(\theta^*)^T \right) = O\left(\frac{n_F}{n_s}\right),
$$

we must have $\lambda_i(\boldsymbol{D}_N(\theta^*)^{-1}) = O(n_s/n_F)$ and hence $\lim_{N\to\infty} \boldsymbol{D}_N(\theta^*)^{-1} = \mathbf{0}$.

Now, multiplying both sides of (1.57) by $\sqrt{n_F+n_s}\boldsymbol{D}_N(\theta^*)^{-1}$, we have

$$\sqrt{n_F+n_s}\boldsymbol{D}_N(\theta^*)^{-1}\left[\hat{\boldsymbol{A}}_F'' + \tilde{\boldsymbol{G}}\cdot\begin{bmatrix}\frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J}\end{bmatrix}\boldsymbol{h}'(\hat{\theta})^T\hat{\boldsymbol{A}}_F''\right](\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}^*)$$

$$= \sqrt{n_F+n_s}\boldsymbol{D}_N(\theta^*)^{-1}\tilde{\boldsymbol{G}}\cdot\begin{bmatrix}\frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J}\end{bmatrix}\boldsymbol{h}'(\hat{\theta})^T\bar{\boldsymbol{A}}_F''\cdot(\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}^*) \tag{1.59}$$

$$+\sqrt{n_F+n_s}\boldsymbol{D}_N(\theta^*)^{-1}[\boldsymbol{g}(\boldsymbol{A}_S'(\bar{\theta})) - \boldsymbol{g}(\boldsymbol{A}_S'(\theta^*))]. \tag{1.60}$$

Using the fact that $\sqrt{n_F+n_s}(\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{A}_F''(\boldsymbol{\eta}^*)^{-1})$ from the standard MLE asymptotic normality theory and the result that

$$\boldsymbol{D}_N(\theta^*)^{-1}\tilde{\boldsymbol{G}}\begin{bmatrix}\frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J}\end{bmatrix}\boldsymbol{h}'(\hat{\theta})^T\hat{\boldsymbol{A}}_F''$$

$$= \boldsymbol{D}_N(\theta^*)^{-1}\left[\boldsymbol{A}_F''(\boldsymbol{\eta}^*) + \tilde{\boldsymbol{G}}\begin{bmatrix}\frac{n_F}{n_1}\boldsymbol{I}_{p_1} & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{n_F}{n_J}\boldsymbol{I}_{p_J}\end{bmatrix}\boldsymbol{h}'(\hat{\theta})^T\hat{\boldsymbol{A}}_F''\right] - \boldsymbol{D}_N(\theta^*)^{-1}\boldsymbol{A}_F''(\boldsymbol{\eta}^*)$$

$$\xrightarrow{\mathbb{P}} \boldsymbol{I}_q,$$

we have terms in (1.59) converges in distribution to $\mathcal{N}(\mathbf{0}, \boldsymbol{A}_F''(\boldsymbol{\eta}^*)^{-1})$ by the Slutsky's theorem. For the terms in (1.60), the standard MLE asymptotic normality theory implies $\sqrt{n_j}(\bar{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, [\boldsymbol{A}_j''(\boldsymbol{\theta}_j^*)]^{-1})$ for $j = 1, \ldots, J$, we must have $\sqrt{n_F + n_s}\boldsymbol{D}_N(\boldsymbol{\theta}^*)^{-1}[\boldsymbol{g}(\boldsymbol{A}_S'(\bar{\boldsymbol{\theta}})) - \boldsymbol{g}(\boldsymbol{A}_S'(\boldsymbol{\theta}^*))] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{0})$. Given the fact that asymptotic normality with variance-covariance matrix being $\mathbf{0}$ is equivalent to convergence in probability, we claim that $\sqrt{n_F + n_s}\, \boldsymbol{D}_N(\boldsymbol{\theta}^*)^{-1}\, [\boldsymbol{g}(\boldsymbol{A}_S'(\bar{\boldsymbol{\theta}})) - \boldsymbol{g}(\boldsymbol{A}_S'(\boldsymbol{\theta}^*))] \xrightarrow{\mathbb{P}} \mathbf{0}$. Therefore, given the assumption $n_s = o(n_F)$ in case ii), we conclude that (1.57) implies

$$\sqrt{n_F}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{A}_F''(\boldsymbol{\eta}^*)^{-1}\right).$$

**Case iii)** $\lim_{N\to\infty} n_F/n_s = \infty$ subject to $n_j < \infty$ for $j = 1, \ldots, J$: Under this case, we have $n_F \to \infty$ and $n_j = o(n_F)$ for $j = 1, \ldots, J$. Dividing both sides of (1.13) by $N$, we see that

$$\frac{1}{N}\boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = \frac{n_F[\boldsymbol{\eta} \cdot \bar{\boldsymbol{T}}_F - A_F(\boldsymbol{\eta})] + o_p(1)}{n_F + o(n_F)}. \tag{1.61}$$

Since the right-hand side of (1.61) only contains $\boldsymbol{\eta}$, we can express $\hat{\boldsymbol{\eta}}$ as

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}\in\mathcal{E}} \frac{n_F}{n_F + o(n_F)}[\boldsymbol{\eta} \cdot \bar{\boldsymbol{T}}_F - A_F(\boldsymbol{\eta})] + o_p(\mathbf{1}),$$

where the first-order optimality condition implies

$$\frac{n_F}{n_F + o(n_F)}[\bar{\boldsymbol{T}}_F - \boldsymbol{A}'_F(\hat{\boldsymbol{\eta}})] + o_p(\mathbf{1}) = \mathbf{0}.$$

Therefore, by the standard MLE asymptotic normality theory on exponential family distribution (Bickel and Doksum, 2001, Theorem 5.4.2), we have

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}^* - \frac{n_F}{n_F + o(n_F)}\boldsymbol{A}''_F(\boldsymbol{\eta}^*)[\bar{\boldsymbol{T}}_F - \boldsymbol{A}'(\boldsymbol{\eta}^*)] + o_p\left(\frac{\mathbf{1}}{\sqrt{n_F}}\right) + o_p(\mathbf{1}),$$

and

$$\sqrt{n_F}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{A}''_F(\boldsymbol{\eta}^*)).$$

□

## 1.4 Numerical Study

### 1.4.1 Synthetic Problem

Here we provide an example that is a simpler version of the work in Zhao and Spall (2016). Assume there exists two subsystems, whose outputs are Bernoulli distributed with $\rho_1^* = 0.4$ and $\rho_2^* = 0.2$, respectively. The full system output are normally distributed with mean $\mu^* = \rho_1^* + \rho_2^* = 0.6$ and variance $\sigma^{2*} = \rho_1^*(1 - \rho_1^*) + \rho_2^*(1 - \rho_2^*) = 0.4$. The first numerical result focuses on the

confidence interval of the estimates as the subsystem sample size goes to infinity. We fixed $n_F = 10$ and each confidence interval below is computed by 0.025 and 0.975 empirical quantiles basing on 1,000 replicates and the estimates are found by the MATLAB "fmincon" function with the initial guess $\hat{\theta}_0 = [0.5, 0.5]$ and parameter space $\Theta = [0, 1]^2$.

It is convincing from the Table 1.1 that subsystem parameters converge to the true parameters and as a consequence, the convergence property also holds for the full system parameters. Another aspect of the convergence argument is to see what happened when the full system sample size goes to infinity. Fixed $n_1 = n_2 = 10$, Table 1.2 verifies that the full system estimates indeed converge to the true parameters. Similar to the first result, we also see that the confidence interval becomes narrower and it is expected that as the full system sample becomes unbounded, our maximum likelihood estimates converge to the true parameters. The last result, Table 1.3, shows that the full system parameters converge to the true parameters when both the full system and subsystems sample sizes go to infinity. For simplicity, we assume the sample sizes are equal across all the subsystems and the full system.

### 1.4.2 Cold-Formed Steel Shear Wall

In a second application, we apply our method to the cold-formed steel (CFS) structural system, where the fasteners are viewed as the subsystems and the

**Table 1.1:** 95% confidence interval for full system and subsystem estimates with fixed full system sample size $n = 10$

| | $\hat{\rho}_1$ | $\hat{\rho}_2$ | $\hat{\mu}$ | $\hat{\sigma}^2$ |
|---|---|---|---|---|
| $n_j = 10$ | [0.151, 0.658] | [0.000, 0.434] | [0.330, 0.904] | [0.244, 0.476] |
| $n_j = 10^2$ | [0.309, 0.496] | [0.122, 0.275] | [0.476, 0.710] | [0.341, 0.439] |
| $n_j = 10^3$ | [0.370, 0.428] | [0.177, 0.225] | [0.563, 0.639] | [0.385, 0.415] |
| $n_j = 10^4$ | [0.390, 0.410] | [0.193, 0.208] | [0.588, 0.613] | [0.395, 0.405] |
| $n_j = 10^5$ | [0.397, 0.403] | [0.197, 0.202] | [0.596, 0.604] | [0.398, 0.401] |

**Table 1.2:** 95% confidence interval for full system estimates with fixed subsystem sample sizes $n_1 = n_2 = 10$

| | $\hat{\mu}$ | $\hat{\sigma}^2$ |
|---|---|---|
| $n = 10$ | [0.347, 0.878] | [0.239, 0.476] |
| $n = 10^2$ | [0.500, 0.724] | [0.313, 0.445] |
| $n = 10^3$ | [0.563, 0.639] | [0.366, 0.425] |
| $n = 10^4$ | [0.587, 0.612] | [0.390, 0.418] |
| $n = 10^5$ | [0.591, 0.604] | [0.396, 0.417] |

**Table 1.3:** 95% confidence interval for full system estimates with the same full system and subsystem sample sizes $n = n_1 = n_2$

| | $\hat{\mu}$ | $\hat{\sigma}^2$ |
|---|---|---|
| $n = n_j = 10$ | [0.345, 0.879] | [0.241, 0.474] |
| $n = n_j = 10^2$ | [0.525, 0.688] | [0.360, 0.433] |
| $n = n_j = 10^3$ | [0.573, 0.626] | [0.387, 0.412] |
| $n = n_j = 10^4$ | [0.592, 0.609] | [0.396, 0.404] |
| $n = n_j = 10^5$ | [0.597, 0.603] | [0.399, 0.401] |

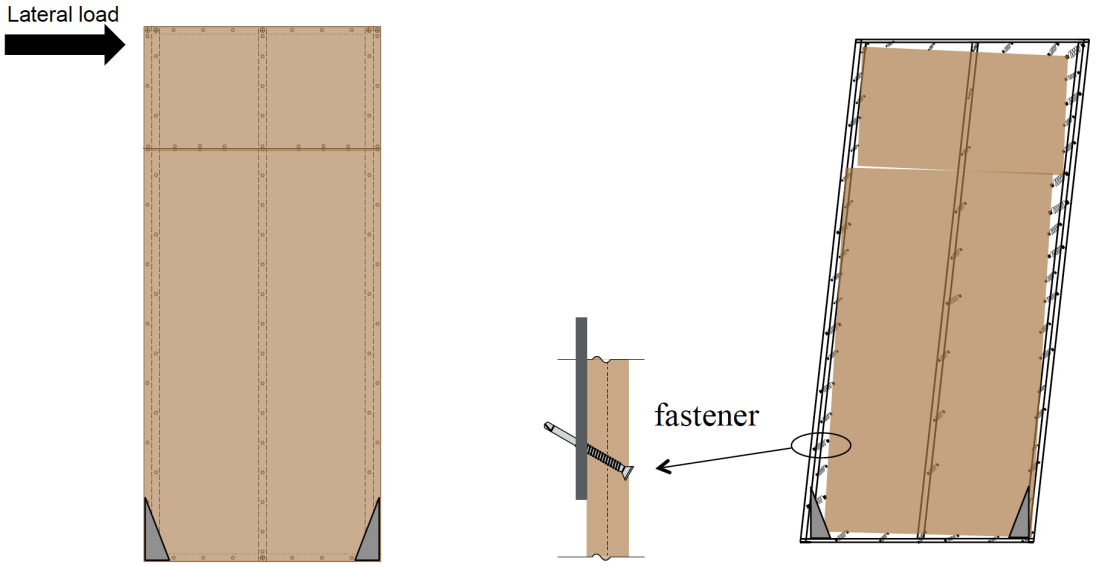
shear wall is viewed as a full system. Using data independently collected from multiple sources, we are able to provide a much better estimation of the shear wall strength when compared with the simple one-sample estimation method (Wang et al., 2018a).

The cold-formed steel (CFS) structural systems are commonly used for low and mid-rise constructions. A typical CFS shear wall building with associated shear walls is shown in Figure 1.2. A shear wall is a lateral force resistance system to bear seismic or wind load. Commonly, wood sheathing, such as oriented strand board (OSB), is screw-fastened to CFS studs and tracks to develop lateral shear stiffness and strength.

**Figure 1.2:** Cold-formed steel building and shear wall

As the wall is sheared, an incompatibility exists between the CFS framing, which is largely deforming as a parallelogram, and the wood sheathing that re-

mains nearly rectangular and primarily undergoes rigid body translation and rotation because of its large in-plane rigidity. The incompatibility between the deformed frame and sheathing causes a relative displacement that must be accommodated at the fasteners, as shown in Figure 1.3.




(a) undeformed shear wall (b) shear wall with deformed shape

**Figure 1.3:** Cold-formed steel shear wall under lateral load

The fastener is the key subsystem in the shear wall system and it is shown in Buonopane et al. (2015) that shear wall capacity is determined by fastener strength. The CFS framed wood sheathed shear wall capacity is defined in the North American Standard for Seismic Design of Cold-Formed Steel Structural Systems (AISI-S400-15) (AISI, 2015). Shear wall capacities are based on previous shear wall experimental tests from the U.S. or Canada. For most shear wall configurations in the design specification, the number of shear wall tests

is very small, and the estimate is simply based on the average of all shear wall test data.

In this experiment, the shear wall capacity is estimated based on combining the shear wall experimental test data (full system) and the fastener test data (subsystem). This combination of data sources should allow for more accurate estimates than the single source above. The full system data is measured during the shear wall loading process and it refers to the shear wall peak strength (capacity) before failure. Shear wall capacity data are from shear wall cyclic test at Branston et al. (2006), which is of particular interest for civil engineers than monotonic loading results.

On the subsystem level, the data being measured are the fastener capacity before its failure. Isolated physical tests for fasteners were conducted at Johns Hopkins University (Peterman et al., 2014). The fasteners in the subsystem tests are in the same loading protocol as they are in the full system shear wall experiment. The fastener test data are provided in Table 1.4, where the configuration is the same as it is in the shear wall.

**Table 1.4:** Subsystem Capacity (kip)

| | | | | |
|---|---|---|---|---|
| Fastener (m=8) | 0.427 | 0.495 | 0.541 | 0.414 |
| | 0.430 | 0.547 | 0.507 | 0.529 |

Assume the test data are collected statistically independently from all data sources. According to Bian et al. (2017), the full system data follows a normal

distribution. Furthermore, after conducting the Lilliefors goodness-of-fit test, all subsystem data are assumed to follow the same log-normal distribution, where only the mean parameter is unknown and the variance is known due to AISI (2015). From a physical point of view, since the loading transferred to the shear wall is carried by each fastener so the capacity of the shear wall is the summation of fastener capacity. From results of structural analysis using civil engineering software OpenSees (McKenna, 2011), we have made the following assumptions on the data and derived the relationships between the full system and subsystem, which are shown as

$$\text{Subsystem: } X_j \sim \mathcal{LN}(\mu_S, \sigma_S^2),$$

$$\text{Full system: } Y_k \sim \mathcal{N}(\mu_F, \sigma_F^2),$$

where $\mathcal{LN}$ denotes the log-normal distribution, $\mu_F = 10.03e^{\mu_S + \frac{1}{2}\sigma_S^2}$ and $\sigma_F^2 = 0.0212e^{4\mu_S + 2\sigma_S^2}$. Those relationships are obtained from a regression of 1,000 simulated systems under cyclic loading with each subsystem capacity following the known capacity mean and standard deviation. According to AISI (2015), the ratio between subsystem standard deviation and mean is set to $0.13$. In terms of the log-normal distribution parameters $\mu_S$ and $\sigma_S^2$, we have $\operatorname{Var} X/(\mathbb{E}[X])^2 = (e^{\sigma_S^2 - 1}e^{2\mu_S + \sigma_S})/(e^{\mu_S + \frac{1}{2}\sigma_S})^2 = 0.13^2$, which gives $\sigma_S^2 = 0.168$.

The `fmincon` function in Matlab is used to find the minimizer of the con-

strained nonlinear function $-\mathcal{L}(\theta)$ (negative log-likelihood function). The final estimated $\hat{\theta}$ is $-1.1548$, which gives the estimated mean of shear wall capacity, $\hat{\mu}_F = 3.1871$ and estimated variance of shear all capacity, $\hat{\sigma}_F^2 = 0.0022$. Using the asymptotic normal distribution, we have $\sqrt{N}(\hat{\mu}_F - \mu_F) \sim N(0, 0.0028)$, which leads to a 95% confidence interval $(3.1696, 3.2047)$ with width $0.0350$. As a comparison, if one only uses the full system sample from G1, it will give estimates $\tilde{\mu}_F = \bar{Y} = 2.9883, \tilde{\sigma}_F^2 = s_Y^2 = 0.0059$ and 95% confidence interval $(2.9080, 3.0686)$ with width $0.1606$. It is worth noting that the simple average estimate falls outside our confidence interval. It is clear that by considering both full system and subsystem data and taking into account the connections between the full system and subsystem, MLE reveals more information on both full system and subsystem level. In terms of the width of confidence interval, our new method improves the accuracy by $(0.1606 - 0.0059)/0.1606 = 96.33\%$.

### 1.4.3 Sensor Network

With recent developments of data collection devices, such as mobile sensors and handheld computers, data in various forms are becoming more plentiful. How to properly integrate all the data, however, is generally an open problem. In this subsection, we study the integration of multilevel data for applications in sensor networks (Wang and Spall, 2020). In particular, the proposed method is applied to locate a target using multiple unmanned aerial vehicles and we

present a spatial searching problem modified from Sun et al. (2016) to illustrate the performance of the proposed method. Consider a mobile sensor network where multiple UAVs and a Doppler radar are deployed to locate a target. Because the Doppler radar can cover a wide range of area, it is viewed as the full system here. Denote $\boldsymbol{Y} = [Y_1, Y_2, Y_3]^T$ as the noisy measurements of the target generated by the Doppler radar, where the components of $\boldsymbol{Y}$ represent the target location in the Cartesian coordinate system with the Doppler radar at the origin. Based on Sun et al. (2016), assume a multivariate normal distribution on the Doppler measurements, i.e., $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}^F, \boldsymbol{\Sigma}^F)$ where $\boldsymbol{\mu}^F$ is the unknown true location of the target and $\boldsymbol{\Sigma}^F$ is the unknown covariance matrix. Although using only the Doppler measurement can provide an estimate of the target location, it is often less accurate and has a limited number of measurements due to operational cost. Hence, we aim to improve the estimation by combining the information from the individual UAV measurements.

We consider the multiple UAVs as the subsystems, where each individual UAV can only detect a small neighborhood of itself and needs to work with other UAVs to locate the target. One potential motivation to have local UAVs is to make them more hidden than the broad signal, which can help to avoid reveal source information too much, and the few signals the better in real operations due to security issues. Similar to the full system outputs, assume the subsystem outputs also follow the multivariate normal distributions, i.e.,

$\boldsymbol{X}^{(j)} = [X_1^{(j)}, X_2^{(j)}, X_3^{(j)}]^T \sim \mathcal{N}(\boldsymbol{\mu}^{(j)}, \boldsymbol{\Sigma}^{(j)})$, which measures the difference of two nearest UAV positions or the difference of the target and its nearest UAV positions in the Cartesian coordinate system. Although UAVs are (of course) allowed to move in space, if the time scale of the sampling is much smaller than the time scale of the movement dynamics, we can think the positions of the UAVs are fixed during the short sampling period. For example, denote the locations of two nearest UAVs as $[l_1^{(1)}, l_2^{(1)}, l_3^{(1)}]$ and $[l_1^{(2)}, l_2^{(2)}, l_3^{(2)}]$, respectively. Then, we have $\mu_1^{(1)} = l_1^{(2)} - l_1^{(1)}$, $\mu_2^{(1)} = l_2^{(2)} - l_2^{(1)}$ and $\mu_3^{(1)} = l_3^{(2)} - l_3^{(1)}$. Figure 1.4 illustrates the framework of this location detection problem with three UAVs and one Doppler radar. Note that we consider the measurements of the different locations between the UAVs as the subsystem measurements, not the locations of the UAVs themselves. This is why Figure 1.4 contains three UAVs, but shows four different $\boldsymbol{X}^{(j)}$ for $j = 1, \ldots, 4$. Note that depending on the location of the target, not all the UAVs are active or reporting information about the target. For simplicity, assume that within the detection range of every UAV, there is only one other UAV in order to maximize the total coverage area.

Under the Cartesian coordinate system, since the Doppler radar measurements can be constructed as $\boldsymbol{Y} = \sum_{j=1}^{4} \boldsymbol{X}^{(j)}$, we link the full system and subsystem parameter by defining the system structure function as $\boldsymbol{\mu}^F = \sum_{j=1}^{J} \boldsymbol{\mu}^{(j)}$ and $\boldsymbol{\Sigma}^F = \sum_{j=1}^{J} \boldsymbol{\Sigma}^{(j)}$. A similar construction can be found in Sun et al. (2016) with the difference being on how UAVs are allowed to communicate with each

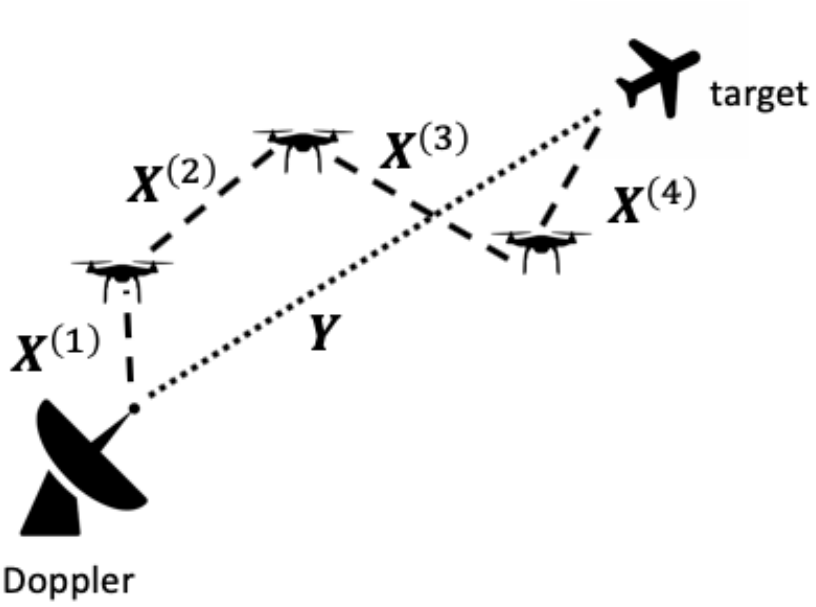



**Figure 1.4:** Conceptual illustration of the sensor network: Doppler radar and UAVs are combined to provided information on target.

other and whether the UAVs are fixed on predetermined orbits. Given a total of $n_F$ measurements for the full system and $n_j$ test measurements for the $j$-th subsystem that are independent from each other, the overall likelihood function has the form

$$\begin{aligned}\mathcal{L}(\theta) = &-\frac{1}{2}\sum_{k=1}^{n_F}(\boldsymbol{Y}_k - \boldsymbol{\mu}^F)^T\boldsymbol{\Sigma}^F(\boldsymbol{Y}_k - \boldsymbol{\mu}^F) - \frac{n_F}{2}\log(\det\boldsymbol{\Sigma}_F)\\ &-\frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\boldsymbol{X}_i^{(j)} - \boldsymbol{\mu}^{(j)})^T\boldsymbol{\Sigma}^{(j)}(\boldsymbol{X}_i^{(j)} - \boldsymbol{\mu}^{(j)})\\ &-\frac{1}{2}\sum_{j=1}^{J}n_j\log(\det\boldsymbol{\Sigma}^{(j)}) + \text{constant}.\end{aligned}$$

We consider a case with 4 subsystems and assume the following true parameter

values,

$$\mu^{*(1)} = [-0.90, 1.42, 5.86]^T, \mu^{*(2)} = [2.35, -0.89, 2.09]^T,$$

$$\mu^{*(3)} = [0.35, 0.32, 1.95]^T, \mu^{*(4)} = [0.10, -0.25, 2.033]^T,$$

$$\mathbf{\Sigma}^{*(1)} = \mathbf{\Sigma}^{*(2)} = \mathbf{\Sigma}^{*(3)} = \mathbf{\Sigma}^{*(4)} = \begin{bmatrix} 12 & 1.2 & 1.2 \\ 1.2 & 12 & 1.2 \\ 1.2 & 1.2 & 12 \end{bmatrix}.$$

The true position values $\mu^*$ are chosen such that the distances between the Doppler radar and each actives UAV are 6km, 8km, and 10km, respectively, matching the orbit radii assumptions in the numerical study of Sun et al. (2016). The distance between the Doppler radar and the target is then chosen to be 12km. The covariance values $\Sigma^*$ are also based on the UAVs measurement noise assumptions in Sun et al. (2016).

Figure 1.5 shows the confidence interval for the target position under $n_F = 5$ for the full system measurements and various subsystem measurements, where for simplicity assume that all the subsystems have the same number of measurements, i.e., $n_1 = n_2 = n_3 = n_4$. The solid lines represent the true parameter values for $\mu_1^F = 1.9, \mu_2^F = 0.6$ and $\mu_3^F = 11.833$. The dotted lines represent the 95% confidence interval for the true target position based on the asymptotic normality from Theorem 1.4.

Based Figure 1.5, it is clear that by combining all the data from different

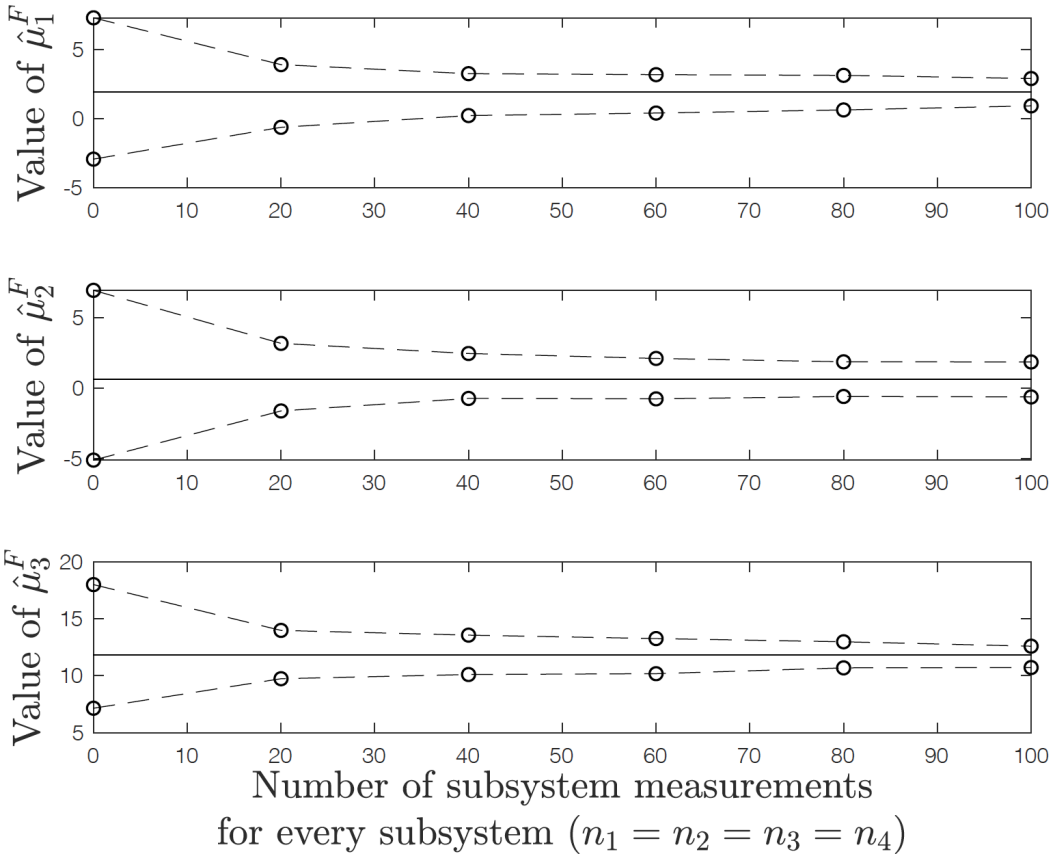



**Figure 1.5:** 95% confidence intervals for the true target position based on the asymptotic normality from Theorem 1.4. All plots assume $n = 5$.

sources, the estimate of the target position becomes much more accurate as the confidence intervals are much narrower. As more data are collected, the final estimates are converging towards the true parameter value and the likelihood of successfully identifying the target is significantly improved.

Another application is based on Duan et al. (2019), where the goal is to detect the position of seven markers on the ground by integrating data from two UAVs, i.e., a tanker UAV and a receiver UAV. Here, we consider a simpler version of detecting only one marker, i.e., the target, on the ground since there is no fundamental difference between the different number of markers in terms of the general methodology. The receiver UAV flying at a lower altitude can collect noisy measurements $\boldsymbol{X}^{(1)} \sim \mathcal{N}(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)})$ of the marker. The tanker UAV flying at a higher altitude can collect noisy measurements $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ of the marker and relative position measurements $\boldsymbol{X}^{(2)} \sim \mathcal{N}(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(2)})$ with respect to the receiver UAV.

In our framework, $\boldsymbol{Y}$ is corresponding to the full system measurements and $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are corresponding to the subsystem measurements. Under the standard Cartesian coordinate system, we have $\mu_Y = \mu^{(1)} + \mu^{(2)}$. Assume the following true parameter values

$$\boldsymbol{\mu}^{(1)} = [20, 10]^T, \boldsymbol{\mu}^{(2)} = [3, -8]^T,$$

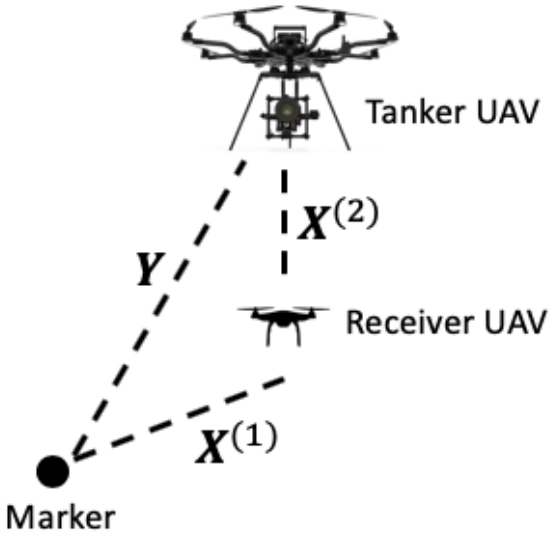



**Figure 1.6:** Tanker UAV and receiver UAV are combined to provided information about the marker.

$$\Sigma^{(1)} = \Sigma^{(2)} = \Sigma_Y = \begin{bmatrix} 5 & 0.1 \\ 0.1 & 5 \end{bmatrix}.$$

We compute the distance between the estimated and the true marker position under various full system sample size $n$ and subsystem sample sizes $n_1$ and $n_2$. The results are presented in Table 1.5 below and it is clear that as more data are collected the estimated position becomes more accurate.

**Table 1.5:** Accuracy of the estimated marker position

| | $n_1 = n_2 = 0$ | $n_1 = n_2 = 10$ | $n_1 = n_2 = 100$ |
|---|---|---|---|
| $n = 10$ | 0.9013 | 0.7291 | 0.3839 |
| $n = 100$ | 0.2906 | 0.2472 | 0.2414 |

## 1.5 Conclusion

In this chapter, we have studies the framework of the system of subsystems, where we generalize the distribution assumptions on both the full system and subsystem outputs to multivariate exponential family distribution. With the results of this chapter, we can now apply the full system and subsystems framework to a much broader class of problems than the previous studies. The convergence and asymptotic distribution of the MLE are also established. The proposed MLE approach has a solid theoretical foundation and works well in this non-trivial setting, where the likelihood function is clearly not a joint density function of i.i.d. data. For future works, one can consider further generalize the model to accommodate the time-varying effects, such as models for deterioration or consumer behaviors. Another direction is to jointly estimate multiple state-space models, where a user can seek to explore the relationships between the models in order to improve the overall estimation accuracy by integrating the data from different models.

# Chapter 2

# Connecting System Identification to Stochastic Optimization

## 2.1 Recursive Maximum Likelihood Estimate

In the previous chapter, we see that system identification is essentially finding the MLE of some likelihood function and studying its limiting behavior. If we denote $n$ as the number of sample points, we can write MLE as $\hat{\theta}_n^{\mathrm{ML}}$ and the corresponding log-likelihood function being $\ell_n(\theta) = (1/n)\sum_{i=1}^{n} \ell(\boldsymbol{x}_i; \theta)$ for some log-likelihood function $\ell(\boldsymbol{x}_i; \theta)$ and the sample point $\boldsymbol{x}_i$. Given the assumption that all the sample points are i.i.d., we see that, under some mild conditions, the SLLN guarantees that $\ell_n(\theta) \overset{\text{a.s.}}{\to} L(\theta)$ for some loss function $L(\theta) = \mathbb{E}[\ell(\boldsymbol{x}; \theta)]$, where the expectation is taken over the randomness in $\boldsymbol{x}$.

Since $\ell_n(\theta)$ is essentially a function of the sample points, it can be viewed as a random variable. If we drop the log-likelihood function requirement on $\ell(\boldsymbol{x}; \theta)$ and consider it as some general function, the MLE procedure is simply seeking to find the maximum value of $L(\theta)$ (or equivalently the minimum of $-L(\theta)$) based on the information of $\ell(\boldsymbol{x}_i; \theta)$. Since $\ell_n(\theta)$ is a random variable, we can express it as $\ell_n(\theta) = L(\theta) + \varepsilon_n(\theta)$ with $\varepsilon_n(\theta)$ being some random noise such that $\mathbb{E}[\varepsilon_n(\theta)] = 0$. In other words, we seek to find the optimal value of $L(\theta)$ based on its noisy measurements, which naturally fits into the framework of stochastic optimization.

In stochastic optimization, a common procedure is to iteratively find the optimal point of $L(\theta)$ with some noisy measurements of either the loss function or its gradient. If we denote $\theta^*$ as the optimal point of $L(\theta)$ and $\hat{\theta}_n^{\text{SO}}$ as the solution found by some stochastic optimization algorithm at iteration $n$, we see that both $\hat{\theta}_n^{\text{ML}}$ and $\hat{\theta}_k^{\text{SO}}$ are seeking to approach to $\theta^*$. Moreover, theoretical analyses are often interested in under what conditions we can have $\hat{\theta}_n^{\text{ML}} \to \theta^*$ or $\hat{\theta}_n^{\text{SO}} \to \theta^*$, and what is the rate of the convergence.

Let us consider the recursive MLE example from Spall (2005, Chapter 3) to better illustrate the similarities and differences between $\hat{\theta}_n^{\text{ML}}$ and $\hat{\theta}_n^{\text{SO}}$. Given a simple linear regression problem, where the sample points are modeled as

$$z_i = \boldsymbol{h}_i^T \theta + v_i \text{ with } v_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Under the independent Gaussian distribution assumption, the MLE approach is equivalent to find the optimal value of the mean-square error such that

$$\hat{\theta}_n^{\mathrm{ML}} = \arg\min \ell_n(\theta) = \arg\min \frac{1}{2n} \sum_{i=1}^{n} (z_i - \boldsymbol{x}_i^T \theta)^2, \tag{2.1}$$

with $\mathbb{E}[\ell_n(\theta)] = L(\theta)$ such that $L(\theta) = 1/(2n)\mathbb{E}[\sum_{i=1}^{n}(z_i - \boldsymbol{x}_i^T\theta)^2]$. It is easy to see that (2.1) has the batch least-squares solution $\hat{\theta}_n^{\mathrm{ML}} = [\boldsymbol{X}_n^T \boldsymbol{X}_n]^{-1} \boldsymbol{X}_n^T \boldsymbol{Z}_n$ with $\boldsymbol{X}_n$ being the concatenated matrix of $\boldsymbol{x}_i^T$ row vector and $\boldsymbol{Z}_n = [z_1, \ldots, z_n]^T$.

To see how $\hat{\theta}_n^{\mathrm{ML}}$ in (2.1) can be related to $\hat{\theta}_n^{\mathrm{SO}}$, we consider the Newton-Raphson method as our stochastic optimization algorithm. From (2.1), we have $\ell_{n+1}(\theta) = n/(n+1)\ell_n(\theta) + 1/[2(n+1)](z_{n+1} - \boldsymbol{x}_{n+1}^T\theta)^2$ and

$$\begin{aligned} \left.\frac{\partial \ell_{n+1}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_n^{\mathrm{SO}}} &= \frac{n}{n+1} \left.\frac{\partial \ell_n(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_n^{\mathrm{SO}}} + \frac{\boldsymbol{x}_{n+1}(\boldsymbol{x}_{n+1}^T \hat{\theta}_n^{\mathrm{SO}} - z_{n+1})}{n+1} \\ &\approx \frac{\boldsymbol{x}_{n+1}(\boldsymbol{x}_{n+1}^T \hat{\theta}_n^{\mathrm{SO}} - z_{n+1})}{n+1}, \end{aligned}$$

where we roughly assume $\partial \ell_n(\theta)/\partial\theta|_{\theta=\hat{\theta}_n^{\mathrm{SO}}} \approx 0$ despite some small discrepancy. To write out the Newton-Raphson iteratively updating formula, we also need to compute the Hessian of $\ell_{n+1}(\theta)$, which gives

$$\frac{\partial^2 \ell_{n+1}(\theta)}{\partial\theta\partial\theta^T} = \frac{1}{n+1} \sum_{i=1}^{n+1} \boldsymbol{x}_i \boldsymbol{x}_i^T = \frac{1}{n+1}(\boldsymbol{P}_{n+1}^{-1} - \boldsymbol{P}_0^{-1}),$$

where $\boldsymbol{P}_{n+1}^{-1} = \boldsymbol{P}_n^{-1} + \boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}^T$ with some user-specified initial conditions of $\boldsymbol{P}_0$.

