SLIM: STOCHASTIC LEARNING AND INFERENCE IN OVERIDENTIFIED MODELS

By

Xiaohong Chen, Min Seong Kim, Sokbae Lee,

Myung Hwan Seo and Myunghyun Song

October 2025

COWLES FOUNDATION DISCUSSION PAPER NO. 2472

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

arXiv:2510.20996v1 [econ.EM] 23 Oct 2025

SLIM: Stochastic Learning and Inference in Overidentified Models*

Xiaohong Chen† Min Seong Kim‡ Sokbae Lee§
Myung Hwan Seo¶ Myunghyun Song‖

October 2025

Abstract

We propose SLIM (Stochastic Learning and Inference in overidentified Models), a scalable stochastic approximation framework for nonlinear GMM. SLIM forms iterative updates from independent mini-batches of moments and their derivatives, producing unbiased directions that ensure almost-sure convergence. It requires neither a consistent initial estimator nor global convexity and accommodates both fixed-sample and random-sampling asymptotics. We further develop an optional second-order refinement and inference procedures based on random scaling and plug-in methods, including plug-in, debiased plug-in, and online versions of the Sargan–Hansen J-test tailored to stochastic learning. In Monte Carlo experiments based on a nonlinear EASI demand system with 576 moment conditions, 380 parameters, and $n = 10^5$, SLIM solves the model in under 1.4 hours, whereas full-sample GMM in Stata on a powerful laptop converges only after 18 hours. The debiased plug-in J-test delivers satisfactory finite-sample inference, and SLIM scales smoothly to $n = 10^6$.

Keywords: Stochastic approximation, generalized method of moments, overidentification test, U-statistics, mini-batch, debiasing, random scaling

*We thank participants at conferences and workshops for helpful comments and suggestions.

†Department of Economics, Yale University and Cowles Foundation for Research in Economics. Email: xiaohong.chen@yale.edu

‡Department of Economics, University of Connecticut. Email: min_seong.kim@uconn.edu

§Department of Economics, Columbia University and Centre for Microdata Methods and Practice. Email: sl3841@columbia.edu

¶Department of Economics, Seoul National University. Email: myunghseo@snu.ac.kr

‖Department of Economics, Columbia University. Email: ms6347@columbia.edu

1 Introduction

Modern machine learning methods have revolutionized data analysis across numerous disciplines. Stochastic approximation, most prominently stochastic gradient descent (SGD), has emerged as a foundational tool behind the success of the recent deep learning revolution. First introduced by Robbins and Monro (1951), SGD is celebrated for its computational simplicity and scalable online implementation. It has become indispensable for modern optimization, particularly in settings involving a large number of parameters, huge datasets, and complex nonlinear objective functions (see, e.g., Bottou et al., 2018, for a review). By iteratively approximating gradients using random samples, SGD achieves substantial computational efficiency, making it highly effective for large-scale optimization problems that arise in contemporary data analysis.

SGD has been widely applied in the context of M-estimation, where the parameter of interest θ_o is defined as the minimizer of a population loss function $\mathbb{E}[\ell(z_i, \theta)]$, with $\ell(z_i, \theta)$ a known real-valued function of the observation z_i and parameter θ. The canonical form of the algorithm is

$$\theta_t = \theta_{t-1} - \gamma_t \frac{\partial}{\partial \theta} \ell(z_t, \theta_{t-1}), \quad \text{with a learning rate } \gamma_t \downarrow 0, \tag{1}$$

where t indexes the iteration count. This approach has proven effective for solving large-scale optimization problems, particularly in machine learning and statistics.

However, the application of SGD in econometrics remains limited. A central reason is that many parameters of interest in economics are not defined as minimizers of objective functions, but rather as solutions to systems of moment conditions:

$$\mathbb{E}[g(z_i, \theta_o)] = 0,$$

where $g(z_i, \theta)$ is a vector-valued function with $\dim(g) \geq \dim(\theta)$. This setting includes overidentified models, where $\dim(g) > \dim(\theta)$, and is typically addressed using the generalized method of moments (GMM) introduced by Hansen (1982). Adapting SGD to such moment-based estimation problems poses significant challenges because the standard notion of a stochastic gradient does not directly apply. This reflects a fundamental methodological distinction between modern scalable optimization techniques and classical GMM estimation frameworks in econometrics.