If we assume $\boldsymbol{P}_0^{-1}$ is small relatively to $\boldsymbol{P}_{n+1}^{-1}$, the Newton-Raphson iteratively updating formula gives

$$\hat{\theta}_{n+1}^{\text{SO}} \approx \hat{\theta}_n^{\text{SO}} - \boldsymbol{P}_{n+1}\boldsymbol{x}_{n+1}(\boldsymbol{x}_{n+1}^T\hat{\theta}_n^{\text{SO}} - z_{n+1}). \tag{2.2}$$

However, we see that (2.2) can also be used to express $\hat{\theta}_{n+1}^{\text{ML}}$ as

$$\hat{\theta}_{n+1}^{\text{ML}} \approx \hat{\theta}_n^{\text{ML}} - \boldsymbol{P}_{n+1}\boldsymbol{x}_{n+1}(\boldsymbol{x}_{n+1}^T\hat{\theta}_n^{\text{ML}} - z_{n+1}) \tag{2.3}$$

with some user-specified initial conditions of $\hat{\theta}_0^{\text{ML}}$ and $\boldsymbol{P}_0$ (Spall, 2005, Section 3.2.4). Therefore, it is reasonable to state that $\hat{\theta}_{n+1}^{\text{ML}}$ and $\hat{\theta}_{n+1}^{\text{SO}}$ are closely related and share several interesting theoretical properties, such as the fast rate of convergence. The rest of this work will focus on stochastic optimization.

## 2.2 Stochastic Optimization

Stochastic Optimization is a large field and there are numerous works on developing algorithms aiming to minimize a loss function $L(\theta)$. However, the primary focus of this work will be on algorithms that only use the noisy function measurements $\ell(\theta, \boldsymbol{v})$, where $\boldsymbol{v}$ denotes the randomness embedded in the function measurement. Introduced in the 1950s, Kiefer and Wolfowitz (1952) and Blum (1954) use the standard finite-difference (FD) gradient approxima-

tion methods to perform a gradient-descent-type algorithm to iteratively find the optimal solution. Mathematically, at the iteration $k$ with the estimate $\hat{\theta}_k$, the two-sided FD gradient approximation uses the following formula to estimate the gradient of the loss function $\boldsymbol{g}(\hat{\theta}_k) = \partial L(\theta)/\partial\theta|_{\theta=\hat{\theta}_k}$ by

$$\hat{\boldsymbol{g}}_k^{\mathrm{FD}}(\hat{\theta}_k) = \frac{1}{2c_k}\begin{bmatrix} \ell(\hat{\theta}_k + c_k\boldsymbol{u}_1, \boldsymbol{v}_k^{(1+)}) - \ell(\hat{\theta}_k - c_k\boldsymbol{u}_1, \boldsymbol{v}_k^{(1-)}) \\ \vdots \\ \ell(\hat{\theta}_k + c_k\boldsymbol{u}_p, \boldsymbol{v}_k^{(p+)}) - \ell(\hat{\theta}_k - c_k\boldsymbol{u}_p, \boldsymbol{v}_k^{(p-)}) \end{bmatrix}, \tag{2.4}$$

where $c_k > 0$ is the perturbation step size, $\boldsymbol{u}_j$ is the vector of all zeros except the $j$-th component being $1$ for $j = 1, \ldots, p$, and $\boldsymbol{v}_k^{(1+)}, \boldsymbol{v}_k^{(1-)}, \ldots, \boldsymbol{v}_k^{(p+)}, \boldsymbol{v}_k^{(p-)}$ represents the total random effects. In order to get an accurate gradient approximation with respect to the true gradient $\boldsymbol{g}(\hat{\theta}_k)$, the perturbation step size $c_k$ is often chosen to be some small scalar value. It is well-known that the bias of the FD gradient approximation satisfies $\mathbb{E}[\hat{\boldsymbol{g}}_k^{\mathrm{FD2}}(\hat{\theta}_k)|\hat{\theta}_k] - \boldsymbol{g}(\hat{\theta}_k) = O(c_k^2)$, where the expectation is taken over the randomness in $\ell(\hat{\theta}_k, \cdot)$. Given the noisy function measurement $\ell(\hat{\theta}_k, \boldsymbol{v}_k)$, we can denote the noisy term $\varepsilon_k(\hat{\theta}_k, \boldsymbol{v}_k)$ as

$$\varepsilon_k(\hat{\theta}_k, \boldsymbol{v}_k) = L(\hat{\theta}_k) - \ell(\hat{\theta}_k, \boldsymbol{v}_k),$$

where it is often assumed that $\mathbb{E}[\varepsilon_k(\hat{\theta}_k + c_k\boldsymbol{u}_j, \boldsymbol{v}_k^{(j+)}) - \varepsilon_k(\hat{\theta}_k - c_k\boldsymbol{u}_j, \boldsymbol{v}_k^{(j-)})|\hat{\theta}_k] = 0$ for $j = 1, \ldots, p$. Note that the assumptions on the noisy function measurements

can be easily satisfied if we assume $\mathbb{E}[\ell(\theta, \boldsymbol{v})] = L(\theta)$ for any $\theta$, which is a common assumption for problems with zero-mean additive noise. One disadvantage of the two-sided FD gradient approximation, however, is that a total of $2p$ function measurements are required to construct one gradient approximation. This makes the method less efficient for high-dimensional problems or when it is expensive to evaluate $\ell(\hat{\theta}_k, \cdot)$. A slightly improved version of the two-sided FD gradient approximation in terms of the number of function measurements is the one-sided FD gradient approximation, which has the formula

$$\hat{\boldsymbol{g}}_k^{\text{FD1}}(\hat{\theta}_k) = \frac{1}{c_k} \begin{bmatrix} \ell(\hat{\theta}_k + c_k \boldsymbol{u}_1, \boldsymbol{v}_k^{(1+)}) - \ell(\hat{\theta}_k, \boldsymbol{v}_k^{(0)}) \\ \vdots \\ \ell(\hat{\theta}_k + c_k \boldsymbol{u}_p, \boldsymbol{v}_k^{(p+)}) - \ell(\hat{\theta}_k, \boldsymbol{v}_k^{(0)}) \end{bmatrix},$$

where $\boldsymbol{v}_k^{(0)}$ is the random effect of the base function measurement $\ell(\hat{\theta}_k, \boldsymbol{v}_k^0)$. Although the number of function measurements is reduced to $p + 1$, the accuracy of the gradient approximation is poorer since the bias now becomes $\mathbb{E}[\hat{\boldsymbol{g}}_k^{\text{FD1}}(\hat{\theta}_k)|\hat{\theta}_k] - \boldsymbol{g}(\hat{\theta}_k) = O(c_k)$. Given the usual condition that the extra error in the gradient approximation cannot be compensated by the benefit of using less function measurements, the one-sided FD gradient approximation is less desirable for many practical problems.

Nonetheless, since the gradient approximation requires a total of $O(p)$ function measurements at each iteration, it is considered to be significantly less

efficient for problems that are expensive to evaluate or have high dimensions. To reduce the required number of function measurements, Spall (1992) proposes the simultaneous perturbation stochastic approximation (SPSA) algorithm, which achieves approximately the same level of accuracy as the standard FD stochastic approximation (FDSA) algorithm, but uses only two function measurements to estimate the gradient. The SP idea proposed is to generate a $p$-dimensional random perturbation vector $\boldsymbol{\Delta}_k = [\Delta_{k1}, \ldots, \Delta_{kp}]^T$, where every component of $\boldsymbol{\Delta}_k$ is independently and identically distributed (i.i.d.) satisfying some regularity conditions specified in Spall (1992, A2′). A common, but not mandatory, choice of $\boldsymbol{\Delta}$ is to have its every component sampled uniformly from $\{-1, +1\}$, i.e., $\Delta_j \overset{\text{i.i.d.}}{\sim}$ Bernoulli$\{-1, +1\}$ for $j = 1, \ldots, p$. Given the current estimate $\hat{\theta}_k$ at the iteration $k$, the function is then evaluated at two design levels, i.e., $\hat{\theta}_k + c_k\boldsymbol{\Delta}_k$ and $\hat{\theta}_k - c_k\boldsymbol{\Delta}_k$ with a non-increasing perturbation step size of $c_k > 0$. Formally, we have

$$\hat{\boldsymbol{g}}_k^{\text{SP}}(\hat{\theta}_k) = \frac{\ell(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k^{(+)}) - \ell(\hat{\theta}_k - c_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k^{(-)})}{2c_k\boldsymbol{\Delta}_k}, \tag{2.5}$$

where $\boldsymbol{\Delta}_k^{-1}$ or $1/\boldsymbol{\Delta}_k$ denotes $[\Delta_{k1}^{-1}, \ldots, \Delta_{kp}^{-1}]^T$, and $\boldsymbol{v}_k^{(+)}, \boldsymbol{v}_k^{(-)} \in \mathbb{R}^q$ represent the random effects at the time of function measurements. The primary advantage of (2.5) over FDSA lies in the dimension-free query requirement per iteration. It is shown that the bias of the two-sided SP gradient approximation is similar

to the two-sided FD gradient approximation, i.e., $\mathbb{E}[\hat{\boldsymbol{g}}_k^{\mathrm{SP}}(\hat{\theta}_k)|\hat{\theta}_k] - \boldsymbol{g}(\hat{\theta}_k) = O(c_k^2)$, where the expectation is taken over both the randomness in $\ell(\hat{\theta}_k, \cdot)$ and $\boldsymbol{\Delta}_k$. Since under a fixed total number of function measurements or a fixed total running time, fewer function measurements leads to more algorithm iterations and hence typically a better final estimate, the advantages of SPSA are apparent. Therefore, SPSA is preferred when $p$ is large or when $\ell(\cdot, \cdot)$ is expensive to evaluate.

To improve the computational cost even further, a variant of SPSA is introduced in Spall (1997), which uses only one function measurement at each iteration and has the form

$$\hat{\boldsymbol{g}}_k^{\mathrm{SP1}}(\hat{\theta}_k) = \frac{\ell(\hat{\theta}_k + c_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k^{(+)})}{c_k\boldsymbol{\Delta}_k}.$$

Note that, unlike the one-sided FD gradient approximation, there is no function measurement of $\ell(\hat{\theta}_k, \boldsymbol{v}_k^{(0)})$ and the bias of the gradient approximation is at the same level as the two-sided SP gradient approximation, i.e., $\mathbb{E}[\hat{\boldsymbol{g}}_k^{\mathrm{SP1}}(\hat{\theta}_k)|\hat{\theta}_k] - \boldsymbol{g}(\hat{\theta}_k) = O(c_k^2)$. Although the one-measurement SP gradient approximation uses the fewest number of function measurements, it is shown to be generally less efficient in total number of function measurements required than the two-sided SP gradient approximation when used to iteratively minimizing the objective function Spall (1997). The method is more useful in real time applications

when it is necessary to get an instantaneous gradient approximation.

To use (2.5) in stochastic optimization algorithms, one can adopt the standard gradient descent scheme to iteratively update the parameter estimate as

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_k^{\mathrm{SP}}(\hat{\theta}_k), \tag{2.6}$$

where $a_k$ is a non-increasing positive gain step size. Spall (1992, Proposition 1) shows that the sequence of the estimate $\{\hat{\theta}_k\}_{k\geq 0}$ generated by SPSA converges almost surely to the optimal point for general stochastic optimization problems where only the noisy function measurements are available. Under some mild conditions, there is also an asymptotic normality for normalized $\hat{\theta}_k$, where the rate of convergence is the same as the FDSA, even though $p$ times fewer function measurements are used.

# Chapter 3

# Mixed Simultaneous Perturbation Stochastic Approximation Algorithm

## 3.1 Introduction

In real-world stochastic optimization, recursive algorithms are often applied to iteratively seek the solutions where the underlying variables are continuous in most scenarios. However, it is also common to have mixed variables (mixture of discrete and continuous variables), especially in engineering design and management science problems. This chapter presents a method for mixed-variable stochastic optimization requiring only noisy values of the loss

function. The method is based on an extension of simultaneous perturbation stochastic approximation (SPSA) and, consequently, is efficient in multidimensional problems.

Some relevant applications of mixed variable stochastic optimization include electric power capacity expansion problem (Shiina and Birge, 2003), optimal decision in energy operations (Wallace and Fleten, 2003), and long-term planning of wind farms (Xiao et al., 2011). In addition, for problems of water resource management (Hemker et al., 2008; Onwunalu and Durlofsky, 2010; Wang et al., 2010), noisy function measurements are generated by the reservoir simulators given a well placement, where the integer variables are the number and type of wells and the continuous variables are the well locations, pumping rates, and other system parameters.

Formally, consider a real-valued mixed-variable loss function $L(\theta) : \mathbb{Z}^d \times \mathbb{R}^{p-d} \to \mathbb{R}$, where $\mathbb{Z}$ and $\mathbb{R}$ represent the spaces of all integers and real numbers, respectively. The vector of interest $\theta$ is $p$-dimensional with the first $d$ components being integer-valued and the last $p - d$ components being real-valued (continuous). Furthermore, assume that $L(\cdot)$ is differentiable with respect to its continuous variables. We are interested in the optimization problem of $\min_{\theta \in \Theta} L(\theta)$ for $\Theta \subseteq \mathbb{Z}^d \times \mathbb{R}^{p-d}$, where the solution set is denoted as

$$\Theta^* = \arg\min_{\theta \in \Theta} L(\theta) = \{\theta^* \in \Theta : L(\theta^*) \leq L(\theta) \text{ for any } \theta \in \Theta\}.$$

Assume that only noisy function measurements $\ell(\boldsymbol{\theta}, \boldsymbol{v}) = L(\boldsymbol{\theta}) + \varepsilon(\boldsymbol{\theta}, \boldsymbol{v})$ are available, where $\varepsilon(\boldsymbol{\theta}, \boldsymbol{v})$ denotes the measurement noise. The explicit form of $L(\boldsymbol{\theta})$ is generally unavailable so that the optimization process is based only on the noisy function measurements $\ell(\boldsymbol{\theta}, \boldsymbol{v})$, which is possibly obtained via Monte Carlo simulations.

Solving stochastic optimization with mixed-variable variables is typically more challenging than with only continuous or integer variables due to the mixture structure of the feasible region and the noise in the function measurements. One possible solution is to naively apply some classical general optimization algorithms that are designed for noise-free problems. Those algorithms include localized random search (Matyas, 1965), adaptive random search (Zabinsky, 2013) and mixed-variables evolutionary programming technique (Cao et al., 2000). Another possible solution is to extend the discrete optimization methods with noisy function measurements to mixed variables by proposing new candidate points containing both the discrete and continuous variables.

A systematic literature review on discrete optimization is available in Wang (2013) and Hill (2014), where general random search algorithms such as stochastic ruler (Yan and Mukai, 1992), stochastic comparison (Gong et al., 2000) and simulated annealing with noise (Gutjahr and Pflug, 1996) are discussed. These iterative methods propose a candidate point based on the current estimate and

then use some metric to determine if the proposed point should be accepted or rejected. In practical applications where function measurements are noisy and/or the search space is high-dimensional, these methods slow down dramatically and become less accurate since it becomes harder to accept a newly proposed point. More importantly, the implicit function structure information, such as the derivatives with respect to the continuous variables are not used. To increase convergence rate with respect to iterations when the function is smooth, a family of retrospective optimization algorithms for mixed-variable variables is proposed in Wang (2012). The term retrospective optimization means a general framework based on the sample-path approach (Chen and Schmeiser, 2001; Pasupathy and Schmeiser, 2009; Pasupathy, 2010; Sandıkçı et al., 2013), which is also called sample average approximation (Kleywegt et al., 2002). Algorithms introduced in Wang (2012) are essentially continuous search procedures embedded in a retrospective optimization framework using dynamic simplex interpolation. During the training process, however, the total number of function measurements is increasing exponentially with respect to the iteration number. Moreover, separate function measurements are required for estimating each component of the gradient of the linear interpolation.

A different approach is related to the stochastic mixed-variable programming, where a good survey can be found in Sen (2010). Although the goal is still to minimize an expected objective function with certain constraints,

many stochastic mixed-variable programming algorithms (see, e.g., Takriti and Birge, 2000; Guan et al., 2009; Ntaimo, 2010) require the noisy measurement to have a finite number of realizations (scenarios) so that the stochastic problem can be converted to a deterministic problem. Hence, any expected values can be written as a finite sum, which makes the problem deterministic. When the noisy measurements are continuous random variables with unknown distributions, these methods become either less attractive or non-applicable.

Therefore, to efficiently handle the general mixed-variable problems with noisy loss measurements or high-dimensional $\theta$, we present a generic stochastic optimization algorithm called the mixed simultaneous perturbation stochastic approximation (MSPSA) that can efficiently minimize a mixed variable problem with only noisy loss function measurements. Our idea is based on the SPSA algorithm (see. e.g., Spall, 1992, 2005; Bhatnagar et al., 2013), which is particularly efficient in solving continuous stochastic optimization problems. The benefits of MSPSA come in five ways: i) implementation is relatively easy; ii) only two noisy loss function measurements are required at each iteration, regardless of problem dimension $p$; iii) loss function structure information, such as gradient or subgradient, is used implicitly; iv) noisy loss measurements are handled properly; v) high-dimensional problems can be solved efficiently in a rigorously justified way. The idea of simultaneous perturbation is also considered in Wang (2013) and Wang and Spall (2014), where the discrete simul-

taneous perturbation stochastic approximation (DSPSA) is proposed to solve problems with only discrete variables.

MSPSA formally connects SPSA and DSPSA such that both SPSA and DSPSA can be viewed as special cases of MSPSA by varying the number of discrete or continuous variables in the proposed algorithm. That said, there are non-trivial aspects in both the algorithm design and the associated theory in order to elegantly merge SPSA and DSPSA to form MSPSA. Based on the preliminary convergence result in Wang et al. (2018b), we derive the formal almost sure convergence and rate of convergence of MSPSA under conditions similar to those of SPSA and DSPSA. The theoretical results not only illustrate an intuitive balance between the discrete and continuous variables, but also achieve the asymptotic bounds at the same rates as SPSA and DSPSA under these respective special cases.

The remainder of the chapter is organized as follows. Section 3.2 provides the description of the algorithm. The almost sure convergence and rate of convergence are derived in Section 3.3 and Section 3.4, respectively. Section 3.5 compares the rate of convergence of MSPSA with DSPSA and SPSA in the respective limiting cases of all discrete parameters and all continuous parameters. Section 3.6 presents some numerical results and the conclusions are given in Section 3.7.

## 3.2 Algorithm Description

Before presenting the details of the MSPSA algorithm, let us first discuss the motivation of the algorithm using a simple one-dimensional discrete case. The mixed variable case is then proposed by generalizing the discrete case to include continuous variables. Figure 3.1 shows a one-dimensional discrete function with a linear interpolation between every pair of neighboring integer points.

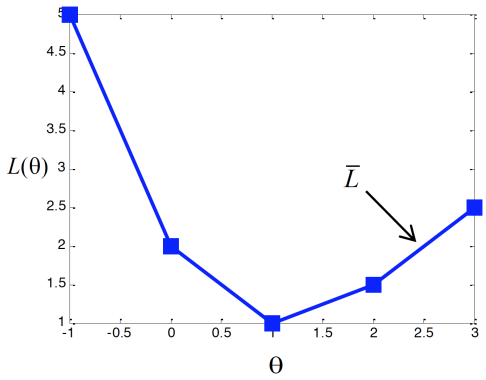



**Figure 3.1:** A Continuous Extension of a One-dimensional Discrete Function

The piecewise linear function $\overline{L}(\theta)$ is a continuous extension of $L(\theta)$ such that $\overline{L}(\theta)$ is continuous for $\theta \in \mathbb{R}$ but only differentiable for $\theta \notin \mathbb{Z}$. Denote $\bar{g}(\theta) = L(m(\theta)+1/2) - L(m(\theta)-1/2)$, where the midpoint $m(\theta) = \lfloor \theta \rfloor + 1/2$ with $\lfloor \cdot \rfloor$ being the flooring function. It is easy to see that $\bar{g}(\theta)$ is exactly the gradient of $\overline{L}(\cdot)$ evaluated at $\theta$ for any $\theta \notin \mathbb{Z}$ and it becomes one of the subgradients of $\overline{L}(\cdot)$ evaluated at $\theta$ for any $\theta \in \mathbb{Z}$. To estimate $\bar{g}(\theta)$ at any $\theta$ with noisy measurements only, we replace $L(\theta)$ with its nosy measurement $\ell(\theta, v)$ and

utilize the idea of SP to construct

$$\hat{g}(\theta) = \frac{\ell(m(\theta) + \Delta/2, v^{(+)}) - \ell(m(\theta) - \Delta/2, v^{(-)})}{\Delta},$$

where $\Delta$ is a random perturbation, $v^{(+)}$ and $v^{(-)}$ denote the random effects. To ensure $m(\theta) \pm \Delta/2 \in \mathbb{Z}$, we assume that $\Delta$ is a symmetrically distributed (around $0$) discrete random variable taking values on all or some of the odd integers (e.g., Bernoulli $\pm 1$ distribution). By this construction, the gradient estimate $\hat{g}(\cdot)$ for the piecewise linear continuous extension $\overline{L}(\cdot)$ is consistent with the gradient estimate in SPSA algorithm. It is shown in He et al. (2003) that the sequence generated by SPSA converges to the optimum for non-differentiable convex continuous functions on a compact and convex domain. Therefore, using the SP idea, there exists a convergent sequence for the continuous extension function $\overline{L}(\cdot)$.

Motivated by the one-dimensional case above, we consider a more general case where $\theta \in \mathbb{Z}^d \times \mathbb{R}^{p-d}$ and the distribution of the random perturbation is relaxed to a general family of distributions. The basic MSPSA algorithm is presented as follows:

- **Step 0 (Initialization):** Set index $k = 0$. Pick an initial guess $\hat{\theta}_0$ and nonnegative coefficients $a, c, A, \alpha$, and $\gamma$ in the gain sequences $a_k = a/(k +$

$1 + A)^{\alpha}$ and $c_k = c/(k+1)^{\gamma}$. Construct the $p$-dimensional vector

$$\boldsymbol{C}_k = [\underbrace{1/2, \ldots, 1/2}_{d \text{ components}}, \quad \underbrace{c_k, \ldots, c_k}_{(p-d) \text{ components}} \quad ]^T.$$

- **Step 1 (Perturbation Vectors):** Generate a $d$-dimensional random perturbation vector $\boldsymbol{\Lambda}_k$ and a $(p-d)$-dimensional random perturbation vector $\boldsymbol{\Pi}_k$ using Monte Carlo algorithms, where all the components of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Pi}_k$ are independently generated from mean-zero symmetric probability distributions satisfying regularity conditions discussed below in Assumption 3.1. An effective (but not mandatory) choice is that all components of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Pi}_k$ are independent symmetric Bernoulli $\pm 1$ distributed with equal probabilities. Construct the $p$-dimensional vector

$$\boldsymbol{\Delta}_k \equiv \begin{bmatrix} \boldsymbol{\Lambda}_k \\ \\ \boldsymbol{\Pi}_k \end{bmatrix}.$$

- **Step 2 (Simultaneous Perturbations):** Compute the simultaneous perturbation estimates $\hat{\theta}_k^{(+)}$ and $\hat{\theta}_k^{(-)}$ around the current estimate $\hat{\theta}_k$,

$$\hat{\theta}_k^{(+)} = \boldsymbol{m}_d(\hat{\theta}_k) + \boldsymbol{C}_k \odot \boldsymbol{\Delta}_k \text{ and } \hat{\theta}_k^{(-)} = \boldsymbol{m}_d(\hat{\theta}_k) - \boldsymbol{C}_k \odot \boldsymbol{\Delta}_k,$$

where $\boldsymbol{m}_d(\cdot)$ is the middle point operator applied to the first $d$ components

of the argument, i.e., $\boldsymbol{m}_d(\theta) = [\lfloor t_1 \rfloor + 1/2, \ldots, \lfloor t_d \rfloor + 1/2, t_{d+1}, \ldots, t_p]^T$ for $\theta = [t_1, \ldots, t_p]$ and $\odot$ is the matrix Hadamard product.

- **Step 3 (Gradient Approximation):** Let $\ell(\hat{\theta}_k, \boldsymbol{v}_k) = L(\hat{\theta}_k) + \varepsilon_k(\hat{\theta}_k, \boldsymbol{v}_k)$ and denote

$$\begin{cases} L_k^{(+)} = L(\hat{\theta}_k^{(+)}) \\ L_k^{(-)} = L(\hat{\theta}_k^{(-)}) \end{cases} \text{ and } \begin{cases} \varepsilon_k^{(+)} = \ell(\hat{\theta}_k^{(+)}, \boldsymbol{v}_k^{(+)}) - L(\hat{\theta}_k^{(+)}) \\ \varepsilon_k^{(-)} = \ell(\hat{\theta}_k^{(-)}, \boldsymbol{v}_k^{(-)}) - L(\hat{\theta}_k^{(-)}) \end{cases}.$$

Generate the simultaneous perturbation gradient approximation,

$$\hat{\boldsymbol{g}}_k(\hat{\theta}_k) = \frac{(L_k^{(+)} + \varepsilon_k^{(+)}) - (L_k^{(-)} + \varepsilon_k^{(-)})}{2\boldsymbol{C}_k \odot \boldsymbol{\Delta}_k}, \tag{3.1}$$

where we use $(\boldsymbol{C}_k \odot \boldsymbol{\Delta}_k)^{-1}$ or $1/(\boldsymbol{C}_k \odot \boldsymbol{\Delta}_k)$ to denote the vector of inverses of the components of $\boldsymbol{C}_k \odot \boldsymbol{\Delta}_k$.

- **Step 4 (Iterative Update):** Use the standard SA form to update $\hat{\theta}_k$,

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_k(\hat{\theta}_k). \tag{3.2}$$

- **Step 5 (Iteration or Termination):** Return to Step 1 with $k+1$ replacing $k$. Terminate the algorithm if there is little change of $\hat{\theta}_k$ in several successive iterates or if the maximum allowable number of iterations has been reached. Return the final estimate by $\text{Proj}_{\Theta}(\hat{\theta}_k)$, where $\text{Proj}_{\Theta}(\cdot)$ is

the projection operator that projects the argument to the domain of the loss function.

**Remark 3.1.** *An important aspect of MSPSA is that the discrete and continuous variables are updated simultaneously in Step 4 in contrast to a cyclic ("see-saw") method, where each subvector (discrete or continuous) is updated in an alternating manner. This approach allows the algorithm to not only take just two loss function measurements per iteration (regardless of the dimension of the problem), but also uncover functional relationships shared between the discrete and continuous variables.*

## 3.3 Convergence

This section discusses the bias in $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ and presents an almost sure convergence proof of the sequence $\{\hat{\theta}_k\}$ generated by the MSPSA algorithm. To separate the discussions on the discrete and continuous components, let $\zeta \in \mathbb{Z}^d$ and $\xi \in \mathbb{R}^{p-d}$ be the first $d$ and the last $p-d$ components of $\theta$, respectively. Similar notations are applied to $\hat{\theta}_k$ and $\hat{\theta}_k^{(\pm)}$ for any $k$ as well, i.e.,

$$\theta = \begin{bmatrix} \zeta \\ \xi \end{bmatrix}, \hat{\theta}_k = \begin{bmatrix} \hat{\zeta}_k \\ \hat{\xi}_k \end{bmatrix} \text{ and } \hat{\theta}_k^{(\pm)} = \begin{bmatrix} \hat{\zeta}_k^{(\pm)} \\ \hat{\xi}_k^{(\pm)} \end{bmatrix}.$$

As mentioned above, the algorithm operates on the full vector in the updating step, not $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ in a cyclic manner.

Because we assume that $L(\cdot)$ is differentiable with respect to its continuous components $\boldsymbol{\xi}$, by fixing the discrete components $\boldsymbol{\zeta}$, we can define a new continuous loss function $L(\cdot|\boldsymbol{\zeta}) : \mathbb{R}^{p-d} \to \mathbb{R}$ such that $L(\boldsymbol{\xi}|\boldsymbol{\zeta}) = L(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \mathbb{Z}^d \times \mathbb{R}^{p-d}$. Given that $L(\cdot|\boldsymbol{\zeta})$ is continuous, we denote its gradient as $\boldsymbol{g}(\cdot|\boldsymbol{\zeta}) : \mathbb{R}^{p-d} \to \mathbb{R}^{p-d}$ such that $\boldsymbol{g}(\boldsymbol{\xi}|\boldsymbol{\zeta}) = \partial L(\boldsymbol{\xi}|\boldsymbol{\zeta})/\partial \boldsymbol{\xi}$ for any $\boldsymbol{\theta} \in \mathbb{Z}^d \times \mathbb{R}^{p-d}$. Due to the existence of the discrete components, however, there is no formal definition for the gradient of $L(\boldsymbol{\theta})$ when $\boldsymbol{\theta} \in \mathbb{Z}^d \times \mathbb{R}^{p-d}$. Hence, we introduce the mean gradient-like expression $\bar{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_k)$ as

$$\bar{g}_i(\hat{\boldsymbol{\theta}}_k) = \begin{cases} \mathbb{E}^{\mathscr{F}_k}\left[\frac{L(\hat{\boldsymbol{\xi}}_k|\hat{\boldsymbol{\zeta}}_k^{(+)}) - L(\hat{\boldsymbol{\xi}}_k|\hat{\boldsymbol{\zeta}}_k^{(-)})}{\Delta_{ki}}\right] & \text{if } i = 1, \ldots, d, \\ \mathbb{E}^{\mathscr{F}_k}\left[\frac{\boldsymbol{g}(\hat{\boldsymbol{\xi}}_k|\hat{\boldsymbol{\zeta}}_k^{(+)}) + \boldsymbol{g}(\hat{\boldsymbol{\xi}}_k|\hat{\boldsymbol{\zeta}}_k^{(-)})}{2}\right] & \text{if } i = d+1, \ldots, p, \end{cases} \tag{3.3}$$

where $\mathbb{E}^{\mathscr{F}_k}[\cdot] = \mathbb{E}[\cdot|\mathscr{F}_k]$ with $\mathscr{F}_k = \{\hat{\boldsymbol{\theta}}_0, \ldots, \hat{\boldsymbol{\theta}}_k\}$ for all $k$, and the expectations are taken over $\boldsymbol{\Delta}_k$. Note that the gradients $\boldsymbol{g}(\hat{\boldsymbol{\xi}}_k|\hat{\boldsymbol{\zeta}}_k^{(+)})$ and $\boldsymbol{g}(\hat{\boldsymbol{\xi}}_k|\hat{\boldsymbol{\zeta}}_k^{(-)})$ in (3.3) are well-defined since $\hat{\boldsymbol{\zeta}}_k^{(+)}, \hat{\boldsymbol{\zeta}}_k^{(-)} \in \mathbb{Z}^d$ and $\boldsymbol{\xi} \in \mathbb{R}^{p-d}$.

The mean gradient-like expression $\bar{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_k)$ in (3.3) is a generalization of the standard gradient for continuous loss function. When all the variables are discrete ($d = p$), $\bar{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_k)$ is reduced to $\mathbb{E}^{\mathscr{F}_k}[(L_k^{(+)} - L_k^{(-)})/\boldsymbol{\Delta}_k]$, which is identical to the mean gradient-like expression defined in Wang and Spall (2013). When all

the variables are continuous ($d = 0$), $\bar{\boldsymbol{g}}(\hat{\theta}_k)$ is reduced to the standard $\boldsymbol{g}(\hat{\theta}_k)$ for continuous problems.

### 3.3.1 Bias of Gradient Estimate

This subsection examines the bias in $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ relative to estimating $\bar{\boldsymbol{g}}(\hat{\theta}_k)$. With the following assumptions, we provide an explicit bound for the bias in $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ and show that $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ is asymptotically unbiased as $k \to \infty$ in Theorem 3.1 below.

**Assumption 3.1** (Discrete Perturbation Vector)**.** *For all $k$ and $i$, $\Lambda_{ki}$ can only take odd integer values. Furthermore, $\Lambda_{ki}$ is independent and identically distributed, symmetrically distributed about $0$, and there exists some constant $B_\Lambda$ such that $1 \leq |\Lambda_{ki}| \leq B_\Lambda$.*

**Assumption 3.2** (Continuous Perturbation Vector)**.** *For all $k$ and $i$, $\Pi_{ki}$ is independent and identically distributed, symmetrically distributed about $0$, and there exists some constants $\kappa_0$ and $\kappa_1$ such that $\mathbb{E}(|\Pi_{ki}|) \leq \kappa_0$ and $\mathbb{E}(|\Pi_{ki}|^{-1}) \leq \kappa_1$.*

**Assumption 3.3** (Measurement Noise)**.** *For all $k$, $\mathbb{E}^{\mathscr{F}_k, \boldsymbol{\Delta}_k}[\varepsilon_k^{(+)} - \varepsilon_k^{(-)}] = 0$.*

**Assumption 3.4** (Loss Function)**.** *Assume that $L(\cdot)$ is defined on $\boldsymbol{\Theta} = \mathbb{Z}^d \times \mathbb{R}^{p-d}$ with a unique minimal point $\theta^*$ and is thrice-differentiable with respect to its continuous components. For almost all $\hat{\theta}_k$, there exists an open neighborhood of $\hat{\xi}_k$ such that for any $\xi$ in that neighborhood, every individual element of $\boldsymbol{H}(\xi|\hat{\zeta}_k) = \partial^2 L(\xi|\hat{\zeta}_k)/\partial\xi\partial\xi^T$ satisfies $|H_{ij}(\xi|\hat{\zeta}_k)| \leq B_H$ for some constant*

*$B_H$ and every individual element of $L^{(3)}(\boldsymbol{\xi}|\hat{\boldsymbol{\zeta}}_k) = \partial^3 L(\boldsymbol{\xi}|\hat{\boldsymbol{\zeta}}_k)/\partial\boldsymbol{\xi}^T\partial\boldsymbol{\xi}^T\partial\boldsymbol{\xi}^T$ satisfies $|L^{(3)}_{ijl}(\boldsymbol{\xi}|\hat{\boldsymbol{\zeta}}_k)| \leq B_T$ for some constant $B_T$.*

**Remark 3.2.** *Assumption 3.1 is to ensure $\hat{\boldsymbol{\zeta}}_k^{(\pm)} \in \mathbb{Z}^d$ for all $k$. Assumptions 3.2 and 3.3 are identical to the random perturbation and noise assumptions in Assumption 2 of Spall (1992). In order to incorporate both the discrete and continuous variables in the loss function, Assumption 3.4 provides a natural extension of the assumptions in Spall (1992, Lemma 1) and Wang and Spall (2011, Theorem 1).*

**Theorem 3.1** (Bias of Gradient Estimate)**.** *Under Assumptions 3.1–3.4, we have*

$$\boldsymbol{b}_k(\hat{\boldsymbol{\theta}}_k) \equiv \mathbb{E}^{\mathscr{F}_k}[\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) - \bar{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_k)] = \boldsymbol{O}(c_k^2) \text{ a.s.} \tag{3.4}$$

*with the component-wise bounds $|b_{ki}(\hat{\boldsymbol{\theta}}_k)| \leq c_k^2 U_i$ such that*

$$U_i = \begin{cases} B_H(p-d)^2\kappa_0^2, & \text{if } i = 1,\ldots,d, \\ \frac{1}{6}B_T\left\{[(p-d)^3 - (p-d-1)^3]\kappa_0^2 + (p-d-1)^3\kappa_0^3\kappa_1\right\}, & \text{if } i = d+1,\ldots,p. \end{cases}$$

**Remark 3.3.** *Theorem 3.1 shows that $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ is asymptotically unbiased as $k \to \infty$ since $c_k \to 0$, as given in Assumption 3.7 below. When all the variables are discrete $(d = p)$, $\hat{\boldsymbol{\theta}}_k^{(\pm)}$ contains no $c_k$ and the gradient estimate $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ becomes unbiased for all $k$. When all the variables are continuous $(d = 0)$, the explicit bound for the bias in $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ becomes identical to Spall (1992, Lemma 1).*

*Proof.* Due to the different assumptions for the discrete and continuous variables in $\hat{\theta}_k$, we proceed below by discussing the first $d$ and last $p-d$ components of $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ separately. All equalities and inequalities hold a.s.

**The First $d$ Components:** We show below that $|b_{ki}(\hat{\theta}_k)| \leq c_k^2 U_i$ for $i = 1, \ldots, d$. Under Assumption 3.4, we may expand $L(\hat{\xi}_k^{(+)} \mid \hat{\zeta}_k^{(+)})$ and $L(\hat{\xi}_k^{(-)} \mid \hat{\zeta}_k^{(-)})$ around $\hat{\xi}_k$ to give

$$L(\hat{\xi}_k^{(+)} \mid \hat{\zeta}_k^{(+)}) = L(\hat{\xi}_k \mid \hat{\zeta}_k^{(+)}) + c_k \boldsymbol{g}(\hat{\xi}_k \mid \hat{\zeta}_k^{(+)})^T \boldsymbol{\Pi}_k + \frac{1}{2} c_k^2 \boldsymbol{\Pi}_k^T \boldsymbol{H}(\bar{\xi}_k^{(+)} \mid \hat{\zeta}_k^{(+)}) \boldsymbol{\Pi}_k, \quad (3.5)$$

$$L(\hat{\xi}_k^{(-)} \mid \hat{\zeta}_k^{(-)}) = L(\hat{\xi}_k \mid \hat{\zeta}_k^{(-)}) - c_k \boldsymbol{g}(\hat{\xi}_k \mid \hat{\zeta}_k^{(\pm)})^T \boldsymbol{\Pi}_k + \frac{1}{2} c_k^2 \boldsymbol{\Pi}_k^T \boldsymbol{H}(\bar{\xi}_k^{(-)} \mid \hat{\zeta}_k^{(-)}) \boldsymbol{\Pi}_k, \quad (3.6)$$

where $\bar{\xi}_k^{(\pm)}$ denotes the points on the line segments between $\hat{\xi}_k$ and $\hat{\xi}_k^{(\pm)}$. Plugging (3.5) and (3.6) into the definition of $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ in (3.1), we have for $i = 1, \ldots, d$,

$$\begin{aligned} \mathbb{E}^{\mathscr{F}_k}[\hat{g}_{ki}(\hat{\theta}_k)] &= \mathbb{E}^{\mathscr{F}_k}\left[\frac{L(\hat{\xi}_k^{(+)} \mid \hat{\zeta}_k^{(+)}) - L(\hat{\xi}_k^{(-)} \mid \hat{\zeta}_k^{(-)})}{\Lambda_{ki}}\right] + \mathbb{E}^{\mathscr{F}_k}\left[\frac{\varepsilon_k^{(+)} - \varepsilon_k^{(-)}}{\Lambda_{ki}}\right] \\ &= \bar{g}_i(\hat{\theta}_k) + c_k \mathbb{E}^{\mathscr{F}_k}\left[\frac{[\boldsymbol{g}(\hat{\xi}_k \mid \hat{\zeta}_k^{(+)}) + \boldsymbol{g}(\hat{\xi}_k \mid \hat{\zeta}_k^{(-)})]^T \boldsymbol{\Pi}_k}{\Lambda_{ki}}\right] \\ &\quad + \frac{1}{2} c_k^2 \mathbb{E}^{\mathscr{F}_k}\left[\frac{\boldsymbol{\Pi}_k^T [\boldsymbol{H}(\bar{\xi}_k^{(+)} \mid \hat{\zeta}_k^{(+)}) - \boldsymbol{H}(\bar{\xi}_k^{(-)} \mid \hat{\zeta}_k^{(-)})]^T \boldsymbol{\Pi}_k}{\Lambda_{ki}}\right], \end{aligned} \quad (3.7)$$

where the last term of the first equality equals $0$ since $\mathbb{E}^{\mathscr{F}_k, \boldsymbol{\Delta}_k}[\varepsilon_k^{(+)} - \varepsilon_k^{(-)}] = 0$ by Assumption 3.3 and

$$\mathbb{E}^{\mathscr{F}_k}\left[\frac{\varepsilon_k^{(+)} - \varepsilon_k^{(-)}}{\Lambda_{ki}}\right] = \mathbb{E}^{\mathscr{F}_k}\left[\frac{\mathbb{E}^{\mathscr{F}_k, \boldsymbol{\Delta}_k}[\varepsilon_k^{(+)} - \varepsilon_k^{(-)}]}{\Lambda_{ki}}\right] = 0. \quad (3.8)$$

We then proceed to discuss the last two terms in (3.7). The second term in (3.7) equals $0$ since both $\Lambda_{ki}$ and $\hat{\boldsymbol{\zeta}}_k^{(\pm)}$ are independent of $\boldsymbol{\Pi}_k$, and $\mathbb{E}[\boldsymbol{\Pi}_k] = \mathbf{0}$ by Assumption 3.1. The last term in (3.7) is bounded in magnitude as

$$\begin{aligned}
&\frac{1}{2}c_k^2\mathbb{E}^{\mathscr{F}_k}\left[\left|\frac{\sum_{m-1}^{p-d}\sum_{n=1}^{p-d}\Pi_{km}[H_{mn}(\bar{\boldsymbol{\xi}}_k^{(+)} \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) - H_{mn}(\bar{\boldsymbol{\xi}}_k^{(-)} \mid \hat{\boldsymbol{\zeta}}_k^{(-)})]\Pi_{kn}}{\Lambda_{ki}}\right|\right] \\
&\leq \frac{1}{2}c_k^2 B_H \mathbb{E}^{\mathscr{F}_k}\left[\sum_{m=1}^{p-d}\sum_{n=1}^{p-d}|\Pi_{km}\Pi_{kn}|\right]\mathbb{E}^{\mathscr{F}_k}\left[\frac{1}{|\Lambda_{ki}|}\right] \\
&\leq \frac{1}{2}c_k^2 B_H (p-d)^2\kappa_0^2 = O(c_k^2),
\end{aligned}$$

where the first inequality holds since $|H_{mn}(\bar{\boldsymbol{\xi}}_k^{(+)} \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) - H_{mn}(\bar{\boldsymbol{\xi}}_k^{(-)} \mid \hat{\boldsymbol{\zeta}}_k^{(-)})| \leq 2B_H$ by Assumption 3.4 and $\boldsymbol{\Pi}_k$ is independent of the $\Lambda_{ki}$.

**The Last $p-d$ Components:** We show below that $|b_{ki}(\hat{\boldsymbol{\theta}}_k)| \leq c_k^2 U_i$ for $i = d+1, \ldots, p$. Similar to (3.5) and (3.6), expanding $L(\hat{\boldsymbol{\xi}}_k^{(+)} \mid \hat{\boldsymbol{\zeta}}_k^{(+)})$ and $L(\hat{\boldsymbol{\xi}}_k^{(-)} \mid \hat{\boldsymbol{\zeta}}_k^{(-)})$ around $\hat{\boldsymbol{\xi}}_k$ to the third order gives

$$\begin{aligned}
L(\hat{\boldsymbol{\xi}}_k^{(+)} \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) &= L(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) + c_k \boldsymbol{g}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(+)})^T\boldsymbol{\Pi}_k + \frac{1}{2}c_k^2\boldsymbol{\Pi}_k^T\boldsymbol{H}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(+)})\boldsymbol{\Pi}_k \\
&\quad + \frac{1}{6}c_k^3 L^{(3)}(\bar{\boldsymbol{\xi}}_k^{(+)} \mid \hat{\boldsymbol{\zeta}}_k^{(+)})[\boldsymbol{\Pi}_k \otimes \boldsymbol{\Pi}_k \otimes \boldsymbol{\Pi}_k], \qquad (3.9) \\
L(\hat{\boldsymbol{\xi}}_k^{(-)} \mid \hat{\boldsymbol{\zeta}}_k^{(-)}) &= L(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(-)}) - c_k \boldsymbol{g}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(-)})^T\boldsymbol{\Pi}_k + \frac{1}{2}c_k^2\boldsymbol{\Pi}_k^T\boldsymbol{H}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(-)})\boldsymbol{\Pi}_k \\
&\quad - \frac{1}{6}c_k^3 L^{(3)}(\bar{\boldsymbol{\xi}}_k^{(-)} \mid \hat{\boldsymbol{\zeta}}_k^{(-)})[\boldsymbol{\Pi}_k \otimes \boldsymbol{\Pi}_k \otimes \boldsymbol{\Pi}_k], \qquad (3.10)
\end{aligned}$$

where $\otimes$ denotes the Kronecker product and $\bar{\boldsymbol{\xi}}_k^{(\pm)}$ are the points on the line segments between $\hat{\boldsymbol{\xi}}_k$ and $\hat{\boldsymbol{\xi}}_k^{(\pm)}$. Plugging (3.9) and (3.10) into the definition of

$\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$ in (3.1), we have for $i = 1, \ldots, p-d$,

$$
\begin{aligned}
&\mathbb{E}^{\mathscr{F}_k}[\hat{g}_{k,d+i}(\hat{\boldsymbol{\theta}}_k)] \\
&= \mathbb{E}^{\mathscr{F}_k}\left[\frac{L(\hat{\boldsymbol{\xi}}_k^{(+)} \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) - L(\hat{\boldsymbol{\xi}}_k^{(-)} \mid \hat{\boldsymbol{\zeta}}_k^{(-)})}{2c_k\Pi_{ki}}\right] + \mathbb{E}^{\mathscr{F}_k}\left[\frac{\varepsilon_k^{(+)} - \varepsilon_k^{(-)}}{2c_k\Pi_{ki}}\right] \\
&= \frac{1}{2c_k}\mathbb{E}^{\mathscr{F}_k}\left[\frac{L(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) - L(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(-)})}{\Pi_{ki}}\right] + \frac{1}{2}\mathbb{E}\left[\frac{[\boldsymbol{g}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) + \boldsymbol{g}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(-)})]^T\boldsymbol{\Pi}_k}{\Pi_{ki}}\right] \\
&\quad + \frac{1}{4}c_k\mathbb{E}^{\mathscr{F}_k}\left[\frac{\boldsymbol{\Pi}_k^T[\boldsymbol{H}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) - \boldsymbol{H}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(-)})]\boldsymbol{\Pi}_k}{\Pi_{ki}}\right] \\
&\quad + \frac{1}{12}c_k^2\mathbb{E}^{\mathscr{F}_k}\left[\frac{[L^{(3)}(\bar{\boldsymbol{\xi}}_k^{(+)} \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) + L^{(3)}(\bar{\boldsymbol{\xi}}_k^{(-)} \mid \hat{\boldsymbol{\zeta}}_k^{(-)})][\boldsymbol{\Pi}_k \otimes \boldsymbol{\Pi}_k \otimes \boldsymbol{\Pi}_k]}{\Pi_{ki}}\right],
\end{aligned} \tag{3.11}
$$

where the last term of the first equality equals $0$, similar to (3.8) . We then proceed to discuss the other terms in the last equality of (3.11). The first term in (3.11) equals $0$ since $\Pi_{ki}$ is independent of $\hat{\boldsymbol{\zeta}}_k^{(\pm)}$ and $\mathbb{E}[\Pi_{ki}^{-1}] = 0$ by Assumption 3.2. Similarly, the second term in (3.11) equals

$$
\begin{aligned}
&\frac{1}{2}\sum_{j=1}^{p-d}\mathbb{E}[g_{d+j}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) + g_{d+j}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(-)})]\mathbb{E}\left[\frac{\Pi_{kj}}{\Pi_{ki}}\right] \\
&= \frac{1}{2}\mathbb{E}[g_{d+i}(\hat{\boldsymbol{\xi}}_k) \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) + g_{d+i}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(-)})] \\
&= \bar{g}_{d+i}(\hat{\boldsymbol{\theta}}_k),
\end{aligned}
$$

and the third term in (3.11) equals

$$
\frac{1}{4}c_k\sum_{m=1}^{p-d}\sum_{n=1}^{p-d}\mathbb{E}[H_{mn}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) - H_{mn}(\hat{\boldsymbol{\xi}}_k \mid \hat{\boldsymbol{\zeta}}_k^{(-)})]\mathbb{E}\left[\frac{\Pi_{km}\Pi_{kn}}{\Pi_{ki}}\right] = 0,
$$

since $\mathbb{E}[\Pi_{kj}/\Pi_{ki}] = 0$ whenever $i \neq j$ by Assumption 3.2. Finally, the last term in (3.11) is bounded in magnitude by

$$\begin{aligned}
&\frac{1}{12}c_k^2\mathbb{E}^{\mathscr{F}_k}\left[\left|\frac{[L^{(3)}(\bar{\boldsymbol{\xi}}_k^{(+)} \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) + L^{(3)}(\bar{\boldsymbol{\xi}}_k^{(-)} \mid \hat{\boldsymbol{\zeta}}_k^{(-)})][\boldsymbol{\Pi}_k \otimes \boldsymbol{\Pi}_k \otimes \boldsymbol{\Pi}_k]}{\Pi_{ki}}\right|\right] \\
&\leq \frac{1}{6}c_k^2 B_T \sum_{l=1}^{p-d}\sum_{m=1}^{p-d}\sum_{n=1}^{p-d}\mathbb{E}^{\mathscr{F}_k}\left[\left|\frac{\Pi_{kl}\Pi_{km}\Pi_{km}}{\Pi_{ki}}\right|\right] \\
&\leq \frac{1}{6}c_k^2 B_T\left\{[(p-d)^3 - (p-d-1)^3]\kappa_0^2 + (p-d-1)^3\kappa_0^3\kappa_1\right\} = O(c_k^2),
\end{aligned}$$

where the first inequality holds since $|L^{(3)}_{lmn}(\bar{\boldsymbol{\xi}}_k^{(+)} \mid \hat{\boldsymbol{\zeta}}_k^{(+)}) - L^{(3)}_{lmn}(\bar{\boldsymbol{\xi}}_k^{(-)} \mid \hat{\boldsymbol{\zeta}}_k^{(-)})| \leq 2B_T$ for $l, m, n = 1, \ldots, p-d$ by Assumption 3.4. □

### 3.3.2 Almost Sure Convergence

In this subsection, we establish the almost sure convergence of $\hat{\theta}_k$ to the optimum $\theta^*$. For the convenience of the main convergence theorem below, we first introduce some useful notation. For all $k$, let $\mathscr{G}_k = \{\boldsymbol{\Delta}_0, \ldots, \boldsymbol{\Delta}_k\}$, $\boldsymbol{\Delta}_k^{-T} = (\boldsymbol{\Delta}_k^{-1})^T$, and $\Omega$ be the set of all possible outcomes of $\hat{\theta}_k$. Define $\mathbb{M}$ as the set of all middle points such that $\mathbb{M} = \{\ldots, -5/2, -3/2, -1/2, 1/2, 3/2, 5/2, \ldots\}$. Recall that $\theta = [t_1, \cdots, t_p]$ and for any $\theta' = [t_1', \cdots, t_p'] \in \mathbb{R}^p$, denote $\boldsymbol{M}(\theta') = \{\theta \in \mathbb{M}^d \times \mathbb{R}^{p-d}$ where $|t_i - t_i'| \leq 1/2$ for $i = 1, \ldots, p\}$. When all the variables are discrete ($d = p$), $\boldsymbol{M}(\theta')$ is the set of middle points of all unit hypercubes containing $\theta'$ such that $\boldsymbol{M}(\theta')$ has exactly one point if $\theta'$ lies strictly within a unit hypercube or at most $2^p$ points if $\theta'$ lies on the boundary of multiple unit hypercubes (Wang

and Spall, 2011). When all the variables are continuous ($d = 0$), $\boldsymbol{M}(\theta')$ is simply the unit hypercube centered at $\theta'$. Below are some conditions used in the convergence result.

**Assumption 3.5** (Continuous Perturbation Vector)**.** *For all $k$ and $i$, assume that $\mathbb{E}[\Pi_{ki}^{-2}] \leq \kappa_2$ for some constant $\kappa_2$.*

**Assumption 3.6** (Bounded Differences)**.** *For all $k$, assume that $\mathbb{E}^{\boldsymbol{\Delta}_k}[(L_k^{(+)} - L_k^{(-)})^2] \leq \sigma_L^2$ and $\mathbb{E}^{\boldsymbol{\Delta}_k}[(\varepsilon_k^{(+)} - \varepsilon_k^{(-)})^2] \leq \sigma_\varepsilon^2$ for some constants $\sigma_L$ and $\sigma_\varepsilon$.*

**Assumption 3.7** (Step-size Sequences)**.** *$\{a_k\}$ and $\{c_k\}$ are sequences that $a_k, c_k > 0$ for all $k$; $a_k \to 0, c_k \to 0$ as $k \to \infty$; further,*

$$\sum_{k=0}^{\infty} a_k = \infty; \text{ and } \sum_{k=0}^{\infty} \frac{a_k^2}{c_k^2} < \infty, \text{ and, if } 1 \leq d \leq p-1, \sum_{k=0}^{\infty} a_k c_k^2 < \infty.$$

**Assumption 3.8** (Iterate Boundedness)**.** $\sup_k \|\hat{\theta}_k\| < \infty$ *a.s.*

**Assumption 3.9** (Search Direction)**.** *For any $\theta' \in \mathbb{R}^p \backslash \{\theta^*\}$, $\bar{\boldsymbol{g}}(\theta)^T(\theta' - \theta^*) > 0$ when $\theta \in \boldsymbol{M}(\theta')$.*

**Remark 3.4.** *Assumptions 3.5–3.8 are standard conditions similar to those in SPSA (see, e.g., Spall, 1992, Proposition 2 and Spall, 2005, Section 7.3). Although the condition $\sum_{k=0}^{\infty} a_k c_k^2 < \infty$ in Assumption 3.7 is not needed when all the variable are continuous ($d = 0$), or when all the variables are discrete (at $d = p$, there is no $c_k$), it is needed here to bound the bias term $\boldsymbol{b}_k(\hat{\theta}_k)$ for mixed*

*variables. This condition also appears in various other stochastic approximation algorithms (see, e.g., Pflug, 2012, Theorem 5.3 and Kushner and Yin, 2003, Theorem 5.2.1). Assumption 3.9 is a generalization of the standard search direction condition for continuous problems (Spall, 2005, Section 4.3.2) and discrete problems (Wang and Spall, 2011, Theorem 1). Section 3.3.3 below also provides a discussion on the connection of Assumption 3.9 to the concept of discrete convexity.*

We present the formal almost sure convergence of $\hat{\theta}_k$ for unconstrained problems in Theorem 3.2 below. For constrained problems with a general explicit inequality constraint set $\boldsymbol{G} = \{\theta \in \mathbb{Z}^d \times \mathbb{R}^{p-d} : q_j(\theta) \leq 0, j = 1, \ldots, s\}$, one can consider adding projection steps to replace $\hat{\theta}_k^{(\pm)}$ with $\text{Proj}_{\boldsymbol{G}}(\hat{\theta}_k^{(\pm)})$ or add the corresponding penalty functions to the original objective function. Similar approaches that are designed for continuous problems have been investigated in Sadegh (1997) and Wang and Spall (2008).

**Theorem 3.2** (Almost Sure Convergence)**.** *Under Assumptions 3.1–3.9, we have*

$$\lim_{k\to\infty} \hat{\theta}_k = \theta^* \text{ a.s.}$$

*Proof.* We first rewrite the standard SA updating form (3.2) into a generalized

Robbins-Monro algorithm,

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_k(\hat{\theta}_k) = \hat{\theta}_k - a_k[\bar{\boldsymbol{g}}(\hat{\theta}_k) + \boldsymbol{b}_k(\hat{\theta}_k) + \boldsymbol{e}_k(\hat{\theta}_k)], \tag{3.12}$$

where the error term $\boldsymbol{e}_k(\hat{\theta}_k)$ is defined as

$$\boldsymbol{e}_k(\hat{\theta}_k) = \hat{\boldsymbol{g}}_k(\hat{\theta}_k) - \mathbb{E}[\hat{\boldsymbol{g}}_k(\hat{\theta}_k) \mid \hat{\theta}_k].$$

Using Assumption 3.8, there exists $\Omega_1 \subseteq \Omega$ such that $P(\Omega_1) = 1$ and $\{\hat{\theta}_k(\omega)\}$ is a bounded sequence for any $\omega \in \Omega_1$. By the Bolzano-Weierstrass theorem, $\{\hat{\theta}_k(\omega)\}$ has at least one convergent subsequence denoted by $\{\hat{\theta}_{k_s}(\omega)\}$. Let $\theta'(\omega)$ be the limiting point of this convergent subsequence such that $\lim_{s\to\infty} \hat{\theta}_{k_s}(\omega) = \theta'(\omega)$. Although different convergent subsequences generally have different limit points, it is shown below that all the convergent subsequences have the same limiting point $\theta^*$, which implies that the sequence $\{\hat{\theta}_k(\omega)\}$ must converge to $\theta^*$ almost surely.

Consider an arbitrary convergent subsequence $\{\hat{\theta}_{k_s}(\omega)\}$ with its limiting point $\theta'(\omega)$. For succinctness, we suppress the notation $\omega$ for the rest of the proof. By the recursive relationship in (3.12), it is easy to get

$$\theta' - \hat{\theta}_{k_s} = -\sum_{i=k_s}^{\infty} a_i[\bar{\boldsymbol{g}}(\hat{\theta}_i) + \boldsymbol{b}_i(\hat{\theta}_i) + \boldsymbol{e}_i(\hat{\theta}_i)] \text{ for } s \in \mathbb{N}. \tag{3.13}$$

Note that $\{\sum_{i=k}^{m} a_i \boldsymbol{e}_i(\hat{\theta}_i)\}_{m\geq k}$ is a martingale sequence since $\mathbb{E}[\boldsymbol{e}_k(\hat{\theta}_k)|\hat{\theta}_k] = \mathbb{E}^{\mathscr{F}_k}[\boldsymbol{e}_k(\hat{\theta}_k)] = 0$. Then for any $\eta > 0$, Doob's martingale inequality (Kushner and Clark, 1978, pp. 27) implies

$$\mathbb{P}(\sup_{m\geq k} \| \sum_{i=k}^{m} a_i \boldsymbol{e}_i(\hat{\theta}_i)\| \geq \eta) \leq \frac{1}{\eta^2}\mathbb{E}[\| \sum_{i=k}^{\infty} a_i \boldsymbol{e}_i(\hat{\theta}_i)\|^2] = \frac{1}{\eta^2} \sum_{i=k}^{\infty} a_i^2 \mathbb{E}[\|\boldsymbol{e}_i(\hat{\theta}_i)\|^2], \quad (3.14)$$

where the last equality is because for all $i < j$,

$$\mathbb{E}[\boldsymbol{e}_i(\hat{\theta}_i)^T \boldsymbol{e}_j(\hat{\theta}_j)] = \mathbb{E}[\mathbb{E}^{\mathscr{F}_j, \mathscr{G}_{j-1}}[\boldsymbol{e}_i(\hat{\theta}_i)^T \boldsymbol{e}_j(\hat{\theta}_j)]] = \mathbb{E}[\boldsymbol{e}_i(\hat{\theta}_i)^T \mathbb{E}^{\mathscr{F}_j}[\boldsymbol{e}_j(\hat{\theta}_j)]] = 0.$$

Since $\mathbb{E}(\|\hat{\boldsymbol{g}}_i(\hat{\theta}_i)\|^2) = \mathbb{E}(\|\boldsymbol{e}_i(\hat{\theta}_i)\|^2) + \mathbb{E}(\|\mathbb{E}(\hat{\boldsymbol{g}}_i(\hat{\theta}_i)|\hat{\theta}_k)\|^2)$, we can further bound (3.14) as

$$\frac{1}{\eta^2} \sum_{i=k}^{\infty} a_i^2 \mathbb{E}(\|\boldsymbol{e}_i(\hat{\theta}_i)\|^2) \leq \frac{1}{\eta^2} \sum_{i=k}^{\infty} a_i^2 \mathbb{E}(\|\hat{\boldsymbol{g}}_i(\hat{\theta}_i)\|^2).$$

Using the definition of $\hat{\boldsymbol{g}}_i(\hat{\theta}_i)$ in (3.1), we get

$$\begin{aligned}\mathbb{E}(\|\hat{\boldsymbol{g}}_i(\hat{\theta}_i)\|^2) = \mathbb{E}[((L_i^{(+)} - L_i^{(-)})^2 + (\varepsilon_i^{(+)} - \varepsilon_i^{(-)})^2)(2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-T}(2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-1}] \\ + 2\mathbb{E}[(L_i^{(+)} - L_i^{(-)})(\varepsilon_i^{(+)} - \varepsilon_i^{(-)})(2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-T}(2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-1}], \quad (3.15)\end{aligned}$$

where the second term on the right-hand size equals $0$, similar to (3.8). Moreover,

$$\mathbb{E}[(2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-T}(2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-1}]$$

$$
\begin{aligned}
&= \mathbb{E}[\Lambda_{i1}^{-2} + \cdots + \Lambda_{id}^{-2} + (2c_i\Pi_{i1})^{-2} + \cdots + (2c_i\Pi_{i,p-d})^{-2}] \\
&\leq d + (p-d)\frac{\kappa_2}{4c_i^2}, \qquad (3.16)
\end{aligned}
$$

where the last inequality is by Assumptions 3.1 and 3.5. To bound the first term on the right-hand size of (3.15), we have

$$
\begin{aligned}
&\mathbb{E}[[(L_i^{(+)} - L_i^{(-)})^2 + (\varepsilon_i^{(+)} - \varepsilon_i^{(-)})^2](2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-T}(2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-1}] \\
&= \mathbb{E}[\{\mathbb{E}^{\boldsymbol{\Delta}_i}[(L_i^{(+)} - L_i^{(-)})^2] + \mathbb{E}^{\boldsymbol{\Delta}_i}[(\varepsilon_i^{(+)} - \varepsilon_i^{(-)})^2]\}(2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-T}(2\boldsymbol{C}_i \odot \boldsymbol{\Delta}_i)^{-1}] \\
&\leq (\sigma_L^2 + \sigma_\varepsilon^2)\left[d + (p-d)\frac{\kappa_2}{4c_i^2}\right].
\end{aligned}
$$

Therefore, after dropping the $1/\eta^2$ multiplier in (3.14), we have

$$
\begin{aligned}
\sum_{i=k}^{\infty} a_i^2 \mathbb{E}(\|\boldsymbol{e}_i(\hat{\boldsymbol{\theta}}_i)\|^2) &\leq \sum_{i=k}^{\infty} a_i^2 \mathbb{E}(\|\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\theta}}_i)\|^2) \\
&\leq (\sigma_L^2 + \sigma_\varepsilon^2)\left[d\sum_{i=k}^{\infty} a_i^2 + (p-d)\kappa_2 \sum_{i=k}^{\infty} \frac{a_i^2}{c_i^2}\right] \\
&< \infty,
\end{aligned}
$$

which further implies $\lim_{k\to\infty} \sum_{i=k}^{\infty} a_i^2 \mathbb{E}(\|\boldsymbol{e}_i(\hat{\boldsymbol{\theta}}_i)\|^2) = 0$. Hence, letting $k \to \infty$ on both sides of (3.14) gives

$$
\lim_{k\to\infty} \mathbb{P}(\sup_{m\geq k} \| \sum_{i=k}^{m} a_i \boldsymbol{e}_i(\hat{\boldsymbol{\theta}}_i)\| \geq \eta) = 0. \qquad (3.17)
$$

Using Theorem 4.1.1 in Chung (2001) on almost sure convergence, (3.17) implies that there exists $\Omega_2 \subseteq \Omega$ such that $P(\Omega_2) = 1$ and $\lim_{s\to\infty} \sum_{i=k_s}^{\infty} a_i \boldsymbol{e}_i(\hat{\theta}_i(\omega)) = \mathbf{0}$ for any $\omega \in \Omega_2$. In addition, using the result $\boldsymbol{b}_k(\hat{\theta}_k) = \boldsymbol{O}(c_k^2)$ from Theorem 3.1 and $\sum_{k=0}^{\infty} a_k c_k^2 < \infty$ from Assumption 3.7, it is easy to see that there exists $\Omega_3 \subseteq \Omega$ such that $P(\Omega_3) = 1$ and $\lim_{s\to\infty} \sum_{i=k_s}^{\infty} a_i \boldsymbol{b}_i(\hat{\theta}_i(\omega)) = \mathbf{0}$ for any $\omega \in \Omega_3$.

Therefore, $P(\Omega_1 \cap \Omega_2 \cap \Omega_3) = 1$ and for any $\omega \in \Omega_1 \cap \Omega_2 \cap \Omega_3$, the terms in (3.13) satisfy

$$\lim_{s\to\infty} \sum_{i=k_s}^{\infty} a_i \boldsymbol{b}_i(\hat{\theta}_i(\omega)) = \mathbf{0} \text{ and } \lim_{s\to\infty} \sum_{i=k_s}^{\infty} a_i \boldsymbol{e}_i(\hat{\theta}_i(\omega)) = \mathbf{0}.$$

Together with $\lim_{s\to\infty} \hat{\theta}_{k_s}(\omega) = \theta'(\omega)$, the above implies

$$\lim_{s\to\infty} \sum_{i=k_s}^{\infty} a_i \bar{\boldsymbol{g}}(\hat{\theta}_i(\omega)) = \mathbf{0}. \tag{3.18}$$

For any convergent subsequence $\{\hat{\theta}_{k_s}(\omega)\}$, we now prove by contradiction that the limiting point $\theta'(\omega)$ is the optimal point $\theta^*$. Suppose $\theta'(\omega)$ is not $\theta^*$. Because $\lim_{s\to\infty} \hat{\theta}_{k_s}(\omega) = \theta'(\omega)$, it is known that for any $\delta > 0$, there exists some $S > 0$ such that $\|\hat{\theta}_{k_s}(\omega) - \theta'(\omega)\| < \delta$ whenever $s > S$. Hence, for any $s > S$, $\boldsymbol{m}_d(\hat{\theta}_{k_s}(\omega)) \in \boldsymbol{M}(\theta'(\omega))$ and $\bar{\boldsymbol{g}}(\boldsymbol{m}_d(\hat{\theta}_{k_s}(\omega)))^T(\theta'(\omega) - \theta^*) > 0$ by Assumption 3.9. Since $\bar{\boldsymbol{g}}(\hat{\theta}_{k_s}(\omega)) = \bar{\boldsymbol{g}}(\boldsymbol{m}_d(\hat{\theta}_{k_s}(\omega)))$, we get $\bar{\boldsymbol{g}}(\hat{\theta}_{k_s}(\omega))^T(\theta'(\omega) - \theta^*) > 0$. Combining

with $\lim_{s\to\infty}\sum_{i=k_s}^{\infty} a_i = \infty$ from Assumption 3.7, it gives

$$\lim_{s\to\infty}\sum_{i=k_s}^{\infty} a_i \bar{\boldsymbol{g}}(\hat{\theta}_i(\omega))^T(\theta'(\omega)-\theta^*) = \infty.$$

On the other hand, multiplying both sides of (3.18) by $\theta'(\omega)-\theta^*$ implies

$$\lim_{s\to\infty}\sum_{i=k_s}^{\infty} a_i \bar{\boldsymbol{g}}(\hat{\theta}_i(\omega))^T(\theta'(\omega)-\theta^*) = 0,$$

leading to a contradiction. Therefore, for any $\omega \in \Omega_1\cap\Omega_2\cap\Omega_3$, all the convergent subsequences $\{\hat{\theta}_{k_s}(\omega)\}$ converge to $\theta^*$.

Now suppose $\{\hat{\theta}_k(\omega)\}$ does not converge to $\theta^*$. In other words, there is some $\eta > 0$ such that infinitely many $\hat{\theta}_k(\omega)$ are at a distance at least $\eta$ away from $\theta^*$, i.e., for all $K \in \mathbb{N}$ there is an $k \geq K$ such that $\|\hat{\theta}_k(\omega)-\theta^*\| \geq \eta$. Denote those infinitely many $\hat{\theta}_k(\omega)$ that are away from $\theta^*$ as a subsequence $\{\hat{\theta}_{k_s}(\omega)\}$. Since $\{\hat{\theta}_{k_s}(\omega)\}$ is a bounded subsequence, the Bolzano-Weierstrass theorem implies that $\{\hat{\theta}_{k_s}(\omega)\}$ must contain a convergent sub-subsequence $\{\hat{\theta}_{k_{s_\ell}}(\omega)\}$. However, note that $\{\hat{\theta}_{k_{s_\ell}}(\omega)\}$ does not converge to $\theta^*$ by construction and hence violates the conclusion that all the convergent subsequences converge to $\theta^*$. Therefore, by contradiction, we must have the sequence $\{\hat{\theta}_k(\omega)\}$ converging and having only one convergent point $\theta^*$, i.e., $\lim_{k\to\infty}\hat{\theta}_k(\omega) = \theta^*$ a.s. □

### 3.3.3 Discrete Convexity

To further understand Assumption 3.9, let us consider a special case, where all variables are discrete ($d = p$) and $\{\Delta_{ki}\}$ are independent Bernoulli random variables taking the values $\pm 1$ with equal probabilities (a common choice of $\Delta_{ki}$ that satisfies Assumption 3.1). Under this special case, Assumption 3.9 provides some connections to the concept of discrete convexity, which is introduced in the early 1970s for non-separable functions and provides a sufficient condition for a local optimum to be a global optimum (Miller, 1971). Other definitions of discrete convex functions are discussed in Favati (1990); Fujishige and Murota (2000), and Murota and Shioura (2001), where Miller's discrete convexity is shown to be the most general. When $p = 1$, any function satisfying Miller's discrete convexity also satisfies Assumption 3.9. When $p > 1$, there are functions that satisfy both Miller's discrete convexity and Assumption 3.9, but it is possible to have functions that satisfy only one or the other.

For any discrete function $L(\cdot)$, consider the following three general continuous extensions: i) a separable function $\overline{L}(\theta) = \sum_{i=1}^{p} \overline{L}_i(t_i)$ with $\overline{L}_i(\cdot)$ being the linear interpolation function within the $i$-th dimension; ii) a piecewise linear function that is linear within each unit hypercube; and iii) a quadratic function $\overline{L}(\theta) = \theta^T \boldsymbol{A} \theta + \boldsymbol{b}^T \theta + c$ with $\boldsymbol{A}$, $\boldsymbol{b}$, and $c$ being a matrix, vector, and scalar, respectively. Assumption 3.9 is guaranteed if the following corresponding cases hold: i) $\overline{L}(\cdot)$ is a strictly convex separable function with $\theta^* \in \mathbb{Z}^p$; ii)

$\overline{L}(\cdot)$ is strictly convex and linear in each unit hypercube (Hill et al., 2004); and iii) $\overline{L}(\theta) = \theta^T \boldsymbol{A}\theta + \boldsymbol{b}^T\theta + c$, where $\boldsymbol{A}$ is a symmetric strictly diagonal dominant matrix with positive diagonal values. Further details on the special cases above can be found in Wang and Spall (2011) and Wang (2013).

### 3.3.4 Constrained Problems

Most discussion here pertains to the unconstrained problem. However, MSPSA can also be used on some constrained problems, although the above theorems do not directly apply with constraints. In particular, suppose $l_i \leq t_i \leq u_i$ with $l_i, u_i \in \mathbb{Z}$ for $i = 1, \ldots, d$ and $l_i, u_i \in \mathbb{R}$ for $i = d+1, \ldots, p$. Under this type of constrained case, the sequence $\{\hat{\theta}_k\}_{k\geq 0}$ generated by MSPSA could be outside the domain. Thus we need to modify the general algorithm to handle the bounded domain case. Let $\psi(\theta) = [\psi_1(t_1), \ldots, \psi_p(t_p)]^T$ be the projection mapping $\theta$ back to the set that is bounded by $l_i$ and $u_i$ such that

$$\psi_i(t_i) = \begin{cases} l_i & \text{if } t_i < l_i \\ t_i & \text{if } l_i \leq t_i \leq u_i \\ u_i - \tau & \text{if } t_i > u_i \text{ and } i \leq d \\ u_i & \text{if } t_i > u_i \text{ and } i > d+1 \end{cases},$$

with $0 < \tau < u_i - l_i$ for all $i$ being a very small positive number (e.g., $\tau = 10^{-10}$). The term $\tau$ is introduced here to make sure that for $i = 1, \ldots, d$, the $i$-the component in $\boldsymbol{m}_d(\boldsymbol{\psi}(\hat{\theta}_k))$ is bounded by $u_i$ since $\lfloor u_i \rfloor + 0.5 > u_i$ but $\lfloor u_i - \tau \rfloor + 0.5 < u_i$. Besides $\boldsymbol{m}_d(\boldsymbol{\psi}(\hat{\theta}_k))$, we also need to project $\hat{\theta}_k^{(\pm)}$ into the feasible domain if necessary to ensure proper function measurements. Therefore, $\hat{\theta}_k^{(\pm)}$ is modified to $\hat{\theta}_k^{(\pm)} = \text{Proj}_{\boldsymbol{\Theta}}(\boldsymbol{m}_d(\boldsymbol{\psi}(\hat{\theta}_k)) \pm \boldsymbol{C}_k \odot \boldsymbol{\Delta}_k)$ and Step 2 of the general MSPSA algorithm in Section 3.2 becomes

> *Step 2 (Modified loss function evaluation):* Compute the simultaneous perturbation estimates around the current estimate $\hat{\theta}_k$,
>
> $$\hat{\theta}_k^{(+)} = \text{Proj}_{\boldsymbol{\Theta}}(\boldsymbol{m}_d(\boldsymbol{\psi}(\hat{\theta}_k)) + \boldsymbol{C}_k \odot \boldsymbol{\Delta}_k),$$
> $$\hat{\theta}_k^{(-)} = \text{Proj}_{\boldsymbol{\Theta}}(\boldsymbol{m}_d(\boldsymbol{\psi}(\hat{\theta}_k)) - \boldsymbol{C}_k \odot \boldsymbol{\Delta}_k),$$
>
> and obtain two noisy measurements of the loss function $\ell(\hat{\theta}_k^{(+)}, \boldsymbol{v}_k^{(+)})$ and $\ell(\hat{\theta}_k^{(-)}, \boldsymbol{v}_k^{(-)})$.

Note that $\hat{\theta}_k$ is still allowed to be outside the feasible domain since no function values are collected directly at $\hat{\theta}_k$. We deliberately make $\hat{\theta}_k$ unconstrained by skipping the unnecessary projections to avoid any potential information loss.

#### 3.3.4.1 Binary Problems

Let $d = p$ and $\boldsymbol{\Theta} = \{0,1\}^p$. We now consider the important special case of binary problems. Similar to the mixed-variable problems, the binary problems also have many practical applications, such as the feature selection problems (Aksakalli and Malekipirbazari, 2016), weighted max-cut problems (Ferrez et al., 2005), and a survey on the binarization of meta-heuristics designed for continuous optimization (Crawford et al., 2017), where the search space is transformed into a binary space.