To illustrate these differences, consider the standard GMM setup. Let $\theta_o \in \Theta \subset \mathbb{R}^d$

denote the true parameter vector, where Θ is the parameter space. Let $z_{1:n} := (z_i)_{i=1}^n$ be a sample of size n, and let $g(z, \theta)$ be a vector of moment functions such that $\theta \mapsto g(z, \theta)$ is differentiable for each z. Define the sample averages of the moment function and its Jacobian as:

$$\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta), \qquad \bar{G}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(z_i, \theta)}{\partial \theta'}. \tag{2}$$

The standard GMM estimator is then given by

$$\hat{\theta}_{n,W_n} := \arg\min_{\theta \in \Theta} \bar{g}_n(\theta)' W_n \bar{g}_n(\theta),$$

where W_n is a possibly data-dependent weighting matrix. When $g(z, \theta)$ is nonlinear in θ, computing $\hat{\theta}_{n,W_n}$ typically requires iterative numerical methods.

Establishing convergence guarantees for algorithms that compute nonlinear GMM estimators has remained a longstanding challenge, particularly in settings where the objective function is nonconvex or a consistent initial estimate is difficult to obtain without exhaustive search. Robinson (1988) analyzes the use of grid search to construct a consistent initial estimator. Given such an estimator $\hat{\theta}_0$, applying t iterations of a Newton–Raphson (NR) procedure yields a refined estimator $\hat{\theta}_t$ that is consistent and asymptotically normal, provided the initial estimator is sufficiently accurate. In Robinson's framework, this initial estimator is obtained through a global search over a regular grid, which becomes computationally infeasible in large-scale problems. Building on this idea, Andrews (1997) proposes a stopping rule to obtain a consistent first-step estimator, followed by a fixed number of NR updates. While this method avoids pre-specifying grid resolution, it still requires a global search to verify the stopping criterion, which limits its scalability. More recently, Forneron and Zhong (2025) study optimization methods for nonlinear GMM without assuming convexity or requiring a consistent initial estimator, focusing on iterative procedures such as gradient descent and Gauss–Newton.

All of the above methods rely on full-sample evaluations of the moment function, its derivative, and related quantities at every update. However, in large-scale settings with very large n, computing the full-sample estimator $\hat{\theta}_{n,W_n}$ becomes computationally burdensome or even infeasible because it requires evaluating $\bar{g}_n(\theta)$ and $\bar{G}_n(\theta)$ over the entire dataset in each step. Moreover, the usual form of SGD in (1) does not directly

apply, since the gradient of the GMM objective function involves a product of two sample averages: $2\bar{G}_n(\theta)'W_n\bar{g}_n(\theta)$. Consequently, neither a single observation nor a single mini-batch yields an unbiased estimate of this gradient.

To address these challenges, we propose a stochastic approximation method that updates the parameter vector using independent mini-batches of the moment function and its Jacobian. The construction is analogous to a U-statistic, yielding unbiased estimates of updating directions that form a martingale difference sequence. This structure permits the use of martingale convergence theory to establish almost sure convergence, while providing substantial scalability advantages for large-scale nonlinear GMM problems.

There are a couple of related papers in the literature. Chen et al. (2025) extend stochastic approximation to GMM, referred to as SGMM, and develop methods for online inference in the context of linear instrumental variable (IV) regression, where $g(z_i, \theta)$ is linear in θ. SGMM employs a second-order optimization method, in the terminology of Bottou et al. (2018, Section 6), and achieves asymptotic efficiency comparable to two-stage least squares (2SLS) and GMM. However, SGMM faces notable limitations when applied to nonlinear GMM: it requires a consistent initial estimator, and its second-order nature limits scalability relative to first-order methods. As a related approach, Leung et al. (2025) develop an online generalized method of moments (OGMM), designed for stationary and ergodic time series data that arrive in batches. They assume that the initial batch size grows to infinity and that a consistent GMM estimator can be constructed from this batch. Moment restrictions are subsequently updated using a first-order Taylor approximation. Although their method accommodates time series dependence, it still relies on the availability of a consistent initial estimator from a large batch.[1]

A key unpleasant assumption in both SGMM and OGMM for nonlinear moment restrictions is the requirement of an initial consistent estimator for θ_o. To address this issue, we develop SLIM (Stochastic Learning and Inference in overidentified Models), which is a modular stochastic approximation framework for nonlinear GMM that

[1]The literature on stochastic approximation for overidentified moment restrictions is sparse. Beyond Chen et al. (2025) and Leung et al. (