**Theorem 3.3.** *Given $L(\cdot)$ is defined on $\boldsymbol{\Theta} = \{0,1\}^p$ with a unique minimal point $\theta^*$, assume that i) $L(\theta)$ is bounded for any $\theta \in \boldsymbol{\Theta}$; ii) for all $k$ and $i$, $\Delta_{ki}$ is independent and identically distributed following a Bernoulli distribution taking values $\pm 1$ with equal probabilities; iii) for all $k$, $\mathbb{E}^{\mathscr{F}_k, \boldsymbol{\Delta}_k}[\varepsilon_k^{(+)} - \varepsilon_k^{(-)}] = 0$ and there exists some constants $\sigma_\varepsilon^2$ such that $\mathbb{E}^{\boldsymbol{\Delta}_k}[(\varepsilon_k^{(+)} - \varepsilon_k^{(-)})^2] \leq \sigma_\varepsilon^2$; iv) $\{a_k\}$ is a sequence that $a_k > 0$; $a_k \to 0$ as $k \to \infty$; $\sum_{k=0}^{\infty} a_k = \infty$; and $\sum_{k=0}^{\infty} a_k^2 < \infty$; v) for any $\theta' \in [0,1]^p \backslash \{\theta^*\}$, $\bar{\boldsymbol{g}}(\mathbf{1}_p/2)^T(\theta' - \theta^*) > 0$, where $\mathbf{1}_p = [1, \ldots, 1]^T$. Then*

$$\text{Proj}_{\boldsymbol{\Theta}}\left(\lim_{k\to\infty} \hat{\theta}_k\right) = \theta^* \text{ a.s.}$$

*Proof.* Following similar arguments in the proof of Theorem 3.2, we have

$$\hat{\theta}_k = \hat{\theta}_0 - \sum_{i=0}^{k} a_i[\bar{\boldsymbol{g}}(\hat{\theta}_i) + \boldsymbol{b}_k(\hat{\theta}_i) + \boldsymbol{e}_i(\hat{\theta}_i)]. \tag{3.19}$$

Denote $\mathrm{sgn}(\cdot)$ as the sign function applied to each component of the argument vector. It is easy to see that $\mathrm{sgn}(\boldsymbol{\theta}' - \boldsymbol{\theta}^*) = \mathbf{1}_p - 2\boldsymbol{\theta}^* \in \{-1, 1\}^p$ for any $\boldsymbol{\theta}' \in [0,1]^p \backslash \{\boldsymbol{\theta}^*\}$ and $\mathrm{sgn}(\bar{\boldsymbol{g}}(\mathbf{1}_p/2)) = \mathbf{1}_p - 2\boldsymbol{\theta}^*$. Combining with $\lim_{k\to\infty} \sum_{i=0}^{k} a_i = \infty$, we have for all $j$,

$$\lim_{k\to\infty} \sum_{i=0}^{k} a_i [\bar{\boldsymbol{g}}(\boldsymbol{m})]_j = \begin{cases} \infty \text{ a.s.} & \text{if } t_j^* = 1, \\ -\infty \text{ a.s.} & \text{if } t_j^* = 0. \end{cases}$$

For the last two terms on the right-hand side (3.19), it is straightforward to check that both $\{\sum_{i=0}^{k} a_i \boldsymbol{b}_i(\hat{\boldsymbol{\theta}}_i)\}_{k\geq 0}$ and $\{\sum_{i=0}^{k} a_i \boldsymbol{e}_i(\hat{\boldsymbol{\theta}}_i)\}_{k\geq 0}$ are martingales. Hence by martingale convergence theorem and the arguments analogous to (3.14) in Theorem 3.2, both the martingale sequences are almost surely finite in the limit. Therefore, we conclude from (3.19) that for all $j$,

$$\lim_{k\to\infty} \hat{\theta}_{kj} = \begin{cases} \infty \text{ a.s.} & \text{if } t_j^* = 1, \\ -\infty \text{ a.s.} & \text{if } t_j^* = 0, \end{cases}$$

and $\mathrm{Proj}_{\boldsymbol{\Theta}}(\lim_{k\to\infty} \hat{\boldsymbol{\theta}}_k) = \boldsymbol{\theta}^*$ a.s. □

Because assumption v) in Theorem 3.3 is rather technical, we provide below a sufficient and easier to understand condition. For any $i$, denote $\mathscr{A}_i = \{\boldsymbol{\Delta}_k \mid \Delta_{ki} = 1 - 2t_i^*\}$ and $\mathscr{B}_i = \{\boldsymbol{\Delta}_k \mid \Delta_{ki} = -1 + 2\theta_i^*\}$. Note that $\mathscr{A}_i \cup \mathscr{B}_i = \{\boldsymbol{\Delta}_k \mid \Delta_{kj} = \pm 1 \text{ for all } j\}$ contains all the possible outcomes for $\boldsymbol{\Delta}_k$. Since $L(\mathbf{1}_p/2 +$

$\boldsymbol{\Delta}_k/2) - L(\mathbf{1}_p/2 - \boldsymbol{\Delta}_k/2) = -[L(\mathbf{1}_p/2 + (-\boldsymbol{\Delta}_k/2)) - L(\mathbf{1}_p/2 - (-\boldsymbol{\Delta}_k/2))]$ holds for every possible outcome of $\boldsymbol{\Delta}_k$, we have for any $k$ and $i$,

$$\begin{aligned}\bar{\boldsymbol{g}}(\mathbf{1}_p/2) &= \mathbb{E}\left[\frac{L(\mathbf{1}_p/2 + \boldsymbol{\Delta}_k/2) - L(\mathbf{1}_p/2 - \boldsymbol{\Delta}_k/2)}{\boldsymbol{\Delta}_k}\right] \\ &= \frac{1}{2^p}\sum_{\boldsymbol{\Delta}_k \in \mathscr{A}_i \cup \mathscr{B}_i} \frac{L(\mathbf{1}_p/2 + \boldsymbol{\Delta}_k/2) - L(\mathbf{1}_p/2 - \boldsymbol{\Delta}_k/2)}{\boldsymbol{\Delta}_k} \\ &= \frac{1}{2^{p-1}}\sum_{\boldsymbol{\Delta}_k \in \mathscr{A}_i} \frac{L(\mathbf{1}_p/2 + \boldsymbol{\Delta}_k/2) - L(\mathbf{1}_p/2 - \boldsymbol{\Delta}_k/2)}{\boldsymbol{\Delta}_k}.\end{aligned}$$

Recall that $\operatorname{sgn}(\bar{\boldsymbol{g}}(\mathbf{1}_p/2)) = 1 - 2\boldsymbol{\theta}^*$, then for any $i$, the condition

$$\begin{aligned}\bar{g}_i(\mathbf{1}_p/2)(1 - 2\theta_i^*) &= \frac{1}{2^{p-1}}\sum_{\boldsymbol{\Delta}_k \in \mathscr{A}_i} \frac{L(\mathbf{1}_p/2 + \boldsymbol{\Delta}_k/2) - L(\mathbf{1}_p/2 - \boldsymbol{\Delta}_k/2)}{1 - 2\theta_i^*}(1 - 2\theta_i^*) \\ &= \frac{1}{2^{p-1}}\sum_{\boldsymbol{\Delta}_k \in \mathscr{A}_i} [L(\mathbf{1}_p/2 + \boldsymbol{\Delta}_k/2) - L(\mathbf{1}_p/2 - \boldsymbol{\Delta}_k/2)] > 0,\end{aligned}$$

is equivalent to $\sum_{\boldsymbol{\Delta}_k \in \mathscr{A}_i} L(\mathbf{1}_p/2 - \boldsymbol{\Delta}_k/2) < \sum_{\boldsymbol{\Delta}_k \in \mathscr{A}_i} L(\mathbf{1}_p/2 + \boldsymbol{\Delta}_k/2)$. Since $\{\mathbf{1}_p/2 - \boldsymbol{\Delta}_k/2 \mid \boldsymbol{\Delta}_k \in \mathscr{A}_i\} = \{\boldsymbol{\theta} \in \{0,1\}^p \mid t_i = t_i^*\}$ and $\{\mathbf{1}_p/2 + \boldsymbol{\Delta}_k/2 \mid \boldsymbol{\Delta}_k \in \mathscr{A}_i\} = \{\boldsymbol{\theta} \in \{0,1\}^p \mid t_i \neq t_i^*\}$, we conclude that in the binary case, iv) is satisfied whenever the sum of loss function values at the points with $i$-th component equaling to $\theta_i^*$ has a smaller value than the sum of the loss function values at the points with $i$-th component not equaling to $\theta_i^*$ for all $i$. A sufficient condition is to have $L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}')$ whenever $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are such that there are more components $t_i = t_i^*$ than $t_i' = t_i^*$.

## 3.4 Rate of Convergence

In this section, we discuss the rate of convergence of MSPSA by deriving a big-$O$ bound to the MSE for $\hat{\theta}_k$. The bound allows for broad objective comparison of MSPSA with other algorithms, not tied to specific problem settings. For general discrete stochastic algorithms, if $\theta^*$ is unique and the sequence $\{\hat{\theta}_k\}$ contains only integer points, it is natural to study the rate of $\mathbb{P}(\hat{\theta}_k \neq \theta^*)$ going to $0$. However, since the points in the sequence $\{\hat{\theta}_k\}$ generated by MSPSA are generally not integer points, we study the mean-square error $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ instead. In the special case of $\boldsymbol{\Theta} = \mathbb{Z}^p$, since $\|\hat{\theta}_k - \theta^*\|^2 \geq 1/4$ when $\mathbf{Proj}_{\boldsymbol{\Theta}}(\hat{\theta}_k) \neq \theta^*$, then

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] &\geq 0^2\mathbb{P}(\mathbf{Proj}_{\boldsymbol{\Theta}}(\hat{\theta}_k) = \theta^*) + \frac{1}{4}\mathbb{P}(\mathbf{Proj}_{\boldsymbol{\Theta}}(\hat{\theta}_k) \neq \theta^*) \\ &= \frac{1}{4}\mathbb{P}(\mathbf{Proj}_{\boldsymbol{\Theta}}(\hat{\theta}_k) \neq \theta^*),\end{aligned}$$

which provides a way to compare with other discrete optimization algorithms. An extensive discussion on how the rate of convergence results can be used to make formal and theoretical comparisons to random search methods such as stochastic ruler and stochastic comparison can be found in Wang (2013, Chapter 5).

### 3.4.1 Explicit Upper Bound for Finite-Sample Performance

In this subsection, we derive an explicit upper bound for the mean-squared error $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ under both the finite-sample and asymptotic settings. Before showing the main theorem, let us first consider the relationship of the mean-squared errors between two consecutive iterations. From the main updating formula $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_k(\hat{\theta}_k)$, we can express $\mathbb{E}[\|\hat{\theta}_{k+1} - \theta^*\|^2]$ as

$$\mathbb{E}[\|\hat{\theta}_{k+1} - \theta^*\|^2] = \mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] - 2a_k \mathbb{E}[(\hat{\theta}_k - \theta^*)^T \hat{\boldsymbol{g}}_k(\hat{\theta}_k)] + a_k^2 \mathbb{E}[\|\hat{\boldsymbol{g}}_k(\hat{\theta}_k)\|^2]. \quad (3.20)$$

After recursively applying (3.20), we can decompose $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ into an error due to the initial condition and an error due to the cumulative effect. With the following two assumptions on gain sequences and search direction, we present the formal finite-sample bound result in Theorem 3.4 below.

**Assumption 3.10** (Step-size Sequences)**.** *For all $k$, assume that the gain sequences have the standard form $a_k = a/(k+1+A)^\alpha$ and $c_k = c/(k+1)^\gamma$, where $a, c, \alpha, \gamma > 0$, $A \geq 0$, and $3\gamma - \alpha/2 \geq 0$.*

**Assumption 3.11** (Search Direction)**.** *For all $k$, there exists some positive constant $\lambda$ such that $1 - \lambda a_k > 0$, and $\mathbb{E}[\bar{\boldsymbol{g}}(\hat{\theta}_k)^T(\hat{\theta}_k - \theta^*)] \geq \lambda \mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$.*

**Remark 3.5.** *Assumption 3.10 provides additional restrictions on the gain sequences beyond the general requirements in Assumption 3.7. In Assumption 3.11, the condition $\mathbb{E}[\bar{\boldsymbol{g}}(\hat{\theta}_k)^T(\hat{\theta}_k - \theta^*)] \geq \lambda \mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ is a stochastic analogue of the*

*definition of strong convexity for continuous case. When all the variables are continuous* $(d = 0)$*, we have* $\bar{\boldsymbol{g}}(\hat{\theta}_k) = \boldsymbol{g}(\hat{\theta}_k)$ *as discussed at the beginning of Section 3.3. Then Assumption 3.11 can be further reduced to* $\mathbb{E}[\boldsymbol{g}(\hat{\theta}_k)^T(\hat{\theta}_k - \theta^*)] \geq \lambda\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$*, which is weaker than the strong convexity assumption* $\boldsymbol{g}(\theta)^T(\theta - \theta^*) \geq \lambda\|\theta - \theta^*\|^2$ *(equivalently,* $\boldsymbol{H}(\theta) \succeq \lambda \boldsymbol{I}$*, i.e., the smallest eigenvalue of* $\boldsymbol{H}(\theta)$ *is bounded below by* $\lambda$ *for any* $\theta$ *(Zhou, 2018)). The condition* $1 - \lambda a_k > 0$ *for all* $k$ *is to guarantee a meaningful recursive relationship between* $\mathbb{E}[\|\hat{\theta}_{k+1} - \theta^*\|^2]$ *and* $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$*, as shown later in (3.26). Since* $a_k \to 0$*, it is easy to find a large* $K$ *such that* $1 - \lambda a_k > 0$ *holds for all* $k \geq K$*. By relabeling* $a_K$ *as* $a_0$*, we can also guarantee the condition holds for all* $k$*.*

**Theorem 3.4** (Finite-Sample Bound)**.** *Under Assumptions 3.1–3.8, 3.10, and 3.11, we have for all* $k$

$$\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] \leq \underbrace{P_k\mathbb{E}[\|\hat{\theta}_0 - \theta^*\|^2]}_{\text{initialization error}} + \underbrace{\frac{\|\boldsymbol{U}\|^2}{\lambda}\frac{a_0c_0^4P_0}{a_1c_1^4P_1}\int_0^k a_xc_x^4\frac{P_k}{P_x}dx}_{\text{bias error from gradient estimate}}$$
$$+ (\sigma_L^2 + \sigma_\varepsilon^2)\frac{a_0^2P_0}{a_1^2P_1}\left[\underbrace{d\int_0^k a_x^2\frac{P_k}{P_x}dx}_{\substack{\text{cumulative error}\\ \text{from discrete comp.}}} + \underbrace{(p-d)\kappa_2\int_0^k \frac{a_x^2}{4c_x^2}\frac{P_k}{P_x}dx}_{\substack{\text{cumulative error}\\ \text{from continuous comp.}}}\right], \tag{3.21}$$

*where*

$$P_x = \begin{cases} \exp\left\{\frac{\lambda a}{1-\alpha}[(1+A)^{1-\alpha} - (x+1+A)^{1-\alpha}]\right\} & \textit{if } \alpha \neq 1, \\ \left(\frac{1+A}{x+1+A}\right)^{\lambda a} & \textit{if } \alpha = 1, \end{cases} \tag{3.22}$$

**Remark 3.6.** *From (3.21), we see that the explicit upper bound for finite-sample performance is decomposed into an error due to the initial condition and an error due to the cumulative effect, which further represents a weighted average between the discrete and continuous components. It can be checked that $P_x$ is continuous in terms of $\alpha$ using L'Hôpital's rule (Wang, 2013, Corollary 3.1). Corollary 3.1 below gives an alternative form of the bound in (3.21) based on solving for the indicated integrals.*

*Proof.* From the main updating formula $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_k(\hat{\theta}_k)$, we can express $\mathbb{E}[\|\hat{\theta}_{k+1} - \theta^*\|^2]$ as (3.20). By adding and subtracting $2a_k\mathbb{E}[(\hat{\theta}_k - \theta^*)^T \bar{\boldsymbol{g}}(\hat{\theta}_k)]$ to the right-hand side of (3.20) and taking expectations on both sides, we have

$$\begin{aligned} \mathbb{E}[\|\hat{\theta}_{k+1} - \theta^*\|^2] &= \mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] - 2a_k\mathbb{E}[(\hat{\theta}_k - \theta^*)^T \bar{\boldsymbol{g}}(\hat{\theta}_k)] \\ &\quad + 2a_k\mathbb{E}[(\hat{\theta}_k - \theta^*)^T(\bar{\boldsymbol{g}}(\hat{\theta}_k) - \hat{\boldsymbol{g}}_k(\hat{\theta}_k))] + a_k^2\mathbb{E}[\|\hat{\boldsymbol{g}}_k(\hat{\theta}_k)\|^2] \\ &\leq \mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] - 2\lambda a_k\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] \\ &\quad + 2a_k\mathbb{E}[(\hat{\theta}_k - \theta^*)^T(\bar{\boldsymbol{g}}(\hat{\theta}_k) - \hat{\boldsymbol{g}}_k(\hat{\theta}_k))] + a_k^2\mathbb{E}[\|\hat{\boldsymbol{g}}_k(\hat{\theta}_k)\|^2]. \end{aligned} \tag{3.23}$$

After dropping the $2a_k$ multiplier, the third term on the right-hand side of (3.23)

is bounded as

$$\begin{aligned}\mathbb{E}[(\hat{\theta}_k - \theta^*)^T(\bar{\boldsymbol{g}}(\hat{\theta}_k) - \hat{\boldsymbol{g}}_k(\hat{\theta}_k))] &\leq \mathbb{E}\left[\|\sqrt{\lambda}(\hat{\theta}_k - \theta^*)\| \left\|\frac{1}{\sqrt{\lambda}}[\bar{\boldsymbol{g}}(\hat{\theta}_k) - \hat{\boldsymbol{g}}_k(\hat{\theta}_k)]\right\|\right] \\ &\leq \frac{\lambda}{2}\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] + \frac{1}{2\lambda}\mathbb{E}[\|\bar{\boldsymbol{g}}(\hat{\theta}_k) - \hat{\boldsymbol{g}}_k(\hat{\theta}_k)\|^2] \\ &\leq \frac{\lambda}{2}\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] + \frac{1}{2\lambda}\|\boldsymbol{U}\|^2 c_k^4, \end{aligned} \tag{3.24}$$

where the first inequality is due to Cauchy-Schwarz inequality, the second inequality is due to the relationship $[\|\sqrt{\lambda}(\hat{\theta}_k - \theta^*)\| - \|[\bar{\boldsymbol{g}}(\hat{\theta}_k) - \hat{\boldsymbol{g}}_k(\hat{\theta}_k)]/\sqrt{\lambda}\|]^2 \geq 0$, and the last inequality is due to $\mathbb{E}[\|\bar{\boldsymbol{g}}(\hat{\theta}_k) - \hat{\boldsymbol{g}}_k(\hat{\theta}_k)\|]^2 \leq \|\boldsymbol{U}\|^2 c_k^4$ from Theorem 3.1. From (3.16), we have that $\mathbb{E}[(2\boldsymbol{C}_k \circ \boldsymbol{\Delta}_k)^{-T}(2\boldsymbol{C}_k \circ \boldsymbol{\Delta}_k)^{-1}] \leq d + (p-d)\kappa_2/(4c_k^2)$. Hence, the last term on the right-hand side of (3.23) is bounded as

$$a_k^2\mathbb{E}[\|\hat{\boldsymbol{g}}_k(\hat{\theta}_k)\|^2] \leq (\sigma_L^2 + \sigma_\varepsilon^2)a_k^2\left[d + (p-d)\frac{\kappa_2}{4c_k^2}\right]. \tag{3.25}$$

Plugging (3.24) and (3.25) into (3.23), the recursive step $k$ to $k+1$ is bounded as

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}_{k+1} - \theta^*\|^2] \leq (1 - \lambda a_k)\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] + \frac{\|\boldsymbol{U}\|^2}{\lambda}a_k c_k^4 \\ + (\sigma_L^2 + \sigma_\varepsilon^2)a_k^2\left[d + (p-d)\frac{\kappa_2}{4c_k^2}\right]. \end{aligned} \tag{3.26}$$

Denote $R_k = 1 - \lambda a_k$. After recursively applying (3.26), we get

$$\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] \leq \left(\prod_{i=0}^{k-1} R_i\right) \mathbb{E}[\|\hat{\theta}_0 - \theta^*\|^2] + \frac{\|\boldsymbol{U}\|^2}{\lambda} \sum_{i=0}^{k-1} \left(\prod_{j=i+1}^{k-1} R_j\right) a_i c_i^4 + (\sigma_L^2 + \sigma_\varepsilon^2) \sum_{i=0}^{k-1} \left(\prod_{j=i+1}^{k-1} R_j\right) a_i^2 \left[d + (p-d)\frac{\kappa_2}{4c_i^2}\right]. \tag{3.27}$$

The rest of the proof is to bound every term on the right-hand side of (3.27) by showing:

i) $\prod_{i=0}^{k-1} R_i \leq P_k$,

ii) $\sum_{i=0}^{k-1} (\prod_{j=i+1}^{k-1} R_j) a_i c_i^4 \leq a_0 c_0^4 P_0/(a_1 c_1^4 P_1) \int_0^k a_x c_x^4 P_k / P_x dx$,

iii) $\sum_{i=0}^{k-1} (\prod_{j=i+1}^{k-1} R_j) a_i^2 \leq a_0^2 P_0/(a_1^2 P_1) \int_0^k a_x^2 P_k / P_x dx$,

iv) $\sum_{i=0}^{k-1} (\prod_{j=i+1}^{k-1} R_j) a_i/(4c_i^2) \leq a_0^2 P_0/(a_1^2 P_1) \int_0^k a_x^2 P_k/(4c_x^2 P_x) dx$.

**Proof of i):** Since $P_0 = 1$, it is sufficient to show that $P_{k+1} \geq P_k R_k$ for all $k$. The second-order Taylor expansion implies that when $\alpha \neq 1$,

$$\exp\left(-\frac{\lambda a}{1-\alpha}(k+1+A)^{1-\alpha}\right)\left[1 - \frac{\lambda a}{(k+1+A)^\alpha}\right] \leq \exp\left(-\frac{\lambda a}{1-\alpha}(k+1+1+A)^{1-\alpha}\right),$$

and when $\alpha = 1$,

$$\frac{1}{(k+1+A)^{\lambda a}}\left[1 - \frac{\lambda a}{k+1+A}\right] \leq \frac{1}{(k+1+1+A)^{\lambda a}},$$

both of which are the same as $P_{k+1} \geq P_k R_k$.

**Proof of ii):** Because the proof of i) implies $\prod_{j=i+1}^{k-1} R_j \leq P_k/P_{i+1}$ for all $i$, it is sufficient to show

$$\sum_{i=0}^{k-1} a_i c_i^4 \frac{P_k}{P_{i+1}} \leq \frac{a_0 c_0^4}{a_1 c_1^4} \frac{P_0}{P_1} \int_0^k a_x c_x^4 \frac{P_k}{P_x} dx$$

for all $k$, where we generalizes the notations of $a_k$ and $c_k$ for $k \in \mathbb{Z}^{\geq 0}$ to $a_x = a/(x+1+A)^\alpha$ and $c_x = c/(x+1)^\gamma$ for $x \in \mathbb{R}^{\geq 0}$. Decomposing $\int_0^k a_x c_x^4 P_k/P_x dx = \sum_{i=0}^{k-1} \int_i^{i+1} a_x c_x^4 P_k/P_x dx$, we then only need to show

$$a_k c_k^4 \frac{1}{P_{k+1}} \leq \frac{a_0 c_0^4}{a_1 c_1^4} \frac{P_0}{P_1} \int_k^{k+1} a_x c_x^4 \frac{1}{P_x} dx$$

for all $k$. Because the mean value theorem for integrals implies $\int_k^{k+1} a_x c_x^4/P_x dx = a_{k'} c_{k'}^4/P_{k'}$ for some $k' \in [k, k+1]$, it is equivalent to show

$$\frac{a_k c_k^4}{a_{k'} c_{k'}^4} \frac{P_{k'}}{P_{k+1}} \leq \frac{a_0 c_0^4}{a_1 c_1^4} \frac{P_0}{P_1}. \tag{3.28}$$

For the first term on the right-hand side of (3.28),

$$\frac{a_k c_k^4}{a_{k'} c_{k'}^2} = \left(\frac{k'+1+A}{k+1+A}\right)^\alpha \left(\frac{k'+1}{k+1}\right)^{4\gamma} \leq \left(\frac{1+1+A}{1+A}\right)^\alpha \left(\frac{1+1}{1}\right)^{4\gamma} = \frac{a_0 c_0^4}{a_1 c_1^4}. \tag{3.29}$$

For the second term on the right-hand side of (3.28), when $\alpha \neq 1$,

$$\begin{aligned}\frac{P_{k'}}{P_{k+1}} &= \exp\left(\frac{\lambda a}{1-\alpha}\left[(k+1+1+A)^{1-\alpha} - (k'+1+A)^{1-\alpha}\right]\right) \\ &\leq \exp\left(\frac{\lambda a}{1-\alpha}\left[(1+1+A)^{1-\alpha} - (1+A)^{1-\alpha}\right]\right) = \frac{P_0}{P_1}, \end{aligned} \tag{3.30}$$

where the inequality is due to the following two Taylor expansions:

$$(k+1+1+A)^{1-\alpha} = (k'+1+A)^{1-\alpha} + \frac{(1-\alpha)(k+1-k')}{(1+A+x)^{\alpha}} \text{ for some } x \in [k', k+1],$$

$$(1+1+A)^{1-\alpha} = (1+A)^{1-\alpha} + \frac{(1-\alpha)}{(1+A+x)^{\alpha}} \text{ for some } x \in [0, 1],$$

and when $\alpha = 1$,

$$\frac{P_{k'}}{P_{k+1}} = \left(\frac{k+1+1+A}{k'+1+A}\right)^{\lambda a} \leq \left(\frac{1+1+A}{1+A}\right)^{\lambda a} = \frac{P_0}{P_1}. \tag{3.31}$$

Therefore, (3.28) holds by combining (3.29)–(3.31).

**Proof of iii) and iv):** Similarly, observe that for any $k' \in [k, k+1]$,

$$\frac{a_k^2}{a_{k'}^2} = \left(\frac{k'+1+A}{k+1+A}\right)^{2\alpha} \leq \left(\frac{1+1+A}{1+A}\right)^{2\alpha} = \frac{a_0^2}{a_1^2}, \tag{3.32}$$

and

$$\frac{a_k^2 c_{k'}^2}{a_{k'}^2 c_k^2} = \left(\frac{k'+1+A}{k+1+A}\right)^{2\alpha}\left(\frac{k+1}{k'+1}\right)^{2\gamma} \leq \left(\frac{1+1+A}{1+A}\right)^{2\alpha} = \frac{a_0^2}{a_1^2}. \tag{3.33}$$

Combining (3.30)–(3.33), we have that iii) and iv) hold.

Finally, plugging the results of i)–iv) into (3.27) completes the proof. □

In Corollary 3.1 below, we provide a finite-sample bound by solving for the integrals in Theorem 3.4 explicitly.

**Corollary 3.1** (Explicit Finite-Sample Bound)**.** *Given the Assumptions of Theorem 3.4, for all* $k$*,*

$$\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] \leq P_k \mathbb{E}[\|\hat{\theta}_0 - \theta^*\|^2] + \frac{\|\boldsymbol{U}\|^2}{\lambda} \frac{a_0 c_0^4 P_0}{a_1 c_1^4 P_1} I_k^B + (\sigma_L^2 + \sigma_\varepsilon^2) \frac{a_0^2 P_0}{a_1^2 P_1} [d I_k^D + (p-d)\kappa_2 I_k^C],$$

*where*

$$\begin{aligned}
I_k^B &= \frac{ac^4}{\lambda a - 4\gamma(1+A)^\alpha} \left[ \frac{1}{(k+1)^{4\gamma}} - P_k \right], \\
I_k^D &= \frac{a^2}{\lambda a - \frac{\alpha}{(1+A)^{1-\alpha}}} \left[ \frac{1}{(k+1+A)^\alpha} - \frac{P_k}{(1+A)^\alpha} \right], \\
I_k^C &= \frac{a^2}{4c^2 \left[ \lambda a - \frac{\alpha - 2\gamma}{(1+A)^{1-\alpha}} \right]} \left[ \frac{1}{(k+1+A)^{\alpha-2\gamma}} - \frac{P_k}{(1+A)^{\alpha-2\gamma}} \right].
\end{aligned}$$

**Remark 3.7.** *The superscripts* $B, D, C$ *in* $I_k^B, I_k^D, I_k^C$ *are used to indicate the integrals in (3.21) pertaining to the bias term, the discrete components, and the continuous components, respectively.*

*Proof.* Corollary 3.1 shows that to get a computable finite-sample bound on the MSE, it is sufficient to bound the integrals in Theorem 3.4. The values of $k'$ in each use below are generally different.

i) When $\alpha \neq 1$, integration by parts implies

$$\int_0^k \frac{\lambda a}{(x+A+1)^{\alpha}(x+1)^{4\gamma}P_x}dx = \frac{1}{(k+1)^{4\gamma}P_k} - 1 + \int_0^k \frac{4\gamma}{(x+1)^{1+4\gamma}P_x}dx, \quad (3.34)$$

where the mean value theorem on the last integral implies

$$\int_0^k \frac{4\gamma}{(x+1)^{1+4\gamma}P_x}dx = \frac{4\gamma(k'+A+1)^{\alpha}}{(k'+1)}\int_0^k \frac{1}{(x+A+1)^{\alpha}(x+1)^{4\gamma}P_x}dx \quad (3.35)$$

for some $k' \in [0,k]$. Plugging (3.35) into (3.34) and combining the integrals yields

$$\int_0^k a_x c_x^4 \frac{P_k}{P_x}dx = \int_0^k \frac{ac^4}{(x+1+A)^{\alpha}(x+1)^{4\gamma}}\frac{P_k}{P_x}dx = \frac{ac^4}{\lambda a - \frac{4\gamma(k'+A+1)^{\alpha}}{k'+1}}\left[\frac{1}{(k+1)^{4\gamma}} - P_k\right].$$

When $\alpha = 1$, applying similar integration by parts approach on $\int_0^k a_x c_x^4 P_k/P_x dx$ gives

$$\int_0^k a_x c_x^4 \frac{P_k}{P_x}dx = \int_0^k \frac{ac^4(x+1+A)^{\lambda a-1}}{(1+A)^{\lambda a}(x+1)^{4\gamma}}P_k dx = \frac{ac^4}{\lambda a - \frac{4\gamma(k'+A+1)}{k'+1}}\left[\frac{1}{(k+1)^{4\gamma}} - P_k\right]$$

for some $k' \in [0,k]$. Hence, $\int_0^k a_x c_x^4 P_k/P_x dx \leq I_k^b$ holds for $1/2 < \alpha \leq 1$,

ii) When $\alpha \neq 1$, integration by parts implies

$$\int_0^k \frac{\lambda a}{(x+1+A)^{2\alpha}P_x}dx = \frac{1}{(k+1+A)^{\alpha}P_k} - \frac{1}{(1+A)^{\alpha}} + \int_0^k \frac{\alpha}{(x+1+A)^{1+\alpha}P_x}dx, \quad (3.36)$$

where mean value theorem on the last integral implies

$$\int_0^k \frac{\alpha}{(x+1+A)^{1+\alpha}P_x}dx = \frac{\alpha}{(k'+1+A)^{1-\alpha}}\int_0^k \frac{1}{(x+1+A)^{2\alpha}P_x}dx \tag{3.37}$$

for some $k' \in [0,k]$. Plugging (3.37) into (3.36) and combining the integrals yields

$$\int_0^k a_x^2\frac{P_k}{P_x}dx = \int_0^k \frac{a^2}{(x+1+A)^{2\alpha}}\frac{P_k}{P_x}dx = \frac{a^2}{\lambda a - \frac{\alpha}{(k'+1+A)^{1-\alpha}}}\left[\frac{1}{(k+1+A)^{\alpha}} - \frac{P_k}{(1+A)^{\alpha}}\right].$$

When $\alpha = 1$, direct evaluation of $\int_0^k a_x^2 P_k/P_x dx$ gives

$$\int_0^k a_x^2\frac{P_k}{P_x}dx = \int_0^k \frac{a^2(x+1+A)^{\lambda a-2}}{(1+A)^{\lambda a}}P_k dx = \frac{a^2}{\lambda a - 1}\left[\frac{1}{k+1+A} - \frac{P_k}{1+A}\right].$$

Hence, $\int_0^k a_x^2 P_k/P_x dx \le I_k^d$ holds for $1/2 < \alpha \le 1$.

iii) Recall that $\beta = \alpha - 2\gamma$ and when $\alpha \neq 1$, integration by parts implies

$$\int_0^k \frac{\lambda a}{(x+1+A)^{\alpha+\beta}P_x}dx = \frac{1}{(k+1+A)^{\beta}P_k} - \frac{1}{(1+A)^{\beta}} + \int_0^k \frac{\beta}{(x+1+A)^{1+\beta}P_x}dx, \tag{3.38}$$

where mean value theorem on the last integral implies

$$\int_0^k \frac{\beta}{(x+1+A)^{1+\beta}P_x}dx = \frac{\beta}{(k'+1+A)^{1-\alpha}}\int_0^k \frac{1}{(x+1+A)^{\alpha+\beta}P_x}dx, \tag{3.39}$$

for some $k' \in [0,k]$. Plugging (3.39) into (3.38) and combining the integrals

yields

$$\int_0^k \frac{a_x^2}{c_x^2}\frac{P_k}{P_x}dx \le \int_0^k \frac{a^2}{4c^2(x+1+A)^{\alpha+\beta}}\frac{P_k}{P_x}dx = \frac{a^2}{4c^2\left[\lambda a - \frac{\beta}{(k'+1+A)^{1-\alpha}}\right]}\left[\frac{1}{(k+1+A)^{\beta}} - \frac{P_k}{(1+A)^{\beta}}\right].$$

When $\alpha = 1$, direct evaluation of $\int_0^k a_x^2 P_k/(4c_x^2 P_x)dx$ gives

$$\int_0^k \frac{a_x^2}{4c_x^2}\frac{P_k}{P_x}dx \le \int_0^k \frac{a^2(x+1+A)^{\lambda a-2+2\gamma}}{4c^2(1+A)^{\lambda a}}P_k dx = \frac{a^2}{4c^2\left[\lambda a - (1-2\gamma)\right]}\left[\frac{1}{(k+1+A)^{1-2\gamma}} - \frac{P_k}{(1+A)^{1-2\gamma}}\right].$$

Hence, $\int_0^k a_x^2 P_k/(4c_x^2 P_x)dx \le I_k^c$ holds for $\alpha \neq 1$. □

### 3.4.2 Asymptotic Performance

Following the discussion of the upper bounds for finite-sample performance in Theorem 3.4 and Corollary 3.1, we now consider the asymptotic performance in the big-$O$ sense. For all sufficiently large $k$, we have

$$P_k = \begin{cases} O\left(k^{-\lambda a}\right) & \text{if } \alpha = 1, \\ O\left(\exp\left(-\frac{\lambda a}{1-\alpha}k^{1-\alpha}\right)\right) & \text{if } \alpha \neq 1. \end{cases} \tag{3.40}$$

Denote $\beta = \alpha - 2\alpha$. Plugging (3.40) into Corollary 3.1, we have

$$\begin{aligned}
&\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] \\
&= \begin{cases} O(k^{-\lambda a}) + O(k^{-4\gamma}) + dO(k^{-1}) + (p-d)O(k^{-(1-2\gamma)}) & \text{if } \alpha = 1, \\ O\left(\exp\left(-\frac{\lambda a}{1-\alpha}k^{1-\alpha}\right)\right) + O(k^{-4\gamma}) + dO(k^{-\alpha}) + (p-d)O(k^{-\beta}) & \text{if } \alpha \neq 1. \end{cases}
\end{aligned}$$

Since the exponential term $O(\exp(-\lambda a k^{1-\alpha}/(1-\alpha)))$ decreases much faster than other terms, we further simplify $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ to

$$\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] = \begin{cases} O(k^{-\lambda a}) + O(k^{-4\gamma}) + dO(k^{-1}) + (p-d)O(k^{-(1-2\gamma)}) & \text{if } \alpha = 1, \\ O(k^{-4\gamma}) + dO(k^{-\alpha}) + (p-d)O(k^{-\beta}) & \text{if } \alpha \neq 1. \end{cases}$$

Because the discussion only pertains to the case where $k$ is sufficiently large and $a_k \to 0$ as $k \to \infty$, we can choose a relatively large $a$ such that $\lambda a > 1$ while the constraint $1 - \lambda a_k > 0$ in Assumption 3.11 is still satisfied. Therefore, we can combine the two cases above to get

$$\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] = O(k^{-4\gamma}) + dO(k^{-\alpha}) + (p-d)O(k^{-\beta}). \tag{3.41}$$

Using the constraints on the values of $\alpha$ and $\gamma$ in Assumption 3.10, we see that the convergence rate of $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ going to $0$ is maximized by choosing $\alpha = 1$ and $\gamma = 1/6$. This optimal convergence rate is consistent with results in SPSA

(Spall, 1992) and DSPSA (Wang and Spall, 2013).

## 3.5 Comparison with DSPSA and SPSA

### 3.5.1 DSPSA: Fully Discrete Case

When all the variables are discrete ($d = p$) as discussed in Remark 3.3, $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ is unbiased for all $k$, i.e., $\boldsymbol{b}_k(\hat{\theta}_k) = \mathbb{E}[\hat{\boldsymbol{g}}_k(\hat{\theta}_k) - \bar{\boldsymbol{g}}(\hat{\theta}_k)|\hat{\theta}_k] = 0$. The recursive formula in (3.23) becomes

$$\mathbb{E}[\|\hat{\theta}_{k+1} - \theta^*\|^2] \leq (1 - 2\lambda a_k)\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] + p(\sigma_L^2 + \sigma_\varepsilon^2)a_k^2, \tag{3.42}$$

where the multiplier for $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ is no longer $1 - \lambda a_k$ since the inner product term $2a_k\mathbb{E}[(\hat{\theta}_k - \theta^*)^T(\bar{\boldsymbol{g}}(\hat{\theta}_k) - \hat{\boldsymbol{g}}_k(\hat{\theta}_k))]$ in (3.23) equals $0$. Denote

$$P_x^D = \begin{cases} \exp\left\{\frac{2\lambda a}{1-\alpha}[(1+A)^{1-\alpha} - (x+1+A)^{1-\alpha}]\right\} & \text{if } 1/2 < \alpha < 1, \\ \left(\frac{1+A}{x+1+A}\right)^{2\lambda a} & \text{if } \alpha = 1, \end{cases}$$

and assume $\mathbb{E}[\Lambda_{ki}^{-2}] = \tau_\Lambda^2$ for all $k$ and $i$. Following the proofs in Theorem 3.4, we can solve (3.42) recursively to get

$$\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] \leq P_k^D\mathbb{E}[\|\hat{\theta}_0 - \theta^*\|^2] + p\tau_\Lambda^2(\sigma_L^2 + \sigma_\varepsilon^2)\frac{a_0^2 P_0^D}{a_1^2 P_1^D}\int_0^k a_x^2 \frac{P_k^D}{P_x^D}dx, \tag{3.43}$$

which is identical to the finite-sample upper bound result for DSPSA (Wang and Spall, 2013, Theorem 1). Furthermore, solving the integral term in (3.43), we get

$$\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] \leq P_k^D \mathbb{E}[\|\hat{\theta}_0 - \theta^*\|^2] + p\tau_\Lambda^2(\sigma_L^2 + \sigma_\varepsilon^2)\frac{a_0^2 P_0^D}{a_1^2 P_1^D}\frac{a^2}{2a\lambda - \frac{\alpha}{(k'+1+A)^{1-\alpha}}}\left[\frac{1}{(k+1+A)^\alpha} - \frac{P_k^D}{(1+A)^\alpha}\right]$$

for some $k' \in [0, k]$, which is identical to the finite-sample upper bound result for DSPSA (Wang and Spall, 2013, Corollary 1).

### 3.5.2 SPSA: Fully Continuous Case

When all the variables are continuous ($d = 0$), we have $\bar{\boldsymbol{g}}(\hat{\theta}_k) = \boldsymbol{g}(\hat{\theta}_k)$ and Assumption 3.11 reduced to $\mathbb{E}[\boldsymbol{g}(\hat{\theta}_k)^T(\hat{\theta}_k - \theta^*)] \geq \lambda\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ for all $k$ as discussed in the Remark 3.5. Following the assumptions of SPSA in Spall (1992), let the eigenvalues of $\boldsymbol{H}(\theta^*)$ be $\lambda_1 \geq \cdots \geq \lambda_p > 0, \mathbb{E}[\Pi_{ki}^2] = \sigma_\Pi^2$ and $\mathbb{E}[\Pi_{ki}^{-2}] = \tau_\Pi^2$ for all $k$ and $i$. Given that $k^{\beta/2}(\hat{\theta}_k - \theta^*)$ is asymptotically normally distributed (Spall, 1992, Proposition 2), it is sufficient to show that $k^\beta\|\hat{\theta}_k - \theta^*\|^2$ is uniformly integrable (Laha and Rohatgi, 1979, p. 138) to achieve the asymptotic mean-squared error $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ as $k \to \infty$. The note after Spall (1992, Proposition 2) also discusses the requirement of uniform integrability, but it was not explicitly shown. Note that it is reasonable to expect that $\lambda \approx \lambda_p$ for sufficiently

large $k$ since $\hat{\theta}_k \to \theta^*$ a.s. and $\boldsymbol{H}(\hat{\theta}_k) \to \boldsymbol{H}(\theta^*)$ a.s.

Following the proof of Theorem 3.4 and the discussion in Section 3.4.2, we have $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] = O(k^{-4\gamma}) + O(k^{-\beta})$ from (3.41) and hence

$$k^{\beta}\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] = O(k^{\alpha-6\gamma}) + O(1).$$

Because $a - 6\gamma \leq 0$ from Assumption 3.10 , we must also have $k^{\beta}\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] < \infty$, i.e., $k^{\beta}\|\hat{\theta}_k - \theta^*\|^2$ is uniformly integrable. Therefore, we have as $k \to \infty$

$$k^{\beta}\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] \to \|\mu\|^2 + \tau_{\Pi}^2\sigma_{\varepsilon}^2\frac{a^2}{4c^2}\sum_{i=1}^{p}\frac{1}{2a\lambda_i - \beta_+}, \tag{3.44}$$

where $\beta_+ = \beta$ if $\alpha = 1$ and $\beta_+ = 0$ if $\alpha < 1$, and

$$\|\mu\|^2 = \begin{cases} 0 & \text{if } 3\gamma - \alpha/2 > 0, \\ a^2c^4\|(a\boldsymbol{H}(\theta^*) - \frac{1}{2}\beta_+\boldsymbol{I})^{-1}\boldsymbol{T}(\theta^*)\|^2 & \text{if } 3\gamma - \alpha/2 = 0, \end{cases}$$

where the $i$-th component of $\boldsymbol{T}(\theta^*)$ is $T_i(\theta^*) = \sigma_{\Pi}^2[L_{iii}^{(3)}(\theta^*) + 3\sum_{j=1,j\neq i}^{p} L_{ijj}^{(3)}(\theta^*)]/6$. Note that the terms $L_{ijl}^{(3)}, i, j, l$ all distinct, do not appear in $\|\mu\|^2$ because the corresponding $\mathbb{E}[\Delta_i\Delta_j\Delta_l/\Delta_i] = 0$. Asymptotically, we see that both (3.41) and (3.44) imply that $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2] = O(k^{-\beta})$.

It is also worth noting that, when all the variables are continuous, there is no explicit differentiability requirement on the loss function. Hence, the con-

vergence results can also be applied to non-differentiable optimization problems, which is related to the work in He et al. (2003) that uses convex analysis to establish convergence of SPSA for non-differentiable loss function.

## 3.6 Numerical Study

### 3.6.1 Skewed-quartic Loss

In this section, we carry out a numerical experiment to check the finite-sample bound in Corollary 3.1. Consider the following skewed-quartic loss function (Spall, 2005, Section 6.7),

$$L(\theta) = \theta^T \boldsymbol{B}^T \boldsymbol{B} \theta + 0.1 \sum_{i=1}^{p} (\boldsymbol{B}\theta)_i^3 + 0.01 \sum_{i=1}^{p} (\boldsymbol{B}\theta)_i^4, \tag{3.45}$$

where $p\boldsymbol{B}$ is an upper triangular matrix of 1's and $(\boldsymbol{B}\theta)_i$ represents the $i$-th component of the vector $\boldsymbol{B}\theta$. The skewed quartic loss function (3.45) has a unique optimal value $L(\theta^*) = 0$ at $\theta^* = \mathbf{0}$ and it has been extensively tested in recent literature (Spall, 2005; Wang and Spall, 2011, 2013; Wang et al., 2018b). To examine the performance in high-dimensional and noisy loss measurement environment, we set $p = 100$ with $\Theta = \mathbb{Z}^{50} \times \mathbb{R}^{50}$ and let $y(\theta) = L(\theta) + \varepsilon(\theta)$ with i.i.d. $\varepsilon(\theta) \sim \mathcal{N}(0, 25)$. We consider three algorithms, the proposed MSPSA, local random search (Spall, 2005, Chapter 2), and stochastic ruler (Yan and Mukai, 1992), to minimize (3.45) with initial estimate $\hat{\theta}_0 = \mathbf{1}_{100}$, where $\mathbf{1}_p$ denotes the

$p$-dimensional all-ones vector. Although the use of stochastic ruler for discrete stochastic optimization has been discussed in Yan and Mukai (1992) and Alrefaei and Andradóttir (2001), no proof of convergence for stochastic ruler in the mixed-variable case appears to exist. Despite the lack of theoretical justification, the stochastic ruler method is well suited to the stochastic mixed variable problem in terms of ease of implementation. In MSPSA, the gain sequence $a_k$ and perturbation sequence $c_k$ are taken as $a_k = a/(1 + A + k)^{\alpha}$ with $a = 0.1, A = 500, \alpha = 0.7$, and $c_k = c/(1 + k)^{\gamma}$ with $c = 0.5, \gamma = 0.167$. Note that $\alpha = 0.7$ and $\gamma = 0.167$ are nearly the smallest allowable values according to Assumptions 3.7 and 3.10. Slower decay rates for the gain sequences often enhance finite-sample performance of stochastic optimization (Spall, 2005, page 189). For all $k$, every component of the perturbation vector $\Delta_k$ follows an independent Bernoulli $\pm 1$ distribution with equal probabilities for each outcome. In local random search, a new candidate point is generated as $\hat{\theta}_k^{\text{cand}} = \text{Proj}_{\Theta}(\hat{\theta}_k + \boldsymbol{d}_k)$ with $\boldsymbol{d}_k \sim \mathcal{N}(\boldsymbol{0}, 0.1\boldsymbol{I}_{100})$. The projection operator moves each of the first 50 components to the nearest integer and makes no changes to the second 50 components. We set $\hat{\theta}_{k+1} = \hat{\theta}_k^{\text{cand}}$ if $y(\hat{\theta}_k^{\text{cand}}) < y(\hat{\theta}_k)$ and $\hat{\theta}_{k+1} = \hat{\theta}_k$ otherwise. In stochastic ruler, a new candidate point $\hat{\theta}_k^{\text{cand}}$ is uniformly sampled from $\Theta$. We then set $\hat{\theta}_{k+1} = \hat{\theta}_k^{\text{new}}$ with probability $\Pr(y(\hat{\theta}_k^{\text{new}}) \leq V)^{M_k}$ with $M_k = \lfloor 0.5 \log(k + 2) \rfloor + 1$ and $V \sim \text{Unif}(0, L(\hat{\theta}_0))$ and $\hat{\theta}_{k+1} = \hat{\theta}_k$ otherwise. The algorithm parameters for stochastic ruler have been tuned to achieve

approximately optimal performance. Given a total of 5000 noisy loss function measures per replicate and averaging over 20 independent replicates, we plot an approximation of the normalized mean-squared error for the estimate $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]/\|\hat{\theta}_0 - \theta^*\|^2$ in Figure 3.2 and the normalized error for the loss function $[L(\text{Proj}_{\Theta}(\hat{\theta}_k)) - L(\theta^*)]/[L(\hat{\theta}_0) - L(\theta^*)]$ in Figure 3.3. It is clear that MSPSA performs the best among the three algorithms.

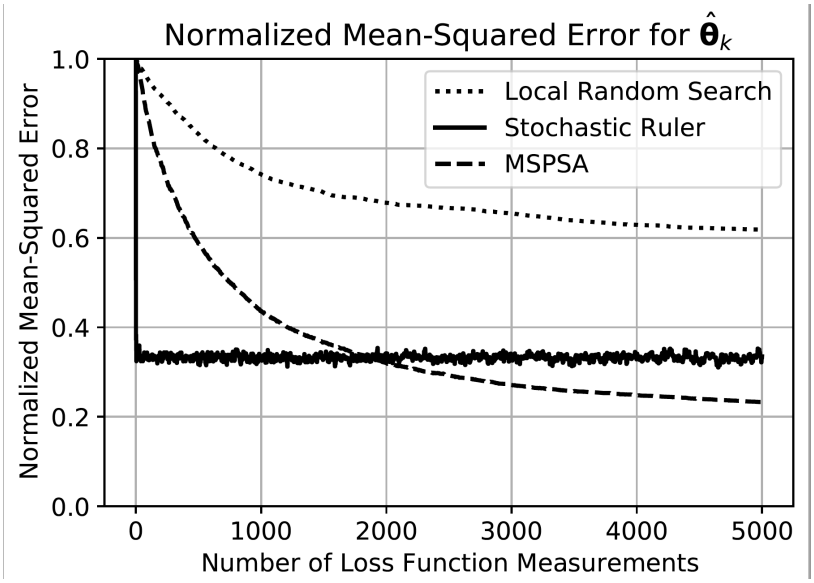



**Figure 3.2:** Performance of local random search, stochastic ruler, and MSPSA for the skewed-quartic function in terms of $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]/\|\hat{\theta}_0 - \theta^*\|^2$ across 5000 noisy function measurements and averaged over 20 independent replicates.

### 3.6.2 Finite-Sample Upper Bound

To further examine the accuracy of the finite-sample upper bound provided, we compare $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ with the computable finite-sample upper bound from Corollary 3.1. Because a very large number of iterations is typically required to

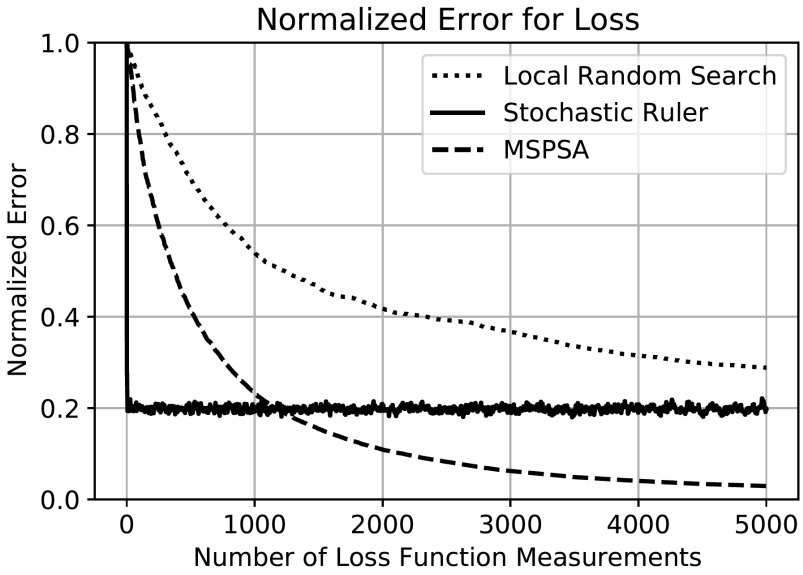



**Figure 3.3:** Performance of local random search, stochastic ruler, and MSPSA for the skewed-quartic function in terms of $[L(\text{Proj}_{\boldsymbol{\Theta}}(\hat{\theta}_k)) - L(\theta^*)]/[L(\hat{\theta}_0) - L(\theta^*)]$ across 5000 noisy function measurements and averaged over 20 independent replicates.

see the finite-sample upper bound approaching the empirical performance for problems with a larger dimension, we choose a modest $p = 10$ with $\boldsymbol{\Theta} = \mathbb{Z}^5 \times \mathbb{R}^5$ for the purpose of illustration. Setting $\hat{\theta}_0 = \mathbf{1}_{10}$, we implement MSPSA with the standard gain sequence $a_k = a/(1 + A + k)^\alpha$ and perturbation sequence $c_k = c/(1 + k)^\gamma$ where $a = 0.1, A = 100, \alpha = 0.7, c = 0.5$ and $\gamma = 0.167$. The mean-squared error $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ is estimated by taking an average of $20$ independent replicates for 5000 iterations. To compute the finite-sample bound, we choose $\kappa_0 = \kappa_1 = \kappa_2 = 1$ since every component of $\Delta_k$ follows an independent Bernoulli $\pm 1$ distribution with equality probabilities and $\sigma_\varepsilon^2 = 2$ since $\varepsilon(\theta) \sim \mathcal{N}(0, 1)$. The rest of the parameters are estimated by computing the corresponding values from $\hat{\theta}_k$ and taking an average of 500 independent replicates. Specifically, we

get the following estimates: $\lambda = 1.5, B_H = B_T = 0.8$, and $\sigma_L^2 = 15$. From Figure 3.4, we see that the finite-sample upper bound provides a reasonable approximation for the empirical MSPSA performance, where the gap becomes smaller as the number of iterations increase. Note that plot of finite-sample upper bound in Figure 3.4 curves backwards in early iterations. This is due to a conservative estimates of $\lambda, B_H, B_T$ and $\sigma_L^2$, which are taking significant effects in the early iterations, but becomes less significant in the later iterations.

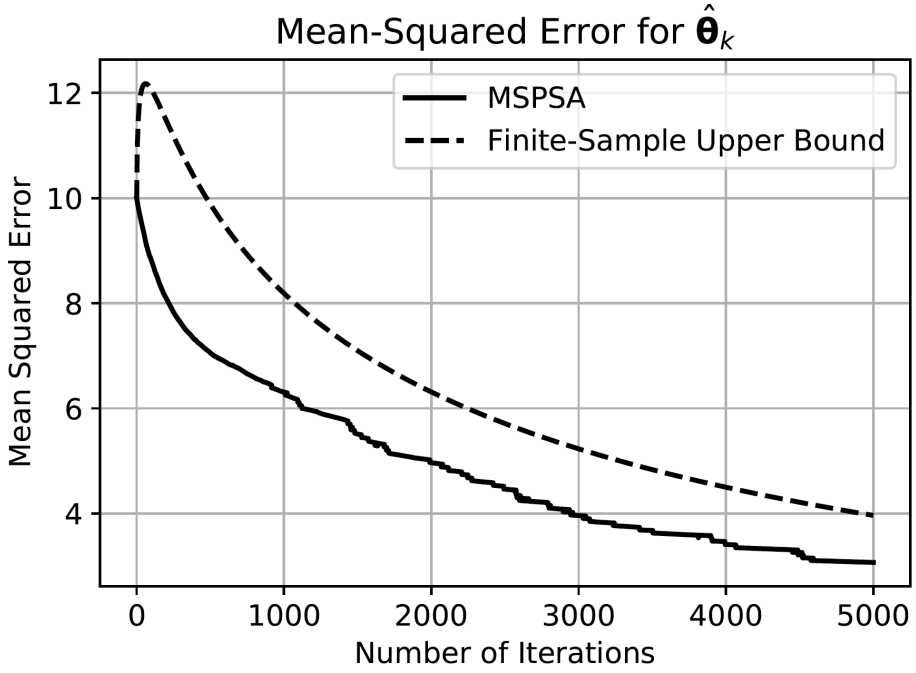



**Figure 3.4:** Performance of MSPSA and finite-sample upper bound for the skewed-quartic function in terms of $\mathbb{E}[\|\hat{\theta}_k - \theta^*\|^2]$ across 5000 iterations and averaged over 20 independent replicates.

### 3.6.3 Pressure Vessel Design

The pressure vessel design problem is a well-known benchmark in the field of structural design. Initially proposed in Sandgren (1988, 1990), this problem has been widely used in the literature to find the optimal design of a pressure vessel under various constraints (see, e.g., Santos Coelho, 2010; Yang and Deb, 2013; Gandomi et al., 2013; Yang et al., 2013 for a summary of main results). There are four parameters in the problem, namely, thickness of shell $t_1$, thickness of head $t_2$, inner radius $t_3$ and length of cylindrical section of vessel $t_4$. The first two parameters are discrete since the thickness can only be integer multiples of 0.0625 such that $t_1/0.0625 \in \{1, \ldots, 99\}$ and $t_2/0.0625 \in \{1, \ldots, 99\}$. The last two parameters are continuous such that $10 \leq t_3 \leq 200$ and $10 \leq t_4 \leq 200$. Denoting $\theta = [t_1, t_2, t_3, t_4]^T$, the objective function and the constraints are as follows

$$L(\theta) = 0.06224 t_1 t_2 t_3 + 1.7781 t_2 t_3^2 + 3.1661 t_1^2 t_4 + 19.84 t_1^2 t_3 \tag{3.46}$$

subject to $h_1(\theta) = -t_1 + 0.0193 t_3 \leq 0$

$$h_2(\theta) = -t_2 + 0.00954 t_3 \leq 0$$

$$h_3(\theta) = -\pi t_3^2 t_4 - \frac{4}{3}\pi t_3^3 + 1296000 \leq 0$$

$$h_4(\theta) = t_4 - 240 \leq 0$$

$$t_1/0.0625 \in \{1, \ldots, 99\}, t_2/0.0625 \in \{1, \ldots, 99\}$$

$$10 \leq t_3 \leq 200, 10 \leq t_4 \leq 200$$

According to Yang et al. (2013), the global optimal value is $L(\theta^*) = 6059.714$ at $\theta^* = [0.8125, 0.4375, 42.0984, 176.6366]^T$. To incorporate the constraints for using MSPSA, we set $\Theta = \{0.0625, \ldots, 6.1875\} \times \{0.0625, \ldots, 6.1875\} \times [10, 200] \times [10, 200]$ and use a penalty function method with Lagrangian multiplier $\lambda$ to minimize $L_\lambda(\theta) = L(\theta) + \lambda[\max\{h_1(\theta), 0\} + \max\{h_2(\theta), 0\} + \max\{h_3(\theta)/12960, 0\}]$ with an increasing $\lambda$. A similar approach of using the Lagrangian multiplier and the $\max$ operation is considered in Kannan and Kramer (1994). The intuition is that we only want to penalize for violating the constraints, and give no reward for feasible solutions. Without the $\max$ operation, any feasible solution has negative values of $h_i(\theta)$ for $i = 1, 2, 3$ contributing to the objective function, which will then dominate the objective value when $\lambda$ is very large. Note that the constraint $h_3(\theta) \leq 0$ is normalized by dividing $12960$ to ensure all the constraints are at the same order of magnitude and the constraint $h_4(\theta) \leq 0$ is dropped since it is always satisfied when $\theta \in \Theta$. Moreover, to test the performance of the algorithms with the noisy function measurements, an i.i.d $\varepsilon(\theta) \sim \mathcal{N}(0, 100)$ noise is added so that only the noisy function measurements $y_\lambda(\theta) = L_\lambda(\theta) + \varepsilon(\theta)$ are available. In practice, the noisy measurements can be collected as outputs of a black-box simulation program for modeling complex structural design, where the program itself is built based on the objective func-

tion in (3.46). The initial condition is set to $\hat{\theta}_0 = [1.125, 0.625, 50, 150]$, as in the previous literature.

The algorithm parameters for the three algorithms have been tuned to achieve approximately optimal performance for each algorithm, where multiple gain sequences are tested for MSPSA, normal proposal distribution with various variances are tested for local random search, and multiple monotonically increasing sequences $\{M_k\}$ are tested in stochastic ruler. We implement MSPSA with the standard gain sequence $a_k = a/(1 + A + k)^\alpha$ and perturbation sequence $c_k = c/(1+k)^\gamma$, where $a = 0.0005$ for the discrete variables, $a = 0.005$ for the continuous variables, $A = 100$, $\alpha = 0.7$, $c = 1$ and $\gamma = 0.1667$. For all $k$, every component of the perturbation vector $\Delta_k$ follows an independent Bernoulli $\pm 1$ distribution with equal probabilities. The Lagrangian multiplier is increasing with respect to the iteration number such that $\lambda_k = 1000 \log(k + 2)$. The use of increasing $\lambda_k$ is also considered in Wang and Spall (2008) to solve constrained stochastic optimization problems using SPSA algorithms. Similar to the numerical study in Section 3.6.1, we choose local random search and stochastic ruler for comparison. In local random search, a new point $\hat{\theta}^{\text{new}}$ is generated as $\hat{\theta}_k^{\text{new}} = \text{Proj}_\Theta(\hat{\theta}_k + \boldsymbol{d}_k)$ with $\boldsymbol{d}_k \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma = \text{diag}(0.025, 0.025, 2.5, 2.5)$ and the projection operator moves the argument to the nearest point within $\Theta$ based on Euclidean distance. We keep generating $\hat{\theta}_k^{\text{new}}$ until all the constraints are satisfied and then let $\hat{\theta}_k^{\text{cand}} = \hat{\theta}_k^{\text{new}}$. Finally, with the noisy function

measurement $y(\theta) = L(\theta) + \varepsilon(\theta)$, we set $\hat{\theta}_{k+1} = \hat{\theta}_k^{\text{cand}}$ if $y(\hat{\theta}_k^{\text{cand}}) < y(\hat{\theta}_k)$ and $\hat{\theta}_{k+1} = \hat{\theta}_k$ otherwise. In stochastic ruler, a new candidate point $\hat{\theta}_k^{\text{cand}}$ is uniformly sampled from $\Theta$ until all the constraints are satisfied. We then set $\hat{\theta}_k = \hat{\theta}_k^{\text{cand}}$ with probability $\Pr(y(\hat{\theta}_k^{\text{cand}}) \leq V)^{M_k}$ with $M_k = \lfloor \log(k+2) \rfloor + 1$ and $V \sim \text{Unif}(6000, L(\hat{\theta}_0))$ and $\hat{\theta}_{k+1} = \hat{\theta}_k$ otherwise. Table 3.1 and 3.2 show the terminal estimates $\text{Proj}_\Theta(\hat{\theta}_K)$ and the corresponding loss values $L(\text{Proj}_\Theta(\hat{\theta}_K))$ along with the normalized loss values $[L(\text{Proj}_\Theta(\hat{\theta}_K)) - L(\theta^*)]/[L(\theta_0) - L(\theta^*)]$. All the results are based on $20,000$ noisy function measures per replicate and averaged over 20 replicates.

**Table 3.1:** Terminal estimate of MSPSA, local random search and stochastic ruler based on 20,000 noisy function measurements per replicate and averaged over 20 independent replicates.

| Algorithm | $\text{Proj}_\Theta(\hat{\theta}_K)$ <br> $\theta^* = [0.8125, 0.4375, 42.0984, 176.6366]^T$ |
|---|---|
| Local Random Search | $[0.8750, 0.5, 45.0079, 148.7725]^T$ |
| Stochastic Ruler | $[1.125, 0.625, 50, 150]^T$ |
| MSPSA | $[0.8125, 0.4375, 41.8324, 182.9006]^T$ |

Evaluating the constraints from the terminal estimate form Table 3.1, we see that the estimate returned by MSPSA satisfies all the constraints with $h_1(\text{Proj}_\Theta(\hat{\theta}_K)) = -0.0051, h_2(\text{Proj}_\Theta(\hat{\theta}_K)) = -0.0384$ and $g_3(\text{Proj}_\Theta(\hat{\theta}_K))/12960 = -1.2468$, which are all negative as required. The MSPSA solution also correctly finds the optimal solution for the discrete variables, i.e., thickness of shell $t_1$ and the thickness of head $t_2$. The third variable, inner radius $t_3$, is also very

**Table 3.2:** Terminal objective function values of MSPSA, local random search and stochastic ruler based on 20,000 noisy function measurements per replicate and averaged over 20 independent replicates.

| Algorithm | $L(\text{Proj}_{\boldsymbol{\Theta}}(\hat{\boldsymbol{\theta}}_K))$ $L(\boldsymbol{\theta}^*) = 6059.714$ | Normalized Loss |
|---|---|---|
| Local Random Search | 6491.868 | 0.113 |
| Stochastic Ruler | 9886.346 | 1 |
| MSPSA | 6160.702 | 0.026 |

close to the optimal value. For the last variable, length of cylindrical $t_4$, we note that, given the initial value of $150$, the estimate is successfully moving towards the optimal value. In contrast, local random search fails to find the optimal value of the discrete variables and the last variable doesn't seem to move at all. Due to the high measurement noise relatively to $L(\boldsymbol{\theta}^*)$, stochastic ruler fails to find any estimate better than the initial estimate. Further, the loss value drops to within $2.6\%$ of the optimal $L(\boldsymbol{\theta}^*)$ relative to $L(\hat{\boldsymbol{\theta}}_0)$.

## 3.7 Conclusion

In this work, we propose the MSPSA algorithm to solve mixed discrete-continuous optimization problems when only noisy values of the loss function are available to carry out the optimization. Special cases of MSPSA include the fully discrete setting of Wang and Spall (2011) and Wang (2013), and the original fully continuous setting of Spall (1992). By taking advantage of "gradient-

type" that is efficiently produced from noisy zeroth-order (loss function) information, MSPSA can efficiently handle mixed-variable stochastic optimization problems. The almost sure convergence and rate of convergence of the sequence generated by MSPSA have been derived. The finite-sample bound and the asymptotic performance are shown to enjoy an intuitive balance between the discrete and continuous components. Further, the general rate of convergence here reduces to the previously known rate in the special cases of all-discrete and all-continuous parameters. The rate allows for objective performance comparison of MSPSA with existing or future other methods for discrete, continuous, or mixed cases using only noisy loss function measurements. Overall, MSPSA appears to be the first algorithm that is formally designed for optimization in the general mixed discrete and continuous case with only noisy "zeroth-order" information of loss functions.

# Chapter 4

# Complex Simultaneous Perturbation Stochastic Approximation Algorithm

## 4.1 Introduction

Besides SPSA-based methods, another method for estimating the gradient is the complex-step (CS) gradient approximation (Lyness and Moler, 1967), which uses complex variables to estimate the gradient at the cost of a single function measurement when $\theta \in \mathbb{R}$. It is shown in Squire and Trapp (1998); Martins et al. (2001, 2003); Abreu et al. (2018); Higham (2018) that CS gradient approximation provides a significant improvement compared with

FDSA. Specifically, the CS gradient approximation will not suffer any numerical errors, as no difference operation is involved. The numerical-error-free characteristic of CS gradient approximation—a reliable result will be returned regardless of the step size—can be beneficial in modern simulation and application experiments. More advantages related to numerical stability can be found in Martins et al. (2003). This contrasts with other methods, where the round-off error becomes dominant for a small step size as discussed in Squire and Trapp (1998).

To extend the CS gradient approximation to multivariate variables and noisy function measurement environment, Nikolovski and Stojkovska (2018) proposes the CS-FDSA, which uses the complex variable to estimate every component of the gradient separately, at the cost of $p$ function measurements per iteration. Several classical testing problems from Krejić et al. (2015) are implemented with additive real-valued Gaussian noise and complex-valued circular noise. Regardless of the noise levels and underlying line-search algorithms, CS-FDSA exhibits better and more robust performance than the standard FDSA algorithm. To overcome the dimension-dependent number of function measurements in CS-FDSA, while maintaining the numerical stability and robustness, we propose the complex-step simultaneous perturbation (CS-SP) gradient approximation. CS-SP combines the ideas of SP and CS gradient approximation to estimate every component of the gradient simul-

taneously by constructing a $p$-dimensional complex-valued random perturbation vector. Based on the CS-SP gradient approximation, we also propose a gradient-descent-type stochastic approximation algorithm called CS-SPSA. Besides the numerical advantages of CS gradient approximation, it is worth noting that although only one noisy function measurement is used in CS-SPSA per iteration, the computational time of complex variable operations is typically slower than the time of real variable operations. Due to the extra imaginary variables, one complex number operation may take two to four times computational cost than one real number operation. Nonetheless, in terms of computational time, CS-SPSA is still considerably more favorable in high-dimensional problems than FDSA and CS-FDSA due to the constant number of loss function measurements at each iteration. When compared with SPSA, however, the computational time per iteration may be slower because of the involvement of complex number operations.

For CS-SPSA to be applicable, the loss function is required to be analytic so that the function measurement can be extended to complex space. When the original functions ($L : \mathbb{R}^p \to \mathbb{R}$ and $\ell : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$) are implemented by numerical algorithms, Martins et al. (2003) explains how the operations can be "complexified" ($L : \mathbb{C}^p \to \mathbb{C}$ and $\ell : \mathbb{C}^p \times \mathbb{R}^q \to \mathbb{C}$) so that the CS gradient approximation yields the correct results. Mathematically, for any $\theta_0 \in \Theta$ with $\Theta$ being an open set in the complex plane, one can write $L(\theta) = \sum_{n=0}^{\infty} \beta_n (\theta - \theta_0)^n$,

where the coefficients $\beta_0, \beta_1, \ldots,$ are real numbers and the series is convergence to $L(\theta)$ for $\theta$ in a neighborhood of $\theta_0$. Because of the nature of complex numbers, it might also be difficult to implement the proposed algorithm in the real world by interacting with the physical system, since the might be no meaningful physical interpretation of complex numbers. However, when the physical system is coded in computer programs (as a black box or complicated software), our algorithm can be useful in finding the numerical value of the parameter of interest that can be used in guiding practical problems. It is also shown that although complex numbers are used, it only remains in the function measurements. The estimated parameter value itself as well as the corresponding gradient approximation are always real-valued and can carry physical meanings. Table 4.1 provides a short summary of cases where the CS gradient approximation can be applied or not. The complex variable operations (e.g., basic arithmetic operations, exponentiation, logarithmic, etc.) are readily available in many modern programming languages, such as Python, MATLAB, R, etc.

Despite the limitations of using a complex-valued parameter for some applications, we summarize some practical problems that have been discussed in previous literature. Several multidisciplinary programs are considered in Martins et al. (2001), including a two-dimensional finite volume solver for Euler equations (Martins et al., 2000), a high-fidelity aero-structural solver for

**Table 4.1:** Examples of Applicable and Not Applicable Applications for CS Gradient Approximation

| Applicable | Not Applicable |
|---|---|
| Simulation-based optimization with complex analytical loss functions | Physical experiments with real-valued only parameters |
| Neural networks training with complex analytical activation functions | Machine learning problems that cannot be evaluated at complex variables |
| Recursive maximum likelihood estimation with complex analytical density function | Computer code or simulations that cannot be evaluated at complex variables |
| Optimal control with computer code that can be evaluated at complex variables | |

wing design optimization problems (Reuther et al., 1999) and supersonic viscous/inviscid solver (Sturdza et al., 1999). The results presented in their work show that the CS gradient approximation is easy to implement and can generate accurate results for sensitivity analysis. Other optimization problems include a first-order linear system with a finite $L_2$ gain cost function and a second-order nonlinear system with a cost function corresponding to the maximum overshoot in response to a unit step reference demand (Kim et al., 2006). More recently, Balzani et al. (2015) considers the numerical calculation of thermo-mechanical problems at large strains. The multiple time-delayed differential equations with non-smoothness sensitivities are investigated in Banks et al. (2015). Stochastic optimization problems with complex-valued noise can also be found in signal processing (Ciblat and Ghogho, 2004) and electrical engi-

neering (Javidi et al., 2010).

The remainder of this chapter is organized as follows. Section 4.2 discusses how CS-SPSA can be applied in general stochastic optimization problems, along with theoretical properties such as convergence and asymptotic distribution. Section 4.3 applies CS-SPSA in model-free control problems, where a time-varying loss function is studied and the convergence of the estimated is shown. Numerical study and conclusion are made in Section 4.4 and 4.5, respectively.

## 4.2 General Stochastic Optimization

In this section, we consider minimizing a general expected loss $L(\theta) = \mathbb{E}[\ell(\theta, \boldsymbol{v})]$, where $\theta \in \mathbb{R}^p$ represents the parameter of interest and $\boldsymbol{v} \in \mathbb{R}^q$ denotes a random variable or vector following some unknown distribution. The variable $\boldsymbol{v}$ commonly stands for the random effect in the process generating the system output or the amalgamation of various random effects. With the presence of $\boldsymbol{v}$, we denote $\ell(\theta, \boldsymbol{v})$ as the noisy function measurement of the expected loss function $L(\theta)$ at some chosen parameter value $\theta$, where the expectation in $L(\theta)$ is taken over all randomness embodied in $\boldsymbol{v}$. To the minimize the expected loss $L(\theta)$, a majority of algorithms resort to finding the root to the gradient function $\boldsymbol{g}(\theta) = \partial L(\theta)/\partial\theta$. Often, neither $L(\theta)$ nor $\boldsymbol{g}(\theta)$ can be computed ex-

plicitly, as both the distribution of $\boldsymbol{v}$ and the form of $\ell(\theta, \boldsymbol{v})$ remain inaccessible. Therefore, this section proposes to use CS-SPSA to minimize $L(\theta)$ using the noisy function measurements $\ell(\theta, \boldsymbol{v})$ only. We show that CS-SPSA converges to the optimal point at an accelerated rate of $k^{-1/2}$, which is faster than the standard convergence rate of $k^{-1/3}$ (Spall, 1992). Furthermore, the asymptotic results suggest that CS-SPSA achieves the same level of accuracy of SPSA as if the user have total control of the noise in the function measurements, i.e., pure common random numbers. In a nutshell, within the class of problems for which complex perturbation are meaningful, CS-SPSA has the following advantages: i) only *one* function measurement is required at each iteration, which is *independent* of the dimension of the problem; ii) there is no round-off errors caused by subtraction, as opposed to the case of noise-free or controlled noise function measurements; iii) faster convergence of of $k^{-1/2}$ is achieved, compared with $k^{-1/3}$ in standard stochastic optimization algorithms.

### 4.2.1 Algorithm Description

The CS gradient approximation estimates the gradient using complex variable, applicable for noise-free evaluation and one-dimensional parameter. To extend the CS gradient approximation to multivariate and noisy function measurement settings, Nikolovski and Stojkovska (2018) proposes the following CS

gradient estimation as

$$\hat{\boldsymbol{g}}_k^{\mathrm{CS}}(\hat{\theta}_k) = \frac{1}{c}\begin{bmatrix} \Im(\ell(\hat{\theta}_k + ic_k\boldsymbol{u}_1, \boldsymbol{v}_k^{(1+)})) \\ \vdots \\ \Im(\ell(\hat{\theta}_k + ic_k\boldsymbol{u}_p, \boldsymbol{v}_k^{(p+)})) \end{bmatrix}, \tag{4.1}$$

where $\Im(\cdot)$ denotes the imaginary part of the complex argument. Similar to FD gradient approximation, (4.1) estimates every component of the gradient individually, giving a total of $p$ function measurements. The CS gradient estimate is also used to solve optimization problems with only noisy function measurements. However, Nikolovski and Stojkovska (2018) only considers the non-monotone line-search methods with numerical illustrations.

Although both the FD and CS gradient approximations incur an approximation error of $O(c_k^2)$, the former requires twice as many function measurements as the latter. Besides the query efficiency, CS gradient approximation can be much more numerically stable for small $c_k$ (Martins et al., 2003). The subtractive cancellation error hinders practitioners from achieving accurate results. When noise-free or controlled noise (i.e., $\boldsymbol{v}_k^{(j+)} = \boldsymbol{v}_k^{(j-)}$ for $j = 1, \ldots, p$ in (2.4) $\boldsymbol{v}_k^{(+)} = \boldsymbol{v}_k^{(-)}$ in (2.5)) function measurements are used, because of the finite machine precision, there exists a constant $\underline{c}$ such that when $c_k \leq \underline{c}$ computers can no longer distinguish between the function measurements at $\hat{\theta}_k + c_k\boldsymbol{u}_j$ and $\hat{\theta}_k - c_k\boldsymbol{u}_j$ for $j = 1, \ldots, p$ in (2.4) or $\hat{\theta}_k + c_k\boldsymbol{\Delta}_k$ and $\hat{\theta}_k - c_k\boldsymbol{\Delta}_k$ in (2.5), which causes

$\hat{\boldsymbol{g}}_k^{\mathrm{FD}}(\hat{\theta}_k) = \mathbf{0}$ or $\hat{\boldsymbol{g}}_k^{\mathrm{SP}}(\hat{\theta}_k) = \mathbf{0}$. Note that $c_k$ is typically very small and converges to $0$ in common optimization algorithms in order to achieve an accurate gradient approximation.

On the contrary, when controlled noise is not possible, we see that the difference of the function measurements $\ell(\hat{\theta}_k + c_k \boldsymbol{u}_j, \boldsymbol{v}_k^{(j+)}) - \ell(\hat{\theta}_k - c_k \boldsymbol{u}_j, \boldsymbol{v}_k^{(j-)})$ for all $j = 1, \ldots, p$ in (2.4) or $\ell(\hat{\theta}_k + c_k \boldsymbol{\Delta}_k, \boldsymbol{v}_k^{(+)}) - \ell(\hat{\theta}_k - c_k \boldsymbol{\Delta}_k, \boldsymbol{v}_k^{(-)})$ in (2.5) does not converge to 0 as $c$ goes to 0 due to the randomness of noisy function measurements. As a consequence, the gradient estimate $\hat{\boldsymbol{g}}_k^{\mathrm{FD}}(\hat{\theta}_k)$ or $\hat{\boldsymbol{g}}_k^{\mathrm{SP}}(\hat{\theta}_k)$ will explode since $c_k$ in the denominator is going to $0$. It is worth noting that such numerical issue can be resolved by not dividing $c_k$ in the gradient estimate and using $a_k/c_k$ as the gain step size. However, the gradient estimate itself remains inaccurate and unstable, especially for sensitivity analysis where the primary interest is the gradient estimate per se (not the solution of an associated optimization problem). Although CS gradient estimate also pertains to dividing by a small step size $c_k$, the gradient estimate itself does not blow up. This numerical superiority is credited to the fact that $\Im[\ell(\hat{\theta}_k + i c_k \boldsymbol{u}_j, \boldsymbol{v}_k^{(j+)})]$ converges to $0$ for $j = 1, \ldots, p$ as $c_k$ converges to $0$. This balances out the decreasing $c_k$ value in the denominator and makes the CS gradient estimate accurate and stable.

To illustrate, we consider a toy example of estimating the gradient of $L(\theta) = \mathbb{E}[(\theta - v)^2]$, where $\theta \in \mathbb{R}$ and $v \sim N(0, \sigma^2)$. Using only the measurement of

$\ell(\theta, \omega) = (\theta - \omega)^2$, we see that CS gradient estimate gives

$$\frac{\Im[\ell(\theta + ic, v)]}{c} = \frac{\Im[(\theta - v)^2 + 2ic(\theta - v) - c^2]}{c} = 2(\theta - v),$$

which is exactly the gradient of $\ell(\theta, v)$ regardless of how small $c$ is. On the other hand, the FD gradient estimate gives

$$\begin{aligned}\frac{\ell(\theta + c, v^{(+)}) - \ell(\theta - c, v^{(-)})}{2c} &= \frac{(2\theta - v^{(+)} - v^{(-)})(2c - v^{(+)} + v^{(-)})}{2c} \\ &= 2\left(\theta - \frac{v^{(+)} + v^{(-)}}{2}\right)\left(1 - \frac{v^{(+)} - v^{(-)}}{2c}\right),\end{aligned}$$

which is less accurate and blows up when $c$ is sufficiently small.

Motivated by the SP and CS gradient approximations, we propose the following CS-SP gradient approximation

$$\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k) = \frac{\Im(\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k))}{c_k\boldsymbol{\Delta}_k}. \tag{4.2}$$

The CS-SP gradient approximation (4.2) inherits the benefits of both SP and CS gradient approximations: query advantage (requiring one function measurement) and numerical stability (avoiding subtraction altogether). Note that $\ell(\hat{\theta}_k, \boldsymbol{v}_k)$ is deemed as a noisy function measurement of $L(\hat{\theta}_k)$. Similarly, we can view $\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)$ as a complex-valued noisy function measurement of

$L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)$ and separate the real part and imaginary part of the noise as

$$\varepsilon_k^{\mathrm{R}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k) = \Re(\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)) - \Re(L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k)), \tag{4.3}$$

$$\varepsilon_k^{\mathrm{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k) = \Im(\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)) - \Im(L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k)), \tag{4.4}$$

where $\Re(\cdot)$ denotes the real part of the argument. It is straightforward that both $\mathbb{E}[\varepsilon_k^{\mathrm{R}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)] = 0$ and $\mathbb{E}[\varepsilon_k^{\mathrm{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)] = 0$ when the expectation is taken over $\boldsymbol{v}_k$. Note that both $\varepsilon_k^{\mathrm{R}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)$ and $\varepsilon_k^{\mathrm{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)$ are real-valued outputs and we can re-express $\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)$ in terms of (4.3) and (4.4) as

$$\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k) = L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k) + \varepsilon_k^{\mathrm{R}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k) + i\varepsilon_k^{\mathrm{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k). \tag{4.5}$$

It is worth noting that the CS-SP gradient approximation relies on a fundamental assumption that the function measurement can be collected at the complex-valued variable $\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k$. To use the CS-SP gradient approximation (4.2) in stochastic optimization algorithms, we proposed the following algorithm to minimize the objective function $L(\theta)$.

- **Step 0 (Initialization):** Set index $k = 0$. Pick an initial guess $\hat{\theta}_0$ and nonnegative coefficients $a, c, A, \alpha$, and $\gamma$ in the gain sequences $a_k = a/(k + 1 + A)^{\alpha}$ and $c_k = c/(k+1)^{\gamma}$.

- **Step 1 (Perturbation Vectors):** Construct a $p$-dimensional random perturbation vector $\boldsymbol{\Delta}_k$ using Monte Carlo algorithms, where all the components of $\boldsymbol{\Delta}_k$ are independently generated from a mean-zero symmetric proability distribution satisfying certain regularity conditions.

- **Step 3 (Gradient Approximation):** Construct the CS-SP gradient approximation as

$$\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k) = \frac{\Im(\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k))}{c_k\boldsymbol{\Delta}_k}$$

- **Step 4 (Iterative Update):** Update the parameter estimate using the standard scheme as

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k) \tag{4.6}$$

- **Step 5 (Iteration or Termination):** Return to Step 1 with $k+1$ replacing $k$. Terminate the algorithm if there is little change of $\hat{\theta}_k$ in several successive iterates or if the maximum allowable number of iterations has been reached.

**Remark 4.1.** *Although the function measurement $\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)$ is a complex-valued scalar, the gradient approximation $\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)$ and the parameter estimate $\hat{\theta}_k$ are always real-valued. We focus on the most fundamental gradient-descent-type updating scheme as in (4.6) for this work and discuss both the the-*

*oretical and empirical performance similar to SPSA, which is also implemented in the standard updating scheme (4.6). It is worth noting that choosing $a_k$ is important for empirical performance, but we think methods for accelerating might best be left for future work.*

### 4.2.2 Convergence

#### 4.2.2.1 Bias of Gradient Estimate

We examine the bias of $\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)$ as an estimator of $\boldsymbol{g}(\hat{\theta}_k)$ and show that the bias goes to $0$ as $k \to \infty$. Similar to Spall (1992, Lemma 1), an explicit bound for the bias is given in Theorem 4.1.

**Assumption 4.1** (Loss Function)**.** *Assume that $L(\cdot)$ can be extended to a complex analytic function. Further assume that, for almost all $\hat{\theta}_k$, independently from $k$, assume that there exists an open neighborhood of $\hat{\theta}_k$ in the complex space such that for any $\theta$ in that neighborhood, $|f(\theta)| \leq B_f$ for some positive constant $B_f$.*

**Assumption 4.2** (Random Perturbation)**.** *For all $k$ and $j$, assume that $\Delta_{kj}$ are independent and identically distributed, symmetrically distributed about 0. Further assume that $|\Delta_{kj}| \leq \kappa_1$ for some positive constant $\kappa_1$ and there exists some positive constant $\delta_1 > 0$ such that $\mathbb{E}[1/|\Delta_{kj}|^{1+\delta_1}] \leq \kappa_2$ for some positive constant $\kappa_2$.*

**Assumption 4.3** (Measurement Noise)**.** *For all $k$, denote the filtration $\mathscr{F}_k = \{\theta_0, \ldots, \theta_k\}$ and assume that $\mathbb{E}[\varepsilon_k^R(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)|\mathscr{F}_k, \boldsymbol{\Delta}_k] = 0$ and $\mathbb{E}[\varepsilon_k^I(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)|\mathscr{F}_k, \boldsymbol{\Delta}_k] = 0$.*

**Remark 4.2.** *Assumption 4.1 contains a necessary extension of $L(\theta)$ to complex space to enable Taylor expansion of analytic complex functions Squire and Trapp (1998). The remaining regularity assumptions on $L^{(3)}(\theta)$ and Assumption 4.2 are identical to the assumptions in (Spall, 1992, Lemma 1). Assumption 4.3 ensures a zero-mean noise, which holds trivially given $L(\theta) = \mathbb{E}[\ell(\theta, \boldsymbol{v})]$ and the definitions of $\varepsilon_k^R(\cdot,\cdot)$ and $\varepsilon_k^I(\cdot,\cdot)$ in (4.3) and (4.4).*

**Theorem 4.1** (Wang and Spall, 2021)**.** *Under Assumptions 4.1–4.3, we have*

$$\boldsymbol{b}_k(\hat{\theta}_k) = \mathbb{E}[\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k) - \boldsymbol{g}(\hat{\theta}_k) \mid \mathscr{F}_k] = O(c_k^2). \tag{4.7}$$

**Remark 4.3.** *The bias of the CS-SP gradient approximation has the same order $O(c_k^2)$ as the SP gradient approximation (Spall, 1992, Lemma 1), where both gradient approximations becomes asymptotically unbiased as $c_k \to \infty$.*

*Proof.* Following the proof of Spall (1992, Lemma 1), applying Taylor expansion to expand $L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k)$ around $\hat{\theta}_k$ gives

$$L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k) = L(\hat{\theta}_k) + ic_k\boldsymbol{\Delta}_k^T\boldsymbol{g}(\hat{\theta}_k) - \frac{1}{2}c_k^2\boldsymbol{\Delta}_k^T\boldsymbol{H}(\hat{\theta}_k)\boldsymbol{\Delta}_k + r_3(\hat{\theta}_k), \tag{4.8}$$

where $\boldsymbol{H}(\theta) = \partial L(\theta)/\partial\theta\partial\theta^T$ is the Hessian matrix of $L(\theta)$ and $r_3(\hat{\theta}_k)$ is the higher-order remainder term of the Taylor expansion such that

$$r_3(\hat{\theta}_k) = \sum_{n=3}^{\infty} \frac{1}{n!}(ic_k)^n L^{(n)}(\hat{\theta}_k)[\boldsymbol{\Delta}_k \otimes^n \boldsymbol{\Delta}_k]$$

with $L^{(n)}(\hat{\theta}_k)$ being the $n$-th derivative of $L(\cdot)$ evaluated at $\hat{\theta}_k$ and $\otimes$ being the Kronecker product. Then taking the imaginary part of (4.8), we obtain

$$\Im(L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k)) = c_k\boldsymbol{\Delta}_k^T\boldsymbol{g}(\hat{\theta}_k) + \Im(r_3(\hat{\theta}_k)). \tag{4.9}$$

Now, plugging (4.5) and (4.9) into the definition of $\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)$ in (4.2) yields

$$\begin{aligned}
\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k) &= \frac{\Im(L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k) + \varepsilon_k^{\text{R}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k) + i\varepsilon_k^{\text{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k))}{c_k\boldsymbol{\Delta}_k} \\
&= \frac{c_k\boldsymbol{\Delta}_k^T\boldsymbol{g}(\hat{\theta}_k) + \Im(r_3(\hat{\theta}_k)) + \varepsilon_k^{\text{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)}{c_k\boldsymbol{\Delta}_k} \\
&= \frac{\boldsymbol{\Delta}_k^T\boldsymbol{g}(\hat{\theta}_k)}{\boldsymbol{\Delta}_k} + \frac{\Im(r_3(\hat{\theta}_k))}{c_k\boldsymbol{\Delta}_k} + \frac{\varepsilon_k^{\text{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)}{c_k\boldsymbol{\Delta}_k}.
\end{aligned} \tag{4.10}$$

Let $\hat{g}_{kj}^{\text{CS-SP}}(\hat{\theta}_k)$ and $g_j(\hat{\theta}_k)$ denote the $j$-th component of $\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)$ and $\boldsymbol{g}(\hat{\theta}_k)$, respectively. Applying the conditional expectation on both sides of (4.10), we have for all $j$ that

$$\mathbb{E}[\hat{g}_{kj}^{\text{CS-SP}}(\hat{\theta}_k)|\hat{\theta}_k] = \mathbb{E}\left[\frac{\boldsymbol{\Delta}_k^T\boldsymbol{g}(\hat{\theta}_k)}{\Delta_{kj}} + \frac{\Im(r_3(\hat{\theta}_k))}{c_k\Delta_{kj}} + \frac{\varepsilon_k^{\text{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)}{c_k\Delta_{kj}}\middle|\hat{\theta}_k\right]$$

$$= g_j(\hat{\theta}_k) + \sum_{m \neq j} \mathbb{E}\left[ g_m(\hat{\theta}_k) \frac{\Delta_{km}}{\Delta_{kj}} \middle| \hat{\theta}_k \right] + \mathbb{E}\left[ \frac{\Im(r_3(\hat{\theta}_k))}{c_k \Delta_{kj}} \middle| \hat{\theta}_k \right]$$
$$= g_j(\hat{\theta}_k) + \mathbb{E}\left[ \frac{\Im(r_3(\hat{\theta}_k))}{c_k \Delta_{kj}} \middle| \hat{\theta}_k \right], \tag{4.11}$$

where the second equality is due to Assumption 4.3 that

$$\mathbb{E}\left[ \frac{\varepsilon_k^{\mathrm{I}}(\hat{\theta}_k + i c_k \boldsymbol{\Delta}_k, \boldsymbol{v}_k)}{c_k \Delta_{kj}} \middle| \hat{\theta}_k \right] = \mathbb{E}\left[ \frac{\mathbb{E}[\varepsilon_k^{\mathrm{I}}(\hat{\theta}_k + i c_k \boldsymbol{\Delta}_k, \boldsymbol{v}_k) | \hat{\theta}_k, \boldsymbol{\Delta}_k]}{c_k \Delta_{kj}} \middle| \hat{\theta}_k \right] = 0,$$

and the last equality in (4.11) is due to

$$\mathbb{E}\left[ g_m(\hat{\theta}_k) \frac{\Delta_{km}}{\Delta_{kj}} \middle| \hat{\theta}_k \right] = g_m(\hat{\theta}_k) \mathbb{E}[\Delta_{km}] \mathbb{E}\left[ \frac{1}{\Delta_{kj}} \right] = 0$$

for any $m \neq j$. For the last term in (4.11), Cauchy's integral formula and the assumption $|\Delta_{km}| \leq \kappa_1$ implies that $|r_3(\hat{\theta}_k)| = O(c_k^3)$ and consequently $|\Im(r_3(\hat{\theta}_k))| = O(c_k^3)$. Hence, using Holder's inequality and Assumption 4.2, there exists some positive constant $\delta_1$ that

$$\left| \mathbb{E}\left[ \frac{\Im(r_3(\hat{\theta}_k))}{c_k \Delta_{kj}} \middle| \hat{\theta}_k \right] \right| \leq \frac{1}{c_k} \mathbb{E}\left[ |\Im(r_3(\hat{\theta}_k))|^{\frac{\delta_1+1}{\delta_1}} \middle| \hat{\theta}_k \right]^{\frac{\delta_1}{\delta_1+1}} \mathbb{E}\left[ \frac{1}{|\Delta_{kj}|^{\delta_1+1}} \right]^{\frac{1}{\delta_1+1}} = O(c_k^2).$$

Therefore, we conclude that $\boldsymbol{b}_k(\hat{\theta}_k) = O(c_k^2)$. □

**Remark 4.4.** *In (4.9), we see that there is no $L(\hat{\theta}_k)$ and $L^{(2)}(\hat{\theta}_k)$ terms since those terms are associated with the real-valued terms in the Taylor expansion,*

*which become exactly* $0$ *after applying the* $\Im(\cdot)$ *operation. This is similar to the two-sided SP gradient approximation, where the contributions of the* $L(\hat{\theta}_k)$ *and* $L^{(2)}(\hat{\theta}_k)$ *terms are also exactly* $0$ *due to the cancellation effects. On the contrary, the one-measurement SP gradient approximation contains* $L(\hat{\theta}_k)$ *in a mean-zero term, which is not exactly* $0$ *and causes the gradient approximation to have a relatively large variance. A more detailed analysis on the variance and mean squared error of FD and SP gradient approximations is available in Blakney and Zhu (2019).*

Theorem 4.1 indicates that as $k \to \infty$, the gradient approximation $\hat{g}_k^{\text{CS-SP}}(\hat{\theta}_k)$ is an asymptotically unbiased approximation of the true gradient $g(\hat{\theta}_k)$. Based on this result, we present the almost sure convergence results of $\hat{\theta}_k$ towards the optimal point $\theta^*$ defined in Assumption 4.7 below.

**Assumption 4.4** (Step-size Sequences)**.** *For all* $k$, $a_k > 0, c_k > 0$ *and* $a_k \to 0, c_k \to 0$ *as* $k \to \infty$*;*

$$\sum_{k=0}^{\infty} a_k = \infty \text{ and } \sum_{k=0}^{\infty} \frac{a_k^2}{c_k^2} < \infty. \tag{4.12}$$

**Assumption 4.5** (Bounded Variances)**.** *For all* $k$ *and* $j$*, assume that* $\mathbb{E}[\Im(L(\hat{\theta}_k + ic_k\Delta_k))^2|\Delta_k] < \sigma_L^2$*,* $\mathbb{E}[\varepsilon_k^I(\hat{\theta}_k + ic_k\Delta_k, v_k)^2|\Delta_k] < \sigma_\varepsilon^2$ *and* $\mathbb{E}[1/\Delta_{kj}^2] \leq \kappa_3$ *for some positive constants* $\sigma_L^2, \sigma_\varepsilon^2$ *and* $\kappa_3$*.*

**Assumption 4.6** (Iterate Boundedness)**.** $\sup_k \|\hat{\theta}_k\| < \infty$ *a.s.*

**Assumption 4.7** (Stable Solution)**.** *Let* $\theta^*$ *be an asymptotically stable solution of the differential equation* $d(\zeta(t))/dt = -\boldsymbol{g}(\zeta)$.

**Assumption 4.8** (Sample Path)**.** *Let* $D(\theta^*) = \{\zeta_0 : \lim_{t\to\infty} \zeta(t|\zeta_0) = \theta^*\}$ *where* $\zeta(t|\zeta_0)$ *denote the solution to the differential equation in Assumption 4.7. Assume that there exists a compact* $S \subseteq D(\theta^*)$ *such that* $\hat{\theta}_k \in S$ *infinitely often for almost all same paths.*

**Remark 4.5.** *All the assumptions here follow Spall (1992, Proposiiton 1), except for the Assumption 4.5. Minor modifications are made to accommodate* $\Im(L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k))$ *and* $\varepsilon_k^I(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)$ *instead of* $L(\hat{\theta}_k \pm c_k\boldsymbol{\Delta}_k)$ *and* $\varepsilon(\hat{\theta}_k \pm c_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k^{\pm})$ *as in the standard SP gradient approximation. Assumption 4.7 also implies that* $\theta^* = \arg\min_{\theta\in\mathbb{R}^p} L(\theta)$.

**Theorem 4.2** (Wang and Spall, 2021)**.** *Under Assumptions 4.1–4.8, we have as* $k \to \infty$*,*

$$\hat{\theta}_k \to \theta^* \text{ a.s.} \tag{4.13}$$

*Proof.* First define the error term $\boldsymbol{e}_k(\hat{\theta}_k)$ as

$$\boldsymbol{e}_k(\hat{\theta}_k) = \hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k) - \mathbb{E}[\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)|\hat{\theta}_k]. \tag{4.14}$$

Then we can rewrite (4.6) using (4.7) and (4.14) as

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k[\boldsymbol{g}(\hat{\theta}_k) + \boldsymbol{b}_k(\hat{\theta}_k) + \boldsymbol{e}_k(\hat{\theta}_k)], \tag{4.15}$$

which gives the standard form of a generalized Robbins-Monro algorithm. Following the proof of Spall (1992, Proposition 1), we proceed to show that

$$\|\boldsymbol{b}_k(\hat{\theta}_k)\| < \infty \text{ for all } k \text{ and } \boldsymbol{b}_k(\hat{\theta}_k) \to 0 \text{ a.s.}, \tag{4.16}$$

and

$$\lim_{k\to\infty} \mathbb{P}(\sup_{m\geq k} \| \sum_{j=k}^{m} a_j \boldsymbol{e}_j(\hat{\theta}_k) \geq \eta\|) = 0 \text{ for any } \eta > 0, \tag{4.17}$$

which are sufficient conditions for the almost sure convergence of $\hat{\theta}_k$ by Kushner and Clark (1978, Lemma 2.2.1 and Theorem 2.3.1).

Using the result from Theorem 4.1 that $\boldsymbol{b}_k(\hat{\theta}_k) = O(c_k^2)$ and Assumption 4.4, it is easy to see that (4.16) holds. To show (4.17), recall that $\mathscr{F}_k = \{\hat{\theta}_0, \ldots, \hat{\theta}_k\}$ and denote another filtration $\mathscr{G}_k = \{\boldsymbol{\Delta}_0, \ldots, \boldsymbol{\Delta}_k\}$ for all $k$. Then note that $\{\sum_{j=k}^{m} a_j \boldsymbol{e}_j(\hat{\theta}_j)\}_{m\geq k}$ is a martingale sequence since $\mathbb{E}[\boldsymbol{e}_k(\hat{\theta}_k)|\hat{\theta}_k] = \mathbb{E}[\boldsymbol{e}_k(\hat{\theta}_k)|\mathscr{F}_k] = 0$. Hence, Doob's martingale inequality implies that for any $\eta > 0$,

$$\mathbb{P}(\sup_{m\geq k} \| \sum_{j=k}^{m} a_j \boldsymbol{e}_j(\hat{\theta}_k) \geq \eta\|) \leq \frac{1}{\eta^2}\mathbb{E}[\| \sum_{j=k}^{\infty} a_j \boldsymbol{e}_j(\hat{\theta}_j)\|^2] = \frac{1}{\eta^2} \sum_{j=k}^{\infty} a_j^2 \mathbb{E}[\|\boldsymbol{e}_j(\hat{\theta}_j)\|^2], \tag{4.18}$$

where the last equality holds since for any $j < m$,

$$\begin{aligned}\mathbb{E}[\boldsymbol{e}_j(\hat{\theta}_j)^T \boldsymbol{e}_m(\hat{\theta}_m)] &= \mathbb{E}[\mathbb{E}[\boldsymbol{e}_j(\hat{\theta}_j)^T \boldsymbol{e}_m(\hat{\theta}_m)|\mathscr{F}_m, \mathscr{G}_{m-1}]] \\ &= \mathbb{E}[\boldsymbol{e}_j(\hat{\theta}_j)^T \mathbb{E}[\boldsymbol{e}_m(\hat{\theta}_m)|\mathscr{F}_m]]\end{aligned}$$

$$= 0.$$

Furthermore, noting that

$$\mathbb{E}[\boldsymbol{e}_j(\hat{\theta}_j)^T \mathbb{E}[\hat{\boldsymbol{g}}_j^{\text{CS-SP}}(\hat{\theta}_j)|\hat{\theta}_j]] = \mathbb{E}[\mathbb{E}[\boldsymbol{e}_j(\hat{\theta}_j)|\hat{\theta}_j]^T \mathbb{E}[\hat{\boldsymbol{g}}_j^{\text{CS-SP}}(\hat{\theta}_j)|\hat{\theta}_j]] = 0,$$

we have the relationship

$$\mathbb{E}[\|\hat{\boldsymbol{g}}_j^{\text{CS-SP}}(\hat{\theta}_j)\|^2] = \mathbb{E}[\|\boldsymbol{e}_j(\hat{\theta}_j)\|^2] + \mathbb{E}[\|\mathbb{E}[\hat{\boldsymbol{g}}_j^{\text{CS-SP}}(\hat{\theta}_j)|\hat{\theta}_j]\|^2]$$

for all $j$. Hence, we can bound the right-hand side of (4.18) as

$$\frac{1}{\eta^2}\sum_{j=k}^{\infty} a_j^2 \mathbb{E}[\|\boldsymbol{e}_j(\hat{\theta}_j)\|^2] \leq \frac{1}{\eta^2}\sum_{j=k}^{\infty} a_j^2 \mathbb{E}[\|\hat{\boldsymbol{g}}_j^{\text{CS-SP}}(\hat{\theta}_j)\|^2],$$

where the $m$-th component of $\hat{\boldsymbol{g}}_j^{\text{CS-SP}}(\hat{\theta}_j)$ can be bounded using Assumption 4.5 as

$$\begin{aligned}
\mathbb{E}[\hat{g}_{jm}^{\text{CS-SP}}(\hat{\theta}_k)^2] &= \mathbb{E}\Big[\frac{[\Im(L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k)) + \varepsilon_k^{\text{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)]^2}{c_k^2\Delta_{km}^2}\Big] \\
&\leq \frac{2}{c_k^2}\mathbb{E}\Big[\frac{\mathbb{E}[\Im(L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k))^2|\boldsymbol{\Delta}_k] + \mathbb{E}[\varepsilon_k^{\text{I}}(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)^2|\boldsymbol{\Delta}_k]}{\Delta_{km}^2}\Big] \\
&\leq \frac{2(\sigma_L^2 + \sigma_\varepsilon^2)\kappa_3}{c_k^2} < \infty.
\end{aligned}$$

Therefore, we conclude that (4.18) is bounded since

$$\mathbb{P}(\sup_{m\geq k} \| \sum_{j=k}^{m} a_j \boldsymbol{e}_j(\hat{\theta}_k) \geq \eta \|) \leq \min \left\{ 1, \frac{2p(\sigma_L^2 + \sigma_\varepsilon^2)\kappa_3}{\eta^2} \sum_{j=k}^{\infty} \frac{a_j^2}{c_j^2} \right\}.$$

The boundedness of the second term in $\{\cdot\}$ above implies that (4.17) holds for any $\eta > 0$. □

### 4.2.3 Asymptotic Distribution

Theorem 4.3 below gives the formal asymptotic normality results of $\hat{\theta}_k$ under the standard gain and perturbation sequences.

**Assumption 4.9** (Standard Step-size Sequences)**.** *For all $k$, assume the gain step size has the following standard form*

$$a_k = \frac{a}{(k+1+A)^\alpha} \text{ and } c_k = \frac{c}{(k+1)^\gamma},$$

*where $a, \alpha, c, \gamma > 0$, $A \geq 0$, and $2\gamma < \alpha < 4\gamma$.*

**Assumption 4.10** (Noisy Gradient)**.** *For all $k$ and given $\boldsymbol{v}_k$, assume that $\ell(\cdot, \boldsymbol{v}_k)$ can be extended to a complex analytic function. Denote $\boldsymbol{G}(\hat{\theta}_k, \boldsymbol{v}_k) = [\partial \ell(\theta, \boldsymbol{v}_k)/\partial \theta]_{\theta=\hat{\theta}_k}$ with its $j$-th component being $G_j(\hat{\theta}_k, \boldsymbol{v}_k)$ for all $j$. Further assume that there exists some random variable $\boldsymbol{v}^*$ such that $\boldsymbol{v}_k \to \boldsymbol{v}^*$ a.s. and $\mathbb{E}[\boldsymbol{G}(\hat{\theta}_k, \boldsymbol{v}_k)\boldsymbol{G}(\hat{\theta}_k, \boldsymbol{v}_k)^T | \mathscr{F}_k]$ $\to \mathbb{E}[\boldsymbol{G}(\theta^*, \boldsymbol{v}^*)\boldsymbol{G}(\theta^*, \boldsymbol{v}^*)^T]$ a.s. as $k \to \infty$.*

**Assumption 4.11** (Bounded Variances)**.** *For all $k$ and $j$, assume that there exists some positive constant $\delta_2 > 0$ such that $\mathbb{E}[\Im(L(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k))^{2+\delta}|\boldsymbol{\Delta}_k] < \sigma_L^2$, $\mathbb{E}[\varepsilon_k^I(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)^{2+\delta}|\boldsymbol{\Delta}_k] < \sigma_\varepsilon^2$ and $\mathbb{E}[1/|\Delta_{kj}|^{2+\delta}] \leq \kappa_2$ for some positive constants $\sigma_L^2, \sigma_\varepsilon^2$ and $\kappa_2$.*

**Remark 4.6.** *Assumption 4.9 is the gain and perturbation sequence assumption used in standard SPSA with common random numbers. Assumption 4.10 is identical to the assumptions in Kleinman et al. (1999, Theorem 2.1), which considers the asymptotic normality of $\{\hat{\theta}_k\}$ generated by SPSA with CRNs (common random numbers), i.e., $\boldsymbol{v}_k^+ = \boldsymbol{v}_k^-$ for all $k$. Assumption 4.11 strengthens Assumption 4.5 used in Theorem 4.2 and is similar to Spall (1992, Assumption A2′) with the extension to incorporate complex variables.*

**Theorem 4.3.** *Assume Assumption 4.1–4.11 hold. Then as $k \to \infty$,*

$$k^{\alpha/2}(\hat{\theta}_k - \theta^*) \xrightarrow{dist} \mathcal{N}(0, \boldsymbol{P}\boldsymbol{M}\boldsymbol{P}^T)\,, \tag{4.19}$$

*where $\boldsymbol{P}$ is an orthogonal matrix satisfying $\boldsymbol{P}^T\boldsymbol{H}(\theta^*)\boldsymbol{P} = a^{-1}\operatorname{diag}(\lambda_1, \ldots, \lambda_p)$, the $mn$-th entry of $\boldsymbol{M}$ is*

$$M_{mn} = a^2[\boldsymbol{P}^T\boldsymbol{\Sigma}\boldsymbol{P}]_{mn}\frac{1}{\lambda_m + \lambda_n - \alpha_+}$$

*with the $mn$-th entry of $\boldsymbol{\Sigma}$ being*

$$\Sigma_{mn} = \begin{cases} \mathbb{E}[G_m(\theta^*, \boldsymbol{v}^*)^2] + \sum_{j \neq m} \sigma_\Delta^2 \tau_\Delta^2 \mathbb{E}[G_j(\theta^*, \boldsymbol{v}^*)^2] & \textit{if } m = n, \\ 2\mathbb{E}[G_m(\theta^*, \boldsymbol{v}^*) G_n(\theta^*, \boldsymbol{v}^*)] & \textit{if } m \neq n, \end{cases}$$

$\mathbb{E}[\Delta_{kj}^2] = \sigma_\Delta^2$, $\mathbb{E}[1/\Delta_{kj}^2] = \tau_\Delta^2$ *for all* $j$ *and* $k$, *and* $\alpha_+ = \alpha < 2 \min_i \lambda_i$ *if* $\alpha = 1$ *and* $\alpha_+ = 0$ *if* $\alpha < 1$.

**Remark 4.7.** *The rate of convergence in (4.19) can achieve an accelerated rate of $k^{-1/2}$ with $\alpha = 1$, which is faster than the max allowable decay rate of $k^{-1/3}$ possible in standard SPSA under typical convergence conditions found using $\alpha = 1$ and $\gamma = 1/6$. It is worth mentioning that an optimal rate of $k^{-1/2}$ has only been known to be achievable only for SPSA with pure common random numbers (CRNs) according to Kleinman et al. (1999, Theorem 2.1). In fact, the covariance matrix of the asymptotic normality in (4.19) is identical to Kleinman et al. (1999, Theorem 2.1). However, using CRNs often requires the user to have the synchronization of the random number streams used in $\boldsymbol{v}_k^{(+)}$ and $\boldsymbol{v}_k^{(-)}$.*

*Proof.* Following Spall (1992, Proposition 2) and Kleinman et al. (1999, Theorem 2.1), this proof relies on verifying the conditions (2.2.1) – (2.2.3) in Fabian (1968). By Assumption 4.1 and Theorem 4.2, there exists an open neighborhood of $\hat{\theta}_k$ in the real space, denoted as $\mathcal{B}(\hat{\theta}_k)$, such that $\boldsymbol{H}(\theta)$ is continuous for all $\theta \in \mathcal{B}(\hat{\theta}_k)$, and for sufficiently large $k$, we have $\theta^* \in \mathcal{B}(\hat{\theta}_k)$. Therefore, ap-

plying Taylor expansion on $\boldsymbol{g}(\hat{\theta}_k)$ around $\theta^*$ and using the fact that $\boldsymbol{g}(\theta^*) = 0$, we have

$$\boldsymbol{g}(\hat{\theta}_k) = \overline{\boldsymbol{H}}_k(\hat{\theta}_k - \theta^*), \tag{4.20}$$

where $\overline{\boldsymbol{H}}_k$ is a matrix with its each row being the corresponding row of $\boldsymbol{H}(\bar{\theta}_k)$ evaluated at a different $\bar{\theta}_k$ that is on the line segment of $\hat{\theta}_k$ and $\theta^*$. Plugging (4.20) into (4.15) and following the notations in Fabian (1968), we rewrite (4.15) as

$$\begin{aligned}
&\hat{\theta}_{k+1} - \theta^* \\
&= \hat{\theta}_k - \theta^* - a_k\overline{\boldsymbol{H}}_k(\hat{\theta}_k - \theta^*) - a_k\boldsymbol{b}_k(\hat{\theta}_k) - a_k\boldsymbol{e}_k(\hat{\theta}_k) \\
&= [\boldsymbol{I} - (k+1+A)^{-\alpha}\boldsymbol{\Gamma}_k](\hat{\theta}_k - \theta^*) + (k+1+A)^{-3\alpha/2}\boldsymbol{T}_k + (k+1+A)^{-\alpha}\boldsymbol{\Phi}_k\boldsymbol{V}_k,
\end{aligned}$$

where $\boldsymbol{\Gamma}_k = a\overline{\boldsymbol{H}}_k, \boldsymbol{T}_k = -a(k+1+A)^{\alpha/2}\boldsymbol{b}_k(\hat{\theta}_k), \boldsymbol{\Phi}_k = -a\boldsymbol{I}$ and $\boldsymbol{V}_k = \boldsymbol{e}_k(\hat{\theta}_k)$. We now proceed to verify Fabian (1968, Conditions 2.2.1 and 2.2.2) by showing $\boldsymbol{\Gamma}_k, \boldsymbol{T}_k$ and $\mathbb{E}[\boldsymbol{V}_k\boldsymbol{V}_k^T|\mathcal{F}_k]$ all converge almost surely.

For the term $\boldsymbol{\Gamma}_k$, by the continuity of $\boldsymbol{H}(\cdot)$ in Assumption 4.1 and the almost sure convergence of $\hat{\theta}_k \to \theta^*$ in Theorem 4.2, it is easy to see that $\overline{\boldsymbol{H}}_k \to \boldsymbol{H}(\theta^*)$ almost surely and that $\boldsymbol{\Gamma}_k \to a\boldsymbol{H}(\theta^*)$. Furthermore, since Theorem 4.1 implies that $\boldsymbol{T}_k = O(k^{\alpha/2-2\gamma})$ and $\alpha - 4\gamma < 0$ in Assumption 4.9, we conclude that $\boldsymbol{T}_k \to 0$ almost surely and hence Fabian (1968, Condition 2.2.1) is satisfied.

For the last term $\mathbb{E}[\boldsymbol{V}_k\boldsymbol{V}_k^T|\mathcal{F}_k]$, we have

$$
\begin{aligned}
&\mathbb{E}[\boldsymbol{V}_k\boldsymbol{V}_k^T|\mathscr{F}_k] \\
&= \mathbb{E}[\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)^T|\mathscr{F}_k] - \mathbb{E}[\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)|\mathscr{F}_k]\mathbb{E}[\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)|\mathscr{F}_k]^T \\
&= \frac{1}{c_k^2}\mathbb{E}[\boldsymbol{\Delta}_k^{-1}(\boldsymbol{\Delta}_k^{-1})^T\Im(\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k))^2|\mathscr{F}_k] - [\boldsymbol{b}_k(\hat{\theta}_k) + \boldsymbol{g}(\hat{\theta}_k)][\boldsymbol{b}_k(\hat{\theta}_k) + \boldsymbol{g}(\hat{\theta}_k)]^T \\
&= \frac{1}{c_k^2}\mathbb{E}[\boldsymbol{\Delta}_k^{-1}(\boldsymbol{\Delta}_k^{-1})^T\mathbb{E}[\Im(\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k))^2|\boldsymbol{\Delta}_k, \mathscr{F}_k]|\mathscr{F}_k] \\
&\quad - [\boldsymbol{b}_k(\hat{\theta}_k) + \boldsymbol{g}(\hat{\theta}_k)][\boldsymbol{b}_k(\hat{\theta}_k) + \boldsymbol{g}(\hat{\theta}_k)]^T. \qquad (4.21)
\end{aligned}
$$

Let us proceed to show that all the terms on the right-hand side of (4.21) converge appropriately. First note that the last term of (4.21) converges to $0$ since Theorem 4.1 and 4.2 imply $\boldsymbol{b}_k(\hat{\theta}_k) = O(k^{-2\gamma})$ and $\boldsymbol{g}(\hat{\theta}_k) \to \boldsymbol{g}(\theta^*) = \boldsymbol{0}$ almost surely. Then, for the first term on the right-hand side of (4.21), similar to how (4.9) is derived, we apply Taylor expansion on $\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k)$ around $\hat{\theta}_k$ and obtain

$$
\begin{aligned}
&\mathbb{E}[\Im(\ell(\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k))^2|\mathscr{F}_k, \boldsymbol{\Delta}_k] \\
&= \mathbb{E}\{[c_k\boldsymbol{\Delta}_k^T\boldsymbol{G}(\hat{\theta}_k, \boldsymbol{v}_k) + \Im(R_3(\hat{\theta}_k, \boldsymbol{v}_k))]^2\} \\
&= c_k^2\mathbb{E}\{[\boldsymbol{\Delta}_k^T\boldsymbol{G}(\hat{\theta}_k, \boldsymbol{v}_k)]^2|\boldsymbol{\Delta}_k, \mathscr{F}_k\} + O(c_k^4),
\end{aligned}
$$

where $R_3(\hat{\theta}_k, \boldsymbol{v}_k)$ is the higher-order remainder term of the Taylor expansion similar to (4.8) and $|R_3(\hat{\theta}_k, \boldsymbol{v}_k)| = O(c_k^3)$. Hence, for the $mn$-th entry of $\mathbb{E}[\boldsymbol{V}_k\boldsymbol{V}_k^T|\mathscr{F}_k]$,

we have

$$\mathbb{E}[\boldsymbol{V}_k\boldsymbol{V}_k^T|\mathscr{F}_k]_{mn} = \sum_{j=1}^{p}\sum_{l=1}^{p}\mathbb{E}\left[\frac{\Delta_{kj}\Delta_{kl}}{\Delta_{km}\Delta_{kn}}\right]\mathbb{E}[G_j(\hat{\boldsymbol{\theta}}_k,\boldsymbol{v}_k)G_l(\hat{\boldsymbol{\theta}}_k,\boldsymbol{v}_k)|\mathscr{F}_k] + o(1),$$

where the $o(1)$ term represents all the terms discussed above that are converging to $0$. Using Assumption 4.2 that all $\Delta_{kj}$ are independent and symmetrically distribution about $0$, we see that when $m = n$

$$\mathbb{E}[\Delta_{kj}\Delta_{kl}/\Delta_{km}\Delta_{kn}] = \begin{cases} 1 & \text{if } j = l = m \\ \sigma_\Delta^2\tau_\Delta^2 & \text{if } j = l \neq m \\ 0 & \text{otherwise}, \end{cases}$$

and when $m \neq n$

$$\mathbb{E}[\Delta_{kj}\Delta_{kl}/\Delta_{km}\Delta_{kn}] = \begin{cases} 1 & \text{if } j = m, l = n \text{ or } j = n, l = m, \\ 0 & \text{otherwise}. \end{cases}$$

Therefore, for all $m$ and $n$,

$$\begin{aligned} &\mathbb{E}[\boldsymbol{V}_k\boldsymbol{V}_k^T|\mathscr{F}_k]_{mn} \\ &= \begin{cases} \mathbb{E}[G_m(\hat{\boldsymbol{\theta}}_k,\boldsymbol{v}_k)^2|\mathscr{F}_k] + \sum_{j\neq m}\sigma_\Delta^2\tau_\Delta^2\mathbb{E}[G_j(\hat{\boldsymbol{\theta}}_k,\boldsymbol{v}_k)^2|\mathscr{F}_k] + o(1) & \text{if } m = n, \\ 2\mathbb{E}[G_m(\hat{\boldsymbol{\theta}}_k,\boldsymbol{v}_k)G_n(\hat{\boldsymbol{\theta}}_k,\boldsymbol{v}_k)|\mathscr{F}_k] + o(1) & \text{if } m \neq n. \end{cases} \end{aligned}$$

By Assumption 4.10, we have $\mathbb{E}[\boldsymbol{G}(\hat{\theta}_k, \boldsymbol{v}_k)\boldsymbol{G}(\hat{\theta}_k, \boldsymbol{v}_k)^T|\mathscr{F}_k] \to \mathbb{E}[\boldsymbol{G}(\theta^*, \boldsymbol{v}^*)\boldsymbol{G}(\theta^*, \boldsymbol{v}^*)^T]$ a.s. as $k \to \infty$, which concludes that Fabian (1968, Condition 2.2.2) is satisfied.

Finally, to verify that Fabian (1968, Condition 2.2.3) holds, using both Markov inequality and Holder's inequality, we have for any $0 < \delta' < \delta/2$ and $r > 0$

$$\lim_{k\to\infty} \mathbb{E}[\mathbb{1}_{\{\|\boldsymbol{V}_k\|^2 \geq rk^\alpha\}}\|\boldsymbol{V}_k\|^2] \leq \limsup_{k\to\infty} \left(\frac{\mathbb{E}[\|\boldsymbol{V}_k\|^2]}{rk^\alpha}\right)^{\frac{\delta'}{1+\delta'}} (\mathbb{E}[\|\boldsymbol{V}_k\|^{2(1+\delta')}])^{\frac{1}{1+\delta'}},$$

where $\delta$ is defined in Assumption 4.5. Following the arguments in Kleinman et al. (1999, Theorem 2.1), we also have

$$\|\boldsymbol{V}_k\|^{2(1+\delta')} \leq 2^{2(1+\delta')}[\|\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)\|^{2(1+\delta')} + \|\boldsymbol{g}(\hat{\theta}_k)\|^{2(1+\delta')} + \|\boldsymbol{b}_k(\hat{\theta}_k)\|^{2(1+\delta')}],$$

where both $\boldsymbol{g}(\hat{\theta}_k)$ and $\boldsymbol{b}_k(\hat{\theta}_k)$ are uniformly bounded for large $k$ by Theorem 4.1. Using the similar arguments in Spall (1992, Proposition 2), we see that Assumption 4.11 and fact that $\delta' < \delta/2$ imply $\mathbb{E}[\|\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)\|^{2(1+\delta')}] = O(1)$. Hence, we conclude that $\mathbb{E}[\|\boldsymbol{V}_k\|^{2(1+\delta')}] = O(1)$, which shows $\lim_{k\to\infty} \mathbb{E}[\mathbb{1}_{\{\|\boldsymbol{V}_k\|^2 \geq rk^\alpha\}}\|\boldsymbol{V}_k\|^2] = 0$ and Fabian (1968, Condition 2.2.3) is verified. □

## 4.3 Model-free Control

This section proposes to use the CS-SPSA to construct model-free controllers for non-linear stochastic systems, where both the state and measurement equa-

tions are unknown. With no stringent model assumptions imposed on the underlying systems, the controllers are constructed by function approximators (FA), such as neural networks or polynomials. The optimal control refers to the optimal parameters within the FA. Since the functional form of the system remains unknown, we need to estimate the gradient to proceed with gradient-type algorithms. This work proposes a novel estimation for the gradient using only the system measurement, which is inspired by the simultaneous perturbation stochastic approximation (SPSA) algorithm and the complex-step gradient approximation. In contrast to prior work that requires at least two measurements per iteration, CS-SPSA requires only one system measurement per iteration, which makes it suitable in tracking transient systems. We also establish an almost sure convergence result for stochastic approximation algorithms with time-varying objective functions.

Complex dynamical systems are widely applied in various fields, and it is inevitable to manipulate the states (variables) through a set of control (variables). Optimal control is designed to locate the optimal controller in minimizing certain objective functions, such as sending a rocket to a target location with minimal fuel consumption or maneuvering robot arms to move certain items. When there is limited knowledge about the system, it is rather challenging to determine the dynamics of the control for a system. This scenario commonly arises in complex physical, biological, or climate systems (Boccara,

2010), where either the entire functional forms of the system equations are unknown or the system equations are too complex to compute any gradient information for optimizing the controller. As a result, we cannot apply the gradient-based algorithms (Lasdon et al., 1967; Khaneja et al., 2005; Luo et al., 2016), and resort to optimization schemes that use only the direct system outputs for optimization algorithms.

Since no information about the system equations is revealed in our optimization procedure, our approach is a *model-free* controller. By "model-free" we mean that no explicit model assumptions is known and our output relies on the system output only. This key characteristic contrasts with prior work in which the hidden or implicit system assumptions are required. For example, several fuzzy controllers (Mamdani, 1976; Buckley, 1992) have rule bases that describe the dynamics of the system in a linguistic-type fashion. As the controller depends on the past system information to generate certain actions, we have to construct a FA to determine the control variable. Note that some variables in the objective function, such as target measurement values, may also be included as the inputs of the controller FA. Popular choices of the FA are, for example, neural networks, polynomials, wavelet functions, or trigonometric series. According to the well-known Stone–Weierstrass approximation theorem (Rudin et al., 1964) and universal approximation theorem (Hornik et al., 1989; Csáji et al., 2001; Zhou, 2020), one can approximate any continuous function

arbitrary well with a certain class of functions. The advantages and disadvantages of various function approximations are summarized in Chen and Chen (1995); Lane et al. (1992) and Poggio and Girosi (1990). Here we consider using polynomial as our controller FA and it can be replaced by any other valid ones.

Given that only the noisy system measurements are accessible, we consider stochastic optimization algorithms to estimate the gradient. In particular, SPSA is a powerful technique to estimate function gradient using only *two* noisy function measurements, regardless of the dimension of the problem itself, which enables it to be applied in optimal control problems, especially when every function measurement requires updating the control. For FDSA that requires dimension-dependent measurements to estimate one gradient, the system state variable may already evolve to a completely different value. SPSA is first applied in Spall and Cristion (1998) to model-free optimal control problems. Later, Zhou et al. (2008) uses an SPSA-based model-free feedback controller in active noise control for periodic disturbances in a duct, Ahmad et al. (2014) analyzes the model-free proportional–integral–derivative tuning of multiple-input and multiple-output systems using SPSA, and Yuan (2008) constructs a model-free automatic tuning method for a restricted structured controller.

Although SPSA-based controller has been popular, it still requires two function measurements, which creates a small discrepancy in the gradient approx-

imation. The error mainly arises from the two function measurements depending on different state variable values. Consider the simple one-step-forward quadratic tracking error $L_k(\theta_k) = \mathbb{E}[(\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T \boldsymbol{A}_k (\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T + \boldsymbol{u}_k^T \boldsymbol{B}_k \boldsymbol{u}_k]$, where $L_k(\cdot)$ denotes the true loss function at time $k$, $\theta_k$ represents the parameter in the controller FA, $\boldsymbol{u}_k$ denotes the control variable, $\boldsymbol{y}_{k+1}$ is the system outputs, $\boldsymbol{t}_{k+1}$ is the target of the state variable, and $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are two given positive semi-definite weighting matrices. To estimate the gradient of $L_k(\cdot)$, SPSA-based methods require two consecutive noisy function measurements of $\ell_k^{(\pm)} = (\boldsymbol{y}_{k+1}^{(\pm)} - \boldsymbol{t}_{k+1})^T \boldsymbol{A}_k (\boldsymbol{y}_{k+1}^{(\pm)} - \boldsymbol{t}_{k+1})^T + \boldsymbol{u}_k^{(\pm)T} \boldsymbol{B}_k \boldsymbol{u}_k^{(\pm)}$ at two simultaneously perturbed parameter values $\theta_k^{(+)}$ and $\theta_k^{(-)}$. However, when $\boldsymbol{y}_{k+1}^{(+)}$ and $\boldsymbol{y}_{k+1}^{(-)}$ are far away from each other, especially in strong transient systems, the gradient approximation becomes less reliable and may cause the system to become unstable.

### 4.3.1 Algorithm Description

Denote $L_k : \mathbb{R}^p \to \mathbb{R}$ as the generic time-varying loss function at time $k$. Given the parameter of interest being $\theta \in \mathbb{R}^p$, we are interested in constructing a parameter estimate $\hat{\theta}_k \in \mathbb{R}^p$ at time $k$ such that true loss function evaluated at $L_k(\hat{\theta}_k)$ is minimized. Consider a general non-linear stochastic system,

$$\boldsymbol{x}_{k+1} = f_k(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{w}_k),$$

$$\boldsymbol{y}_k = h_k(\boldsymbol{x}_k, \boldsymbol{v}_k),$$

where, for the state equation, $\boldsymbol{x}_k$ represents the state of the system, $f_k(\cdot)$ is a linear or non-linear state function, $\boldsymbol{u}_k$ is the control variable, $\boldsymbol{w}_k$ is the state noise, and for the measurement equation, $\boldsymbol{y}_k$ is the observed quality of the state variable, $h_k(\cdot)$ is a linear or non-linear measurement function, and $\boldsymbol{v}_k$ s the measurement noise. A common loss in optimal control problems is the one-step-forward quadratic tracking error

$$L_k(\theta_k) = \mathbb{E}[(\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T \boldsymbol{A}_k (\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T + \boldsymbol{u}_k^T \boldsymbol{B}_k \boldsymbol{u}_k], \tag{4.22}$$

where the given matrices $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are often introduced to reflect the cost on the deviation of $\boldsymbol{y}_{k+1}$ from the target $\boldsymbol{t}_{k+1}$ and penalize for large value of $\boldsymbol{u}_k$, respectively. Note that different indices of $\boldsymbol{y}_{k+1}$ and $\boldsymbol{u}_k$ are used in the loss function (4.22). This is due to fact that the observation $\boldsymbol{y}_{k+1}$ is only available from the unknown state variable $\boldsymbol{x}_{k+1}$, which is generated by evolving the system with $\boldsymbol{x}_k$ and $\boldsymbol{u}_k$. Since the control variable $\boldsymbol{u}_k$ is a direct output of our model-free FA, which is parameterized by $\theta_k$, the true loss function at time $k$ is denoted as $L_k(\cdot)$ evaluated at $\theta_k$. Given the system output $\boldsymbol{y}_{k+1}$, a natural

choice of the noise loss function at time $k$ becomes

$$\ell_k(\theta_k) = (\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T \boldsymbol{A}_k (\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T + \boldsymbol{u}_k^T \boldsymbol{B}_k \boldsymbol{u}_k.$$

Therefore, with only evaluations of $\ell_k(\cdot)$, we seek to find the minimum of the loss function $L(\cdot)$.

Given the loss function $L_k$ in (4.22), denote its gradient as

$$\boldsymbol{g}_k(\theta_k) = \frac{\partial L_k}{\partial \theta_k} = \frac{\partial \boldsymbol{u}_k^T}{\partial \theta_k} \frac{\partial L_k}{\partial \boldsymbol{u}_k}.$$

Since the problem of minimizing a loss function can often be solved by using its gradient information, we seek to find the solution $\theta_k^*$ of the gradient equation $\boldsymbol{g}_k(\theta_k) = 0$. To solve the optimization problem, first consider the parameter estimate $\hat{\theta}_k$ at time $k$. In order to estimate the optimal parameter $\theta_{k+1}^*$ at the very next time step $k+1$, we use the standard gradient-descent-type algorithm as

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_{k+1}(\hat{\theta}_k)$$

where $a_k$ is the non-negative gain coefficient and $\hat{\boldsymbol{g}}_{k+1}(\cdot)$ is denoted as the gradient approximation of the true gradient $\boldsymbol{g}_{k+1}(\cdot)$. Note that different indices are used for $\hat{\boldsymbol{g}}_{k+1}(\cdot)$ and $\hat{\theta}_k$, which is to emphasize the gradient is of the loss function $L_{k+1}(\cdot)$, and we propose to use both the noise loss function $\ell_{k+1}(\cdot)$ and

the current parameter estimate $\hat{\theta}_k$ to construct the gradient approximation. Specifically, we have

$$\hat{\boldsymbol{g}}_{k+1}(\hat{\theta}_k) = \frac{\Im(\ell_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1}))}{c_{k+1}\boldsymbol{\Delta}_{k+1}}, \tag{4.23}$$

where $c_k$ is the non-negative perturbation step-size coefficient, $\boldsymbol{\Delta}_{k+1}$ is a $p$-dimensional random perturbation vector $\boldsymbol{\Delta}_{k+1} = [\Delta_{k+1,1}, \ldots, \Delta_{k+1,p}]^T$, and $\Im(\cdot)$ denotes the imaginary part of the complex argument. We use $\boldsymbol{\Delta}_{k+1}^{-1}$ or $1/\boldsymbol{\Delta}$ to denote $[\Delta_{k+1,1}^{-1}, \ldots, \Delta_{k+1,p}^{-1}]^T$, where each component of $\boldsymbol{\Delta}_{k+1}$ is independently and identically distributed satisfying some regularity conditions specified in Assumption 4.14 below.

It is worth noting that the gradient approximation in (4.23) requires a noisy function evaluation at the complex value since the perturbed parameter estimate $\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1} \in \mathbb{C}^p$. However, the gradient approximation $\hat{\boldsymbol{g}}_{k+1}(\hat{\theta}_k)$ is always real-valued since $\ell_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1}) \in \mathbb{C}$ and $\Im(\ell_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1})) \in \mathbb{R}$. With the real-valued $a_k, c_k$ and $\boldsymbol{\Delta}_k$ for all $k$, it is also guaranteed that $\hat{\theta}_k \in \mathbb{R}^p$ for all $k$.

To illustrate how the proposed gradient approximation (4.23) is used, we present the following step-by-step guide.

**Step 0 (Initialization):** Set index $k = 0$. Pick an initial parameter estimate $\hat{\theta}_0$ and the non-negative gain coefficient $a_k$ and perturbation step-size

coefficient $c_k$.

**Step 1 (Perturbation Vector):** Generate a $p$-dimensional random perturbation vector $\boldsymbol{\Delta}_{k+1}$, where all the components of $\boldsymbol{\Delta}_{k+1}$ are independently generated from mean-zero symmetric probability distributions satisfying the regularity conditions discussed in Assumption 4.14 below.

**Step 2 (Gradient Estimate):** Construct the gradient approximate $\hat{\boldsymbol{g}}_{k+1}(\hat{\theta}_k)$ according to (4.23).

**Step 3 (Iterative Update):** Update the parameter estimate using the standard gradient-descent-type algorithm as

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_{k+1}(\hat{\theta}_k)\,,$$

**Step 4 (Iteration or Termination):** Return to Step 1 with $k+1$ replacing $k$. Terminate the algorithm if the maximum allowable number of iterations has been reached.

### 4.3.2 Convergence

This second provides the assumptions of the almost sure convergence of $\hat{\theta}_k$ and the formal convergence theorem.

**Assumption 4.12** (Step-size Sequences)**.** *For all $k$, assume $a_k > 0, c_k > 0$ and*

$a_k \to 0, c_k \to 0$ ***as*** $k \to \infty$***;***

$$\sum_{k=0}^{\infty} a_k = \infty, \sum_{k=0}^{\infty} \frac{a_k^2}{c_k^2} < \infty \textit{ and } \sum_{k=0}^{\infty} a_k c_k^2 < \infty.$$

**Assumption 4.13** (Loss Function)**.** *Assume that $L_k(\cdot)$ can be extended to a complex analytic function and $L_k(\theta) \in \mathbb{R}^p$ whenever $\theta \in \mathbb{R}^p$. Further assume that, for almost all $\hat{\theta}_k$, there exists an open neighborhood of $\hat{\theta}_k$ in the complex space such that for any $\theta$ in that neighborhood, $|f(\theta)| \leq B_f$ for some positive constant $B_f$.*

**Assumption 4.14** (Perturbation Vector)**.** *Far all $k$ and $j$, assume that $\Delta_{kj}$ are independent and identically distributed, symmetrically distributed about 0. Further assume that $|\Delta_{kj}| \leq \kappa_0$ for some positive constant $\kappa_0$ and there exists some positive $\delta$ such that $\mathbb{E}[1/|\Delta_{kj}|^{2+\delta}] < \kappa_1$ for some positive constant $\kappa_1$.*

**Assumption 4.15** (Noisy Measurement)**.** *Denote the noise term as $\varepsilon_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1}) = \ell_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1}) - L_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1})$. For all $k$, assume that $\mathbb{E}[\varepsilon_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1})|\hat{\theta}_k, \boldsymbol{\Delta}_{k+1}] = 0$. Further assume that there exists some positive constant $\delta$ such that $\mathbb{E}[\Im(L_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1}))^{2+\delta}|\boldsymbol{\Delta}_{k+1}] < \sigma_L^2$ and $\mathbb{E}[\Im(\varepsilon_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1}))^{2+\delta}|\boldsymbol{\Delta}_{k+1}] < \sigma_\varepsilon^2$ for some positive constants $\sigma_L^2$ and $\sigma_\varepsilon^2$.*

**Assumption 4.16** (Strong Convexity)**.** *For some $K < \infty$, assume that there exists some positive constant $\rho > 0$ such that, for all $k \geq K$, we have $(\theta - \theta^*)^T \boldsymbol{g}_{k+1}(\theta) \geq \delta_k(\rho)$ with some constant $\delta_k(\rho)$ satisfying $\sum_{k=0}^{\infty} a_k \delta_k(\rho) = \infty$ when-*

*ever* $\|\theta - \theta^*\|_2^2 \geq \rho$.

**Assumption 4.17** (No Bouncing Around)**.** *Denote* $\bar{\boldsymbol{g}}_{k+1}(\hat{\theta}_k) = \mathbb{E}[\hat{\boldsymbol{g}}_{k+1}(\hat{\theta}_k)|\hat{\theta}_k]$. *For any* $j = 1, \ldots, p$ *and* $\rho > 0$*, assume the conditional probability* $P(\{\bar{g}_{k+1,j}(\hat{\theta}_k) \geq 0$ *i.o.*$\} \cap \{\bar{g}_{k+1,j}(\hat{\theta}_k) < 0$ *i.o.*$\}|\{|\hat{\theta}_{kj} - [\theta^*]_j| \geq \rho$ *for all* $k\}) = 0$.

**Assumption 4.18** (Non-negligible Contribution)**.** *For any* $\tau < 0$ *and non-empty* $S \subset \{1, 2, \ldots, p\}$*, there exists a* $\lambda > \tau$ *such that*

$$\limsup_{k \to \infty} \left| \frac{\sum_{j \notin S} (\theta - \theta^*) g_{k+1,j}(\theta)}{\sum_{j \in S} (\theta - \theta^*) g_{k+1,j}(\theta)} \right| < 1 \text{ a.s.} \tag{4.24}$$

*for all* $|[\theta]_j - [\theta^*]_j| < \tau$ *where* $j \notin S$ *and* $|[\theta]_j - [\theta^*]_j| \geq \lambda$ *when* $j \in S$.

Before presenting the main almost sure convergence theorem of $\hat{\theta}_k$, let us first discussion the assumptions here and show how they are related to the assumptions proposed in the previous SP-based stochastic optimization literature. Assumption 4.12 is a standard gain and perturbation step-size sequence condition for SP-based stochastic optimization algorithm (Spall, 1992; Spall and Cristion, 1998; Spall, 2005). Although the condition $\sum_{k=0}^{\infty} a_k c_k^2 < \infty$ is not need in the original SPSA, it is required here to bound the bias of the gradient approximation. This condition also appears in various stochastic optimization algorithms (see, e.g., Pflug, 2012, Theorem 5.3 and Kushner and Clark, 1978, Theorem 5.2.1). The requirement of complex analytic function in Assumption 4.13 is a much stricter assumption since it automatically im-

plies the function to be infinitely differentiable. However, it is necessary here to allow an appropriate Taylor expansion of $L_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1})$, which then guarantees the accuracy of the SP gradient approximation. Nonetheless, the complex analytic function assumption is still a common assumption for using the complex-step gradient approximation (see, e.g., Martins et al., 2003; Higham, 2018; Nikolovski and Stojkovska, 2018). Assumption 4.14 on the perturbation vector $\boldsymbol{\Delta}_k$ is similar to the one used in the basic SPSA algorithm, which can be easily satisfied by choosing $\boldsymbol{\Delta}_{kj}$ to be independently Bernoulli $\pm 1$ distributed with equal probabilities for all $k$ and $j$. Assumption 4.15 is also a standard assumption on the noise of the function measurements, which prevents the variability of $\hat{\boldsymbol{g}}_{k+1}(\hat{\theta}_k)$ to be too large and ensures the stability of the algorithm. Similar assumptions with $L_{k+1}(\hat{\theta}_k + c_{k+1}\boldsymbol{\Delta}_{k+1})$ replacing $\Im(L_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1}))$ and $\varepsilon_{k+1}(\hat{\theta}_k + c_{k+1}\boldsymbol{\Delta}_{k+1})$ replacing $\Im(\varepsilon_{k+1}(\hat{\theta}_k + ic_{k+1}\boldsymbol{\Delta}_{k+1}))$ can be found in Spall (1992). Assumption 4.16 requires the loss function to be strongly convex if we choose $\delta_k(\rho) = \rho\|\theta - \theta^*\|_2^2$. This assumption provides a key improvement from previous work since it is a direct assumption of the loss function itself. In comparison, the assumption in Spall and Cristion (1998) requires $(\theta - \theta^*)^T \bar{\boldsymbol{g}}_{k+1}(\theta) \geq \delta_k(\rho)$ with $\bar{\boldsymbol{g}}_{k+1}(\theta) = \mathbb{E}[\hat{\boldsymbol{g}}_{k+1}(\hat{\theta}_k)|\hat{\theta}_k]$. The involvement of $\bar{\boldsymbol{g}}_k(\theta)$, however, makes the assumption depends on the estimate of the algorithm and it is much harder to be examined for practical problems. As stated in Spall and Cristion (1998), the assumption involving $\bar{\boldsymbol{g}}_{k+1}(\theta)$ may also

be violated when there is a strong system transition between two consecutive noisy function measurements. In our proposed algorithm, however, since only one noisy function measurement is collected at each iteration, there is no such concern as long as the function is strongly convex. Assumptions 4.17 and 4.18 are both weak technical conditions to ensure the almost convergence of $\hat{\theta}_k$. Generally speaking, Assumption 4.17 prevents $\hat{\theta}_k$ from bouncing around $\theta^*$ that causes the elements of $\bar{\boldsymbol{g}}_{k+1}(\hat{\theta}_k)$ to change sign infinitely often. Assumption 4.18 ensures that no element of $\boldsymbol{g}(\theta)$ makes a negligible contribution to $(\theta - \theta^*)^T \boldsymbol{g}_{k+1}(\theta)$ whenever $[\theta]_j - [\theta^*]_j \neq 0$ for any $j$. A common sufficient condition is to let $\boldsymbol{g}_{k+1}(\theta)$ to be uniformly bounded between $0$ and $\infty$ whenever $[\theta]_j - [\theta^*]_j \neq 0$ for all $j$.

**Theorem 4.4** (Wang et al., 2021)**.** *Let Assumptions 4.12–4.18 hold and there exists a unique* $\theta^*$ *such that* $thetabmstar_k \to \theta^*$ *as* $k \to \infty$*. We have*

$$\hat{\theta}_k \to \theta^* \ a.s.$$

*Proof.* Due to the page limit, we provide the sketch of the proof. It is worth noting that the proof here largely follows Spall and Cristion (1998) with modifications due to different assumptions and the newly proposed complex-step gradient approximation. Denote $\tilde{\theta}_k = \hat{\theta}_k - \theta^*$, we then proceed to show the almost sure convergence by the following main three steps: i) $\mathbb{P}(\limsup_{k\to\infty} \|\tilde{\theta}_k\|_2 =$

$\infty) = 0$; ii) $\tilde{\theta}_k \to \tilde{\theta}^*$ a.s. for some unique $\tilde{\theta}^*$; iii) the limit $\tilde{\theta}^*$ must be the constant zero.

Recall that Theorem 4.1 gives $\bar{\boldsymbol{g}}_{k+1}(\hat{\theta}_k) - \boldsymbol{g}_{k+1}(\hat{\theta}_k) = O(c_{k+1}^2)$. To prove the first step, we rely on the martingale convergence theorem. From Assumptions 4.12–4.15 and using the relationship that $\mathbb{E}[\hat{\boldsymbol{g}}_{k+1}(\hat{\theta}_k) - \bar{\boldsymbol{g}}_{k+1}(\hat{\theta}_k)] = 0$, it is easy to see that the sequence $\{\sum_{t=0}^k a_t[\hat{\boldsymbol{g}}_{t+1}(\hat{\theta}_t) - \bar{\boldsymbol{g}}_{t+1}(\hat{\theta}_t)]\}_{k\geq 0}$ represents a martingale sequence, where $\mathbb{E}[\| \sum_{t=0}^k a_t[\hat{\boldsymbol{g}}_{t+1}(\hat{\theta}_t) - \bar{\boldsymbol{g}}_{t+1}(\hat{\theta}_t)]\|_2^2] \leq \sum_{t=0}^k \mathbb{E}[\|a_t[\hat{\boldsymbol{g}}_{t+1}(\hat{\theta}_t) - \bar{\boldsymbol{g}}_{t+1}(\hat{\theta}_t)]\|_2^2] < \infty$. Observing that

$$\begin{aligned} &\hat{\theta}_0 - \theta^* - \sum_{t=0}^k a_t[\hat{\boldsymbol{g}}_{t+1}(\hat{\theta}_t) - \bar{\boldsymbol{g}}_{t+1}(\hat{\theta}_t)] \\ &= \tilde{\theta}_{k+1} + \sum_{t=0}^k a_t \bar{\boldsymbol{g}}_{t+1}(\hat{\theta}_t), \end{aligned}$$

the martingale convergence theorem implies that $\tilde{\theta}_{k+1} + \sum_{t=0}^k a_t \bar{\boldsymbol{g}}_{t+1}(\hat{\theta}_t)$ converges almost surely to some integrable random variable.

Since we have shown that $\bar{\boldsymbol{g}}_{k+1}(\hat{\theta}_k) - \boldsymbol{g}_{k+1}(\hat{\theta}_k) = O(c_{k+1}^2)$, which converges to 0 as $c_k \to 0$ by Assumption 4.12, we see that (4.24) implies

$$\limsup_{k\to\infty} \left|\frac{\sum_{j\notin S}(\theta - \theta^*)\bar{g}_{k+1,j}(\theta)}{\sum_{j\in S}(\theta - \theta^*)\bar{g}_{k+1,j}(\theta)}\right| < 1 \text{ a.s.}$$

Hence, using the constant $\lambda$ in Assumption 4.18, we have

$$\cup_S\{\tilde{\theta}_{k,j} \to \infty \text{ for } j \in S\} \subseteq \cup_{S,\tau>0}\{\mathcal{A} \cup \mathcal{B}\},$$

where the event $\mathcal{A}$

$$\mathcal{A} = \{\tilde{\theta}_{k,j} \geq \lambda \text{ for } j \in S, \tilde{\theta}_{k,j} \leq \tau \text{ and } j \notin S\}$$
$$\cap \limsup_{k\to\infty} \{a_k \bar{g}_{k+1,j}(\hat{\theta}_k) < 0 \text{ for } j \in S\}, \tag{4.25}$$

and the event $\mathcal{B}$

$$\mathcal{B} = \{\tilde{\theta}_{k,j} \to \infty \text{ for } j \in S\}$$
$$\cap \liminf_{k\to\infty} \{a_k \bar{g}_{k+1,j}(\hat{\theta}_k) < 0 \text{ for } j \in S\}^c. \tag{4.26}$$

For event $\mathcal{A}$, there exists a subsequence $\{k_s\}$ such that the event $\mathcal{C} = \{\tilde{\theta}_{k_s,j} \geq \lambda \text{ for } i \in S\} \cap \{a_{k_s} \bar{g}_{k_s+1,j}(\hat{\theta}_{k_s}) < 0 \text{ for } i \in S\}$ is true. However, with $\bar{\boldsymbol{g}}_{k_s+1}(\hat{\theta}_{k_s}) - \boldsymbol{g}_{k_s+1}(\hat{\theta}_{k_s}) \to 0$, the event $\mathcal{C}$ also implies $\tilde{\theta}_{k_s}^T \boldsymbol{g}_{k_s+1}(\hat{\theta}_{k_s}) < 0$ for any sufficiently large $k_s$, which contradicts Assumption 4.16. Hence, we must have $\mathbb{P}(\mathcal{A}) = 0$ for any $\tau$ and S. For event $\mathcal{B}$, another contradiction argument can be made by drawing conclusions from the almost sure convergence of $\tilde{\theta}_{k+1} + \sum_{t=0}^{k} a_t \bar{\boldsymbol{g}}_{t+1}(\hat{\theta}_t)$ and Assumption 4.17 in a similar fashion as the proof in Spall and Cristion (1998, Proposition). Therefore, we can show $\mathbb{P}(\mathcal{B}) = 0$, which further implies that $\mathbb{P}(\limsup_{k\to\infty} \|\tilde{\theta}_k\|_2 = \infty) = 0$.

To prove the second step, it is sufficient to show that for any $j$ and $a < b$,

$$\mathbb{P}(\liminf_{k\to\infty} \tilde{\theta}_{k,j} < a < b < \limsup_{k\to\infty} \tilde{\theta}_{k,j}) = 0. \tag{4.27}$$

From the result of the first step, we note that there exists a sub-subsequence $\{k_{s_l}\}$ such that

$$\limsup_{l\to\infty} |\sum_{t=0}^{k_{s_l}} a_t \bar{g}_{t+1,j}(\hat{\theta}_t)| < \infty \text{ a.s.}$$

If there exists some constants $a$ and $b$ such that $\liminf_{k\to\infty} \tilde{\theta}_{k,j} < a < b < \limsup_{k\to\infty} \tilde{\theta}_{k,j}$, Assumption 4.17 implies that for any $\rho > 0$ we can choose $m > n$ sufficiently large such that $|\sum_{t=k_{s_n}}^{k_{s_m-1}} a_t \bar{g}_{t+1,j}(\hat{\theta}_t)| \leq \rho$ and $|\tilde{\theta}_{k_{s_m},j} - \tilde{\theta}_{k_{s_n},j} + \sum_{t=k_{s_n}}^{k_{s_m-1}} a_t \bar{g}_{t+1,j}(\hat{\theta}_t)| \leq (b-a)/3$. By picking $\rho < (b-a)/3$, we have

$$|\tilde{\theta}_{k_{s_m},j} - \tilde{\theta}_{k_{s_n},j}| \leq \frac{2(b-a)}{3}. \tag{4.28}$$

However, recall that we assume $\tilde{\theta}_{k_{s_n},j} < a < b < \tilde{\theta}_{k_{s_m},j}$, which implies $\tilde{\theta}_{k_{s_m},j} - \tilde{\theta}_{k_{s_n},j} > b - a$. Hence, by contradicting with (4.28), we must have (4.27) hold for the second step.

For the last step, it is sufficient show that

$$\mathbb{P}(\lim_{k\to\infty} \tilde{\theta}_k \neq 0 \text{ and } \|\sum_{t=0}^{\infty} a_t \bar{\boldsymbol{g}}_{t+1}(\hat{\theta}_t)\|_2^2 < \infty) = 0. \tag{4.29}$$

Assume the event in (4.29) does hold and denote $J = \{j : \tilde{\theta}_{k,j} \not\to 0 \text{ and } j \in$

$\{1, \ldots, p\}\}$. From the result of second step, we see that there exists some constants $L$ and $U$, and a sufficiently large $K$ such that for any $k \geq K$, we have $L \leq |\tilde{\theta}_{k,j}| \leq U$ when $j \in J$ and $|\tilde{\theta}_{k,j}| \leq L$ when $j \notin J$. Now, with $\bar{g}_{k+1,j}(\hat{\theta}_k) = g_{k+1,j}(\hat{\theta}_k) + O(c_k^2)$ and Assumption 4.16, we have

$$\sum_{t=K+1}^{n} a_k \sum_{j \in J} \tilde{\theta}_{k,j} \bar{g}_{k+1,j}(\hat{\theta}_k) \geq \sum_{t=K+1}^{n} a_k \delta_k(a).$$

However, since $\bar{g}_{k+1,j}(\hat{\theta}_k)$ can only change sign a finite number of times by Assumption 4.17, there exists some $j \in J$ such that

$$\limsup_{n \to \infty} \left| \frac{\sum_{t=K+1}^{n} a_k \delta_k(a)}{\sum_{t=K+1}^{n} a_k \bar{g}_{k+1,j}(\hat{\theta}_k)} \right| < \infty.$$

Hence, we can must have $|\sum_{t=K+1}^{n} a_k \bar{g}_{k+1,j}(\hat{\theta}_k)| = \infty$ given that $\sum_{t=K+1}^{n} a_k \delta_k(a) = \infty$ by Assumption 4.16. Finally, after observing that the existence of such $j \in J$ contradicts the event in (4.29), we must have (4.29) holds, which completes the proof. □

## 4.4 Numerical Study

### 4.4.1 Synthetic Problem

In this subsection, we test the performance of the proposed CS-SPSA on a simple synthetic problem. This loss function has also been examined in Klein-

man et al. (1999) and Spall (2005, Example 14.8). Specifically, we consider the following noisy function measurement,

$$\hat{L}(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\boldsymbol{\theta} + \sum_{j=1}^{p} \exp(-X_j t_j), \tag{4.30}$$

where $\boldsymbol{\theta} = [t_1, \ldots, t_p]^T \in [0,\infty)\times\cdots\times[0,\infty)$, and $X_j \sim \exp(\eta_j)$ is an exponentially distributed random variable with the rate parameter $\eta_j$ for $j = 1, \ldots, J$. Using the basic property of exponential distribution, it is easy to see that the expected loss function has the following formula

$$L(\boldsymbol{\theta}) = \mathbb{E}[\ell(\boldsymbol{\theta})] = \boldsymbol{\theta}^T\boldsymbol{\theta} + \sum_{j=1}^{p} \frac{\eta_j}{\eta_j + t_j}. \tag{4.31}$$

Note that both (4.30) and (4.31) satisfy the requirements of being analytic functions so that whenever $\boldsymbol{\theta} \in \mathbb{C}^p$, the function outputs $L(\boldsymbol{\theta}) \in \mathbb{C}$ and $\ell(\boldsymbol{\theta}) \in \mathbb{C}$. Moreover, one can also check that the CS-SP gradient approximation returns an appropriate estimate for the true gradient when using (4.30) and (4.31). Based on the values provided in Kleinman et al. (1999), we first consider the case where $p = 10$ and $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_p]^T = [1.10254, 1.69449, 1.47894, 1.92617, 0.750471, 1.32673, 0.842822, 0.724652, 0.769311, 1.3986]^T$. The optimal estimate $\boldsymbol{\theta}^*$ is computed algebraically by solving $\partial L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$.

Setting the initial estimate $\hat{\boldsymbol{\theta}}_0 = [1, \ldots, 1]^T$, we compare the proposed CS-SPSA with FDSA, SPSA and CS-FDSA. The gain sequence $\{a_k\}$ and perturba-

tion sequence $\{c_k\}$ are chosen to have the standard form $a_k = a/(k+1+A)$ and $c_k = c/(k+1)^\gamma$ for scalars $a > 0$, $A \geq 0$, $c > 0$ and $\gamma > 0$. Following Spall (1992), for all $k$, the distribution of the random perturbation vectors $\boldsymbol{\Delta}_k$ in SPSA and CS-SPSA is set to be independent Bernoulli $\pm 1$ distribution with equal probabilities. After tuning for optimal performance in terminal loss function values, we choose $a = 0.02, A = 100, \alpha = 0.668$, $c = 0.2$ and $\gamma = 0.167$ for all the algorithms. Since each algorithm uses a different number of function measurement per iteration, for a fair comparison under a fixed total budget or computational resource, we present the results in terms of the number of function measurements. All the algorithms are implemented using a total of 50,000 function measurements per replicate and the results are averaged over 20 independent replicates. Figure 4.1 shows the performance of all the algorithms in terms of the normalized loss function error $[L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}^*)]/[L(\hat{\boldsymbol{\theta}}_0) - L(\boldsymbol{\theta}^*)]$. Figure 4.2 shows the performance of all the algorithm in terms of the normalized parameter estimate error $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|/\|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|$.

It is clear from Figure 4.1 and 4.2 that CS-SPSA performs the best among all the algorithms. The advantage of using only one function measurement per iteration makes CS-SPSA converge to the optimal value much faster.

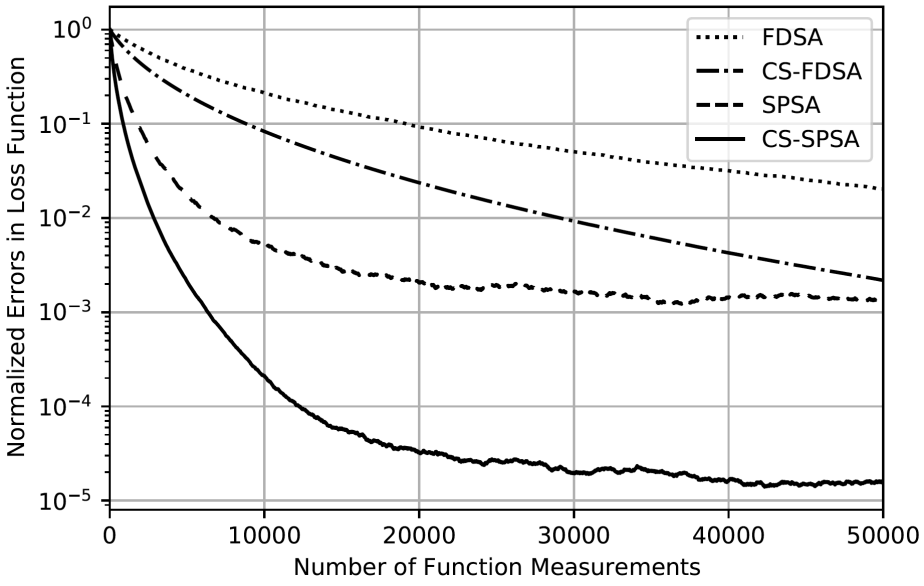



**Figure 4.1:** Performance of FDSA, SPSA, CS-FDSA and CS-SPSA in terms of $[L(\hat{\theta}_k) - L(\theta^*)]/[L(\hat{\theta}_0) - L(\theta^*)]$ across 50,000 function measurements and averaged over 20 independent replicates.

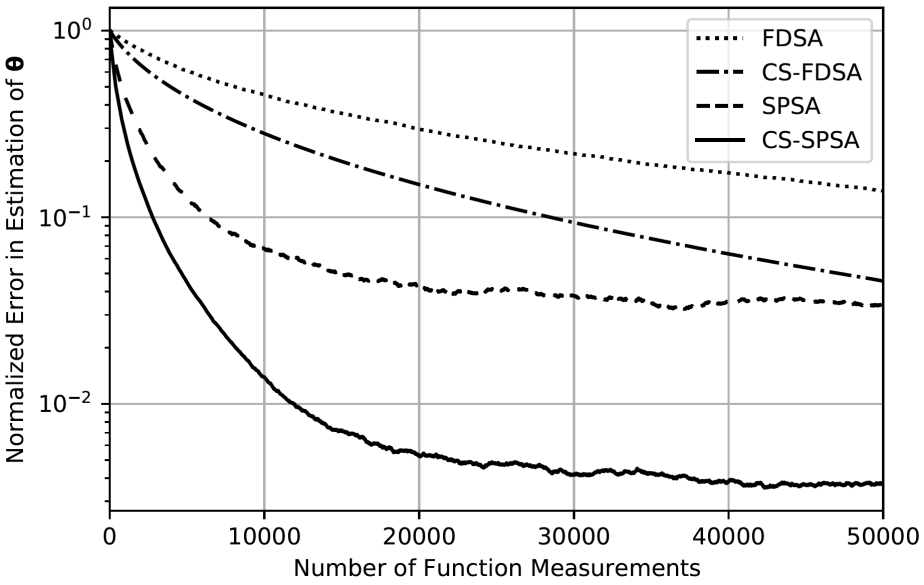



**Figure 4.2:** Performance of FDSA, SPSA, CS-FDSA and CS-SPSA in terms of $\|\hat{\theta}_k - \theta^*\|/\|\hat{\theta}_0 - \theta^*\|$ across 50,000 function measurements and averaged over 20 independent replicates.

### 4.4.2 A Data-Driven Linear-quadratic Regulator

To examine the performance of CS-SPSA on a real-world problem, we consider the following well-known data-drive linear-Quadratic regulator (LQR) problem. Mathematically, we have

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t + \boldsymbol{w}_t,$$

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{v}_t,$$

where $\boldsymbol{x}_t \in \mathbb{R}^n$ is the state vector, $\boldsymbol{u} \in \mathbb{R}^m$ is the control vector, $\boldsymbol{w}_t \in \mathbb{R}^m$ is the state noise, $\boldsymbol{y}_t \in \mathbb{R}^n$ is the measurement vector and $\boldsymbol{v}_t \in \mathbb{R}^n$ is the measurement noise, and the matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}, \boldsymbol{B} \in \mathbb{R}^{n\times m}$ are both system matrices. Consider a linear feedback control law $\boldsymbol{u}_t = -\boldsymbol{K}\boldsymbol{x}_t$ with $\boldsymbol{K} \in \mathbb{R}^{m\times n}$. We can define the parameter of interest $\theta$ as the vector form of $\boldsymbol{K}$ such that $\theta = \text{vec}(\boldsymbol{K}) \in \mathbb{R}^p$ with $p = mn$. Our goal is to find an optimal $\theta$ such that the following linear-quadratic regulator (LQR) cost is minimized,

$$L(\theta) = \mathbb{E}\left[\boldsymbol{x}_T^T\boldsymbol{Q}\boldsymbol{x}_T + \sum_{t=0}^{T-1}(\boldsymbol{x}_t^T\boldsymbol{Q}\boldsymbol{x}_t + \boldsymbol{u}_t^T\boldsymbol{R}\boldsymbol{u}_t)\right],$$

where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{R} \in \mathbb{R}^{m\times m}$ are both given positive definite matrices determined by the user to reflect the relative weights to put on the cost associated with the state vector and the control vector. The LQR cost is generally a non-

convex function (Fazel et al., 2018). Since the state vector $\boldsymbol{x}_t$ is inaccessible and a user can only observe the measurement vector $\boldsymbol{y}_t$, we rely on the following noisy function measurement

$$\ell(\theta) = \boldsymbol{y}_T^T \boldsymbol{Q} \boldsymbol{y}_T + \sum_{t=0}^{T-1} (\boldsymbol{y}_t^T \boldsymbol{Q} \boldsymbol{y}_t + \boldsymbol{u}_t^T \boldsymbol{R} \boldsymbol{u}_t)$$

to find the optimal $\theta$. It is also worth noting that the proposed CS-SPSA requires noisy function measurements at the complex-valued $\hat{\theta}_k + ic_k\boldsymbol{\Delta}_k$ in order to construct the gradient approximation $\hat{\boldsymbol{g}}_k^{\text{CS-SP}}(\hat{\theta}_k)$. Therefore, it might not be possible to collect data by conducting physical experiments as in real world systems, especially when the real systems only accept real-valued inputs. The CS-SPSA can be useful, however, when one can simulate the system in computer code so that complex-valued inputs are acceptable. Since the optimal solution found by CS-SPSA are always real values, the solution can still be used to guide the corresponding real systems.

Following Al-Abri et al. (2020), we choose an asymptotically stable system with $\boldsymbol{K} \in \mathbb{R}^{3\times 4}$ and the parameter values,

$$\boldsymbol{A} = \begin{bmatrix} -2.5 & 1.2 & 4.3 & 0.1 \\ 0.97 & -10.3 & 0.4 & -6.1 \\ -9.2 & 1.1 & -4.9 & 0.3 \\ 1.1 & 0.9 & -3.4 & -0.9 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} 1.1 & 0.4 & -0.2 \\ -3.2 & 1.4 & 0 \\ -0.8 & 0.1 & 3.0 \\ -1.1 & -0.9 & 5.2 \end{bmatrix}, \boldsymbol{Q} = \boldsymbol{I}_4 \text{ and } \boldsymbol{R} = \boldsymbol{I}_3.$$

Both the state and measurement noises are assumed to be independent Gaussian distributed such that $\boldsymbol{w}_t \sim \mathcal{N}_4(\mathbf{0}, 0.1^2\boldsymbol{I}_4)$ and $\boldsymbol{v}_t \sim \mathcal{N}_4(\mathbf{0}, 0.1^2\boldsymbol{I}_4)$. The LQR cost is computed with roll out length $T = 100$. Using the initial state $\boldsymbol{x}_0 = [20, 40, -20, -10]^T$ and the initial estimate $\hat{\theta}_0 = [2, \ldots, 2]^T$, we consider the comparison between SPSA and CS-SPSA. With the random perturbation vector $\boldsymbol{\Delta}_k$ being independent Bernoulli $\pm 1$ distributed with equal probabilities. After tuning for optimal performance of SPSA, the gain and perturbation sequence parameters are set to be $a = 10^{-4}, A = 100, c = 0.5, \alpha = 0.668$ and $\gamma = 0.167$ for both SPSA and CS-SPSA. We implement both algorithms for 500 iterations each. Since the total computational times are about the same, i.e., SPSA and CS-SPSA takes 218.51 and 216.09 seconds, respectively, we present the final result in terms of the number of iterations, not the number of function measurements. Figure 4.3 shows the normalized loss function value, i.e., $[L(\hat{\theta}_k) - L(\theta^*)]/[L(\hat{\theta}_0) - L(\theta^*)]$, where $\theta^*$ is the minimizer of $L(\theta^*)$ such that $L(\theta^*) = 4149.3895$. We can see that CS-SPSA outperforms SPSA after just 100 iterations and the advantages persist for all later iterations.

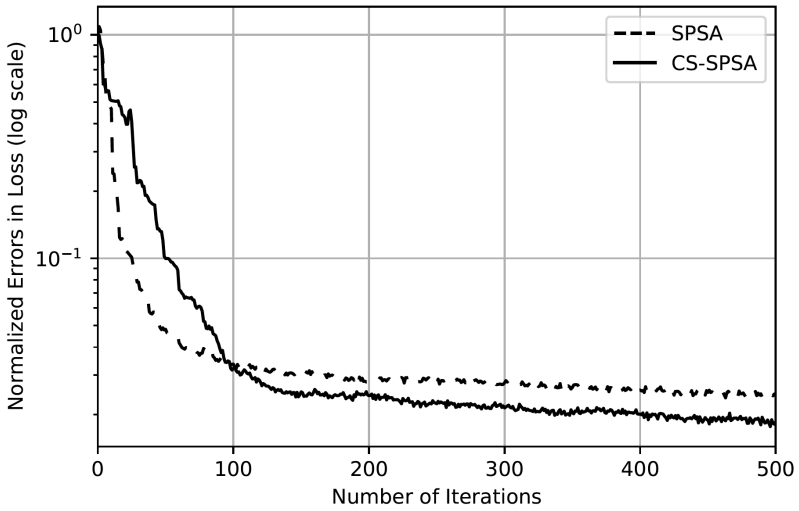



**Figure 4.3:** Performances of SPSA and CS-FDSA in terms of normalized errors in loss, i.e., $[L(\hat{\theta}_k) - L(\theta^*)]/[L(\hat{\theta}_0) - L(\theta^*)]$, across 500 iterations and averaged over 20 independent replicates.

### 4.4.3 Non-additive Noise Model

For the numerical study, we consider the non-additive noise model introduced in Yaz (1987). Mathematically, we have

$$\boldsymbol{y}_{k+1} = \begin{pmatrix} -0.5 & 0.3 \\ 0 & 1.1 \end{pmatrix} \boldsymbol{y}_{k+1} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \boldsymbol{u}_k + \begin{pmatrix} \|\boldsymbol{y}_k\|_2 \\ 0 \end{pmatrix} w_k, \tag{4.32}$$

where $\boldsymbol{y}_k, \boldsymbol{u}_k \in \mathbb{R}^2$, and $w_k$ is independently Bernoulli $\pm 0.5$ distributed with equal probabilities. As mentioned in Spall and Cristion (1998), beside the non-additive noise term, the system can only be affected by the second components of the control variable and the first component of $\boldsymbol{y}_k$ can only be affected by the control variable after a delay of one time period. Note that since work

considers a model-free approach, there is no prior knowledge of the dynamics (4.32). Hence, although the first component of $\boldsymbol{u}_k$ has no affect on the system, we still need to construct a full controller to generate both components of $\boldsymbol{u}_k$. To test the performance of the tracking ability, we use a periodic square wave target sequence, i.e., $\boldsymbol{t}_k = (1, 0)^T$ for the first five iterations of the period and $\boldsymbol{t}_k = (-1, 0)^T$ for the second five iterations of the period. A quadratic loss function,

$$L_k(\theta_k) = \mathbb{E}[(\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T \boldsymbol{A}_k (\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T]$$

with $\boldsymbol{A}_k = 0.5\boldsymbol{I}_2$ is considered here. The long-run best possible root-mean-square (RMS) error, i.e., $\{\mathbb{E}[(\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T \boldsymbol{A}_k (\boldsymbol{y}_{k+1} - \boldsymbol{t}_{k+1})^T]\}^{1/2}$ is $1/\sqrt{2} \approx 0.7071$. To construct the controller, we use a second-order polynomial with the most recent system variable and the target as the inputs, i.e., $\boldsymbol{u}_k = f(\theta_k; \boldsymbol{y}_k, \boldsymbol{t}_{k+1})$. There are a total of $p = 30$ parameters.

For our proposed CS-SPSA algorithm, we choose a decaying gain sequence $a_k = a/(k + 1 + A)^{0.667}$ with $a = 0.1, A = 10$ and $c_k = c/(k + 1)^{0.167}$ with $c = 0.05$. The initial state is set to $\boldsymbol{y}_0 = [2, 1]^T$. Table 4.2 shows the RMS error at various iterations. We can see that CS-SPSA successfully minimizes the loss function values and, even at iteration 50, the estimated loss function error is already close the optimal value. For SPSA and the one-measurement versions SPSA, both algorithm fail to converge after 500 iterations. Note that the controller FA here is much simpler than the third-order polynomial used in Spall and

Cristion (1998), which has 70 parameters.

**Table 4.2:** Estimated RMS error for the non-additive noise model

| **Iteration Number** | **CS-SPSA** |
|---|---|
| 0 | 2.236 |
| 50 | 0.790 |
| 100 | 0.709 |
| 500 | 0.708 |

## 4.5 Conclusion

Borrowing some ideas from SPSA and CS-FDSA, CS-SPSA is proposed and established to solve stochastic optimization problems with only noisy function measurements. Using the complex-valued noisy function measurement to estimate the gradient, theoretical results suggest that the proposed CS-SPSA algorithm converges almost surely at an accelerated rate of $k^{-1/2}$, which is faster than the standard convergence rate of $k^{-1/3}$ and is only achievable as if CRNs are used in SPSA. The numerical studies on a synthetic problem and the data-driven LQR problem show that CS-SPSA delivers the best results in terms of accuracy and robustness compared with FDSA, CS-FDSA, and SPSA. Practically, the algorithm is also shown to perform well for constructing a model-free controller of a non-additive noise model. We note that the newly-proposed CS-

SPSA is applicable only when the loss function can be extended to evaluate complex parts, which may restrict its use when interacting with real-world systems. Otherwise, SPSA remains to be the "gold standard" in general problems when the loss function evaluation outputs real-valued scalar only. For future directions, it is also possible to consider the more generalized forms of CS gradient approximation listed in Abreu et al. (2013) to construct the gradient approximation, such as $\hat{\boldsymbol{g}}_k(\hat{\theta}_k) = \Im(\ell(\hat{\theta}_k + \tilde{c}_k\tilde{\boldsymbol{\Delta}}_k + ic_k\boldsymbol{\Delta}_k, \boldsymbol{v}_k))/c_k\boldsymbol{\Delta}_k$ for a new pair of $\tilde{c}_k$ and $\tilde{\boldsymbol{\Delta}}_k$. In this work, however, as the first attempt to utilize both the SP and CS gradient approximations, we stick to the most basic and widely used version (4.2) for gradient approximation.

# Bibliography

Abreu, R., Stich, D., and Morales, J. (2013). On the generalization of the complex step method. *Journal of Computational and Applied Mathematics*, 241:84–102.

Abreu, R., Su, Z., Kamm, J., and Gao, J. (2018). On the accuracy of the complex-step-finite-difference method. *Journal of Computational and Applied Mathematics*, 340:390–403.

Ahmad, M. A., Azuma, S.-i., and Sugie, T. (2014). Performance analysis of model-free pid tuning of mimo systems based on simultaneous perturbation stochastic approximation. *Expert Systems with Applications*, 41(14):6361–6370.

AISI, A. (2015). S400-15, North American standard for seismic design of cold-formed steel structural systems. *American Iron and Steel Institute, Washington, DC*.

Aksakalli, V. and Malekipirbazari, M. (2016). Feature selection via binary si-

multaneous perturbation stochastic approximation. *Pattern Recognition Letters*, 75:41–47.

Al-Abri, S., Lin, T. X., Tao, M., and Zhang, F. (2020). A derivative-free optimization method with application to functions with exploding and vanishing gradients. *IEEE Control Systems Letters*, 5(2):587–592.

Alrefaei, M. H. and Andradóttir, S. (2001). A modification of the stochastic ruler method for discrete stochastic optimization. *European Journal of Operational Research*, 133(1):160–182.

Apostol, T. M. (1974). *Mathematical Analysis*. Addison Wesley Publishing Company.

Balzani, D., Gandhi, A., Tanaka, M., and Schröder, J. (2015). Numerical calculation of thermo-mechanical problems at large strains based on complex step derivative approximation of tangent stiffness matrices. *Computational Mechanics*, 55(5):861–871.

Banks, H., Bekele-Maxwell, K., Bociu, L., Noorman, M., and Tillman, K. (2015). The complex-step method for sensitivity analysis of non-smooth problems arising in biology. *Eurasian Journal of Mathematical and Computer Applications*, 3:15–68.

Bhatnagar, S., Prasad, H., and Prashanth, L. (2013). *Stochastic Recursive Al-*

*gorithms for Optimization: Simultaneous Perturbation Methods*, volume 434. Springer.

Bian, G., Chatterjee, A., Buonopane, S. G., Arwade, S. R., Moen, C. D., and Schafer, B. W. (2017). Reliability of cold-formed steel framed shear walls as impacted by variability in fastener response. *Engineering Structures*, 142:84–97.

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics: Basic and Selected Topics 1 (updated printing)*. Pearson Prentice Hall.

Blakney, A. and Zhu, J. (2019). A comparison of the finite difference and simultaneous perturbation gradient estimation methods with noisy function evaluations. In *Proceedings of the Annual Conference on Information Sciences and Systems*, pages 1–6, Baltimore, MD.

Blum, J. R. (1954). Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744.

Boccara, N. (2010). *Modeling Complex Systems*. Springer Science & Business Media.

Boem, F., Ferrari, R. M., Keliris, C., Parisini, T., and Polycarpou, M. M. (2017). A distributed networked approach for fault detection of large-scale systems. *IEEE Transactions on Automatic Control*, 62(1):18–33.

BIBLIOGRAPHY

Branston, A. E., Boudreault, F. A., Chen, C. Y., and Rogers, C. A. (2006). Light-gauge steel-frame wood structural panel shear wall design method. *Canadian Journal of Civil Engineering*, 33(7):872–889.

Brockett, R. W. (2010). On the control of a flock by a leader. *Proceedings of the Steklov Institute of Mathematics*, 268(1):49–57.

Buckley, J. J. (1992). Universal fuzzy controllers. *Automatica*, 28(6):1245–1248.

Buonopane, S. G., Bian, G., Tun, T. H., and Schafer, B. W. (2015). Computationally efficient fastener-based models of cold-formed steel shear walls with wood sheathing. *Journal of Constructional Steel Research*, 110:137–148.

Cao, Y., Jiang, L., and Wu, Q. H. (2000). An evolutionary programming approach to mixed-variable optimization problems. *Applied Mathematical Modelling*, 24(12):931–942.

Cárdenas, A. A., Amin, S., Lin, Z.-S., Huang, Y.-L., Huang, C.-Y., and Sastry, S. (2011). Attacks against process control systems: risk assessment, detection, and response. In *Proceedings of the ACM Symposium on Information, Compuer and Communications Security*, pages 355–366, Hong Kong, China.

Chen, H. and Schmeiser, B. W. (2001). Stochastic root finding via retrospective approximation. *IIE Transactions*, 33(3):259–275.

Chen, T. and Chen, H. (1995). Universal approximation to nonlinear operators

by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917.

Chung, K. L. (2001). *A Course in Probability Theory*. Academic Press.

Ciblat, P. and Ghogho, M. (2004). Harmonic retrieval in non-circular complex-valued multiplicative noise: Cramer-Rao bound. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 489–492, Montreal, Canada.

Crawford, B., Soto, R., Astorga, G., García, J., Castro, C., and Paredes, F. (2017). Putting continuous metaheuristics to work in binary search spaces. *Complexity*, 2017(8404231):1–19.

Csáji, B. C. et al. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24(48):7.

Duan, H., Xin, L., and Chen, S. (2019). Robust cooperative target detection for a vision-based UAVs autonomous aerial refueling platform via the contrast sensitivity mechanism of eagle's eye. *IEEE Aerospace and Electronic Systems Magazine*, 34(3):18–30.

Fabian, V. (1968). On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332.

BIBLIOGRAPHY

Favati, P. (1990). Convexity in nonlinear integer programming. *Ricerca Operativa*, 53:3–44.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the International Conference on Machine Learning*, pages 1467–1476, Stockholm, Sweden.

Ferrez, J.-A., Fukuda, K., and Liebling, T. M. (2005). Solving the fixed rank convex quadratic maximization in binary variables by a parallel zonotope construction algorithm. *European Journal of Operational Research*, 166(1):35–50.

Fleming, W. H. (2012). *Functions of Several Variables*. Springer Science & Business Media.

Fujishige, S. and Murota, K. (2000). Notes on L-/M-convex functions and the separation theorems. *Mathematical Programming*, 88(1):129–146.

Gandomi, A. H., Yang, X.-S., and Alavi, A. H. (2013). Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. *Engineering with Computers*, 29(1):17–35.

Gong, W.-B., Ho, Y.-C., and Zhai, W. (2000). Stochastic comparison algorithm

for discrete optimization with estimation. *SIAM Journal on Optimization*, 10(2):384–404.

Guan, Y., Ahmed, S., and Nemhauser, G. L. (2009). Cutting planes for multi-stage stochastic integer programs. *Operations Research*, 57(2):287–298.

Guo, J., Li, Z. S., and Jin, J. J. (2018). System reliability assessment with multilevel information using the Bayesian melding method. *Reliability Engineering & System Safety*, 170:146–158.

Guo, J. and Wilson, A. G. (2013). Bayesian methods for estimating system reliability using heterogeneous multilevel information. *Technometrics*, 55(4):461–472.

Gutjahr, W. J. and Pflug, G. C. (1996). Simulated annealing for noisy cost functions. *Journal of Global Optimization*, 8(1):1–13.

He, Y., Fu, M. C., and Marcus, S. I. (2003). Convergence of simultaneous perturbation stochastic approximation for non-differentiable optimization. *IEEE Transactions on Automatic Control*, 48(8):1459–1463.

Hemker, T., Fowler, K. R., Farthing, M. W., and von Stryk, O. (2008). A mixed-integer simulation-based optimization approach with surrogate functions in water resources management. *Optimization and Engineering*, 9(4):341–360.

Hernández, K. and Spall, J. C. (2015). System identification for multi-sensor

data fusion. In *Proceedings of the American Control Conference*, pages 3931–3936, Chicago, IL.

Higham, N. (2018). Differentiation with (out) a difference. *SIAM News*, 51(5):2.

Hill, S. D. (2014). Discrete optimization with noisy objective function measurements. *Wiley Encyclopedia of Operations Research and Management Science*.

Hill, S. D., Gerencsér, L., and Vágó, Z. (2004). Stochastic approximation on discrete sets using simultaneous difference approximations. In *Proceedings of the American Control Conference*, pages 2795–2798, Portland, OR.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press.

Hornik, K., Stinchcombe, M., White, H., et al. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Javidi, S., Mandic, D. P., and Cichocki, A. (2010). Complex blind source extraction from noisy mixtures using second-order statistics. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 57(7):1404–1416.

Kannan, B. and Kramer, S. (1994). An augmented Lagrange multiplier based method for mixed integer discrete continuous optimization and its applications to mechanical design. *Journal of Mechanical Design*, 116(2):405–411.

BIBLIOGRAPHY

Khaneja, N., Reiss, T., Kehlet, C., Schulte-Herbrüggen, T., and Glaser, S. J. (2005). Optimal control of coupled spin dynamics: Design of NMR pulse sequences by gradient ascent algorithms. *Journal of Magnetic Resonance*, 172(2):296–305.

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

Kim, J., Bates, D. G., and Postlethwaite, I. (2006). Nonlinear robust performance analysis using complex-step gradient approximation. *Automatica*, 42(1):177–182.

Kleinman, N. L., Spall, J. C., and Naiman, D. Q. (1999). Simulation-based optimization with stochastic approximation using common random numbers. *Management Science*, 45(11):1570–1578.

Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502.

Knorn, S. and Middleton, R. H. (2013). String stability analysis of a vehicle platoon with communication range 2 using the two-dimensional induced operator norm. In *Proceedings of the European Control Conference*, pages 3354–3359, Zurich, Switzerland.

BIBLIOGRAPHY

Krejić, N., Lužanin, Z., Nikolovski, F., and Stojkovska, I. (2015). A non-monotone line search method for noisy minimization. *Optimization Letters*, 9(7):1371–1391.

Kushner, H. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media.

Kushner, H. J. and Clark, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, volume 26. Springer Science & Business Media.

Laha, R. and Rohatgi, V. (1979). *Probability Theory*. John Wiley.

Lane, S. H., Handelman, D. A., and Gelfand, J. J. (1992). Theory and development of higher-order CMAC neural networks. *IEEE Control Systems Magazine*, 12(2):23–30.

Lasdon, L., Mitter, S., and Waren, A. (1967). The conjugate gradient method for optimal control problems. *IEEE Transactions on Automatic Control*, 12(2):132–138.

Li, M., Zhang, W., Hu, Q., Guo, H., and Liu, J. (2017). Design and risk evaluation of reliability demonstration test for hierarchical systems with multilevel information aggregation. *IEEE Transactions on Reliability*, 66(1):135–147.

BIBLIOGRAPHY

Luo, B., Liu, D., Wu, H.-N., Wang, D., and Lewis, F. L. (2016). Policy gradient adaptive dynamic programming for data-based optimal control. *IEEE Transactions on Cybernetics*, 47(10):3341–3354.

Lyness, J. N. and Moler, C. B. (1967). Numerical differentiation of analytic functions. *SIAM Journal on Numerical Analysis*, 4(2):202–210.

Mamdani, E. H. (1976). Advances in the linguistic synthesis of fuzzy controllers. *International Journal of Man-Machine Studies*, 8(6):669–678.

Maranzano, C. J. and Spall, J. C. (2010a). Implementation and application of maximum likelihood reliability estimation from subsystem and full system tests. In *Proceedings of the Performance Metrics for Intelligent Systems Workshop*, pages 146–153, Baltimore, MD.

Maranzano, C. J. and Spall, J. C. (2010b). Robust test design for reliability estimation with modeling error when combining full system and subsystem tests. In *Proceedings of the American Control Conference*, pages 3741–3746, Baltimore, MD.

Maranzano, C. J. and Spall, J. C. (2011). Framework for estimating system reliability from full system and subsystem tests with dependence on dynamic inputs. In *Proceedings of the Joint IEEE Conference on Decision and Control and European Control Conference*, pages 6666–6671, Orlando, FL.

BIBLIOGRAPHY

Martins, J., Kroo, I., and Alonso, J. (2000). An automated method for sensitivity analysis using complex variables. In *Proceedings of the Aerospace Sciences Meeting and Exhibit*, number 689, Reno, NV.

Martins, J., Sturdza, P., and Alonso, J. (2001). The connection between the complex-step derivative approximation and algorithmic differentiation. In *Proceedings of the Aerospace Sciences Meeting and Exhibit*, number 921, Reno, NV.

Martins, J. R., Sturdza, P., and Alonso, J. J. (2003). The complex-step derivative approximation. *ACM Transactions on Mathematical Software*, 29(3):245–262.

Matyas, J. (1965). Random optimization. *Automation and Remote Control*, 26(2):246–253.

McKenna, F. (2011). Opensees: a framework for earthquake engineering simulation. *Computing in Science & Engineering*, 13(4):58–66.

Miller, B. L. (1971). On minimizing nonseparable functions defined on the integers with an inventory application. *SIAM Journal on Applied Mathematics*, 21(1):166–185.

Murota, K. and Shioura, A. (2001). Relationship of M-/L-convex functions with

discrete convex functions by Miller and Favati–Tardella. *Discrete Applied Mathematics*, 115(1-3):151–176.

Nekouei, E., Skoglund, M., and Johansson, K. H. (2018). Privacy of information sharing schemes in a cloud-based multi-sensor estimation problem. In *Proceedings of the American Control Conference*, pages 998–1002, Milwaukee, WI.

Nikolovski, F. and Stojkovska, I. (2018). Complex-step derivative approximation in noisy environment. *Journal of Computational and Applied Mathematics*, 327:64–78.

Ntaimo, L. (2010). Disjunctive decomposition for two-stage stochastic mixed-binary programs with random recourse. *Operations Research*, 58(1):229–243.

Oncu, S., Van de Wouw, N., Heemels, W. M. H., and Nijmeijer, H. (2012). String stability of interconnected vehicles under communication constraints. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2459–2464, Maui, HI.

Onwunalu, J. E. and Durlofsky, L. J. (2010). Application of a particle swarm optimization algorithm for determining optimum well location and type. *Computational Geosciences*, 14(1):183–198.

Pasupathy, R. (2010). On choosing parameters in retrospective-approximation

algorithms for stochastic root finding and simulation optimization. *Operations Research*, 58(4-1):889–901.

Pasupathy, R. and Schmeiser, B. W. (2009). Retrospective-approximation algorithms for the multidimensional stochastic root-finding problem. *ACM Transactions on Modeling and Computer Simulation*, 19(2):1–36.

Peng, W., Huang, H.-Z., Xie, M., Yang, Y., and Liu, Y. (2013). A Bayesian approach for system reliability analysis with multilevel pass-fail, lifetime and degradation data sets. *IEEE Transactions on Reliability*, 62(3):689–699.

Peterman, K. D., Nakata, N., and Schafer, B. W. (2014). Hysteretic characterization of cold-formed steel stud-to-sheathing connections. *Journal of Constructional Steel Research*, 101:254–264.

Pflug, G. C. (2012). *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*, volume 373. Springer Science & Business Media.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497.

Reese, C. S., Wilson, A. G., Guo, J., Hamada, M. S., and Johnson, V. E. (2011). A Bayesian model for integrating multiple sources of lifetime information in

system-reliability assessments. *Journal of Quality Technology*, 43(2):127–141.

Reuther, J., Alonso, J., Martins, J., and Smith, S. (1999). A coupled aero-structural optimization method for complete aircraft configurations. In *Proceedings of the Aerospace Sciences Meeting and Exhibit*, number 187, Reno, NV.

Rudin, W. et al. (1964). *Principles of Mathematical Analysis*, volume 3. McGraw-Hill, New York.

Sadegh, P. (1997). Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation. *Automatica*, 33(5):889–892.

Sandgren, E. (1988). Nonlinear integer and discrete programming in mechanical design. In *Proceeding of the ASME Design Technology Conference*, pages 95–105, Kissimee, FL.

Sandgren, E. (1990). Nonlinear integer and discrete programming in mechanical design optimization. *Journal of Mechanical Design*, 112(2):223–229.

Sandıkçı, B., Kong, N., and Schaefer, A. J. (2013). A hierarchy of bounds for stochastic mixed-integer programs. *Mathematical Programming*, 138(1):253–272.

BIBLIOGRAPHY

Santos Coelho, L. d. (2010). Gaussian quantum-behaved particle swarm optimization approaches for constrained engineering design problems. *Expert Systems with Applications*, 37(2):1676–1683.

Sen, S. (2010). Stochastic mixed-integer programming algorithms: Beyond Benders' decomposition. *Wiley Encyclopedia of Operations Research and Management Science*.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

Sharp, C., Schaffert, S., Woo, A., Sastry, N., Karlof, C., Sastry, S., and Culler, D. E. (2005). Design and implementation of a sensor network system for vehicle tracking and autonomous interception. In *Proceedings of the European Workshop on Wireless Sensor Networks*, pages 93–107, Istanbul, Turkey.

Shiina, T. and Birge, J. R. (2003). Multistage stochastic programming model for electric power capacity expansion problem. *Japan Journal of Industrial and Applied Mathematics*, 20(3):379–397.

Son, B., Her, Y.-s., and Kim, J.-G. (2006). A design and implementation of forest-fires surveillance system based on wireless sensor networks for South Korea mountains. *International Journal of Computer Science and Network Security*, 6(9):124–130.

BIBLIOGRAPHY

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341.

Spall, J. C. (1997). A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica*, 33(1):109–112.

Spall, J. C. (2005). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley & Sons.

Spall, J. C. (2008). Reliability estimation and confidence regions from subsystem and full system tests via maximum likelihood. In *Proceedings of the Performance Metrics for Intelligent Systems Workshop*, pages 9–16, Gaithersburg, MD.

Spall, J. C. (2009). System reliability estimation and confidence regions from subsystem and full system tests. In *Proceedings of the American Control Conference*, pages 5067–5072, St. Louis, MO.

Spall, J. C. (2010). Convergence analysis for maximum likelihood-based reliability estimation from subsystem and full system tests. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2017–2022, Atlanta, GA.

Spall, J. C. (2012). Asymptotic normality and uncertainty bounds for reliabil-

ity estimates from subsystem and full system tests. In *Proceedings of the American Control Conference*, pages 56–61, Montreal, Canada.

Spall, J. C. (2013a). Maximum likelihood-based estimation of parameters in systems with binary subsystems. In *Proceedings of the Conference on Information Sciences and Systems*, pages 1–6, Baltimore, MD.

Spall, J. C. (2013b). Parameter estimation for systems with binary subsystems. In *Proceedings of the American Control Conference*, pages 83–88, Washington, DC.

Spall, J. C. (2014). Identification for systems with binary subsystems. *IEEE Transactions on Automatic Control*, 59(1):3–17.

Spall, J. C. and Cristion, J. A. (1998). Model-free control of nonlinear stochastic systems with discrete-time measurements. *IEEE Transactions on Automatic Control*, 43(9):1198–1210.

Squire, W. and Trapp, G. (1998). Using complex variables to estimate derivatives of real functions. *SIAM Review*, 40(1):110–112.

Sturdza, P., Manning, V., Kroo, I., and Tracy, R. (1999). Boundary layer calculations for preliminary design of wings in supersonic flow. In *Proceedings of the Applied Aerodynamics Conference*, number 3104, Norfolk, VA.

Sun, T., Xin, M., and Jia, B. (2016). Distributed estimation in general directed

sensor networks based on batch covariance intersection. In *Proceedings of the American Control Conference*, pages 5492–5497, Boston, MA.

Takriti, S. and Birge, J. R. (2000). Lagrangian solution techniques and bounds for loosely coupled mixed-integer stochastic programs. *Operations Research*, 48(1):91–98.

Wallace, S. W. and Fleten, S.-E. (2003). Stochastic programming models in energy. *Handbooks in Operations Research and Management Science*, 10:637–677.

Wang, H. (2012). Retrospective optimization of mixed-integer stochastic systems using dynamic simplex linear interpolation. *European Journal of Operational Research*, 217(1):141–148.

Wang, H., Ciaurri, D. E., and Durlofsky, L. J. (2010). Use of retrospective optimization for placement of oil wells under uncertainty. In *Proceedings of the Winter Simulation Conference*, pages 1750–1757, Baltimore, MD.

Wang, I.-J. and Spall, J. C. (2008). Stochastic optimisation with inequality constraints using simultaneous perturbations and penalty functions. *International Journal of Control*, 81(8):1232–1238.

Wang, L., Bian, G., Spall, J. C., and Schafer, B. W. (2018a). Combining subsystem and full system data with application to cold-formed steel shear wall. In

*Proceedings of the American Control Conference*, pages 272–277, Milwaukee, WI.

Wang, L. and Spall, J. C. (2017). Beyond the identification of reliability for system with binary subsystems. In *Proceedings of the American Control Conference*, pages 158–163, Seattle, WA.

Wang, L. and Spall, J. C. (2020). Multilevel data integration with application in sensor networks. In *Proceedings of the American Control Conference*, pages 5213–5218, Denver, CO (online conference).

Wang, L. and Spall, J. C. (2021). Improved SPSA using complex variables with applications in optimal control problems. In *Proceedings of the American Control Conference*, New Orleans, LA. in press.

Wang, L., Zhu, J., and Spall, J. C. (2018b). Mixed simultaneous perturbation stochastic approximation for gradient-free optimization with noisy measurements. In *Proceedings of the American Control Conference*, pages 3774–3779, Milwaukee, WI.

Wang, L., Zhu, J., and Spall, J. C. (2021). Model-free optimal control using SPSA with complex variables. In *Proceedings of the Annual Conference on Information Sciences and Systems*, Baltimore, MD. in press.

Wang, Q. (2013). Optimization with discrete simultaneous perturbation

stochastic approximation using noisy loss function measurements. *arXiv preprint arXiv:1311.0042*.

Wang, Q. and Spall, J. C. (2011). Discrete simultaneous perturbation stochastic approximation on loss function with noisy measurements. In *Proceedings of the American Control Conference*, pages 4520–4525, San Francisco, CA.

Wang, Q. and Spall, J. C. (2013). Rate of convergence analysis of discrete simultaneous perturbation stochastic approximation algorithm. In *Proceedings of the American Control Conference*, pages 4771–4776, Washington, DC.

Wang, Q. and Spall, J. C. (2014). Discrete simultaneous perturbation stochastic approximation for resource allocation in public health. In *Proceedings of American Control Conference*, pages 3639–3644, Portland, OR.

Wilson, A. G., McNamara, L. A., and Wilson, G. D. (2007). Information integration for complex systems. *Reliability Engineering & System Safety*, 92(1):121–130.

Wolfe, J., Chichka, D., and Speyer, J. (1996). Decentralized controllers for unmanned aerial vehicle formation flight. In *Proceedings of the Guidance, Navigation, and Control Conference*, number 3833, San Diego, CA.

Xiao, J., Hodge, B.-M. S., Liu, A. L., Pekny, J. F., and Reklaitis, G. V. (2011).

Long-term planning of wind farm siting in the electricity grid. *Computer Aided Chemical Engineering*, 29:1804–1808.

Yan, D. and Mukai, H. (1992). Stochastic discrete optimization. *SIAM Journal on Control and Optimization*, 30(3):594–612.

Yang, X.-S. and Deb, S. (2013). Multiobjective cuckoo search for design optimization. *Computers and Operations Research*, 40(6):1616–1624.

Yang, X.-S., Huyck, C., Karamanoglu, M., and Khan, N. (2013). True global optimality of the pressure vessel design problem: a benchmark for bio-inspired optimisation algorithms. *International Journal of Bio-Inspired Computation*, 5(6):329–335.

Yaz, E. (1987). A control scheme for a class of discrete nonlinear stochastic systems. *IEEE Transactions on Automatic Control*, 32(1):77–80.

Yuan, Q. (2008). A model free automatic tuning method for a restricted structured controller by using simultaneous perturbation stochastic approximation. In *Proceedings of the American Control Conference*, pages 1539–1545, Seattle, WA.

Zabinsky, Z. B. (2013). *Stochastic Adaptive Search for Global Optimization*, volume 72. Springer Science & Business Media.

Zhao, X. and Spall, J. C. (2016). Estimating travel time in urban traffic by

modeling transportation network systems with binary subsystems. In *Proceedings of the American Control Conference*, pages 803–808, Boston, MA.

Zhou, D. H., He, X., Wang, Z., Liu, G.-P., and Ji, Y. D. (2012). Leakage fault diagnosis for an internet-based three-tank system: an experimental study. *IEEE Transactions on Control Systems Technology*, 20(4):857–870.

Zhou, D.-X. (2020). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794.

Zhou, X. (2018). On the Fenchel duality between strong convexity and Lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*.

Zhou, Y.-L., Zhang, Q.-Z., Li, X.-D., and Gan, W.-S. (2008). On the use of an SPSA-based model-free feedback controller in active noise control for periodic disturbances in a duct. *Journal of Sound and Vibration*, 317(3-5):456–472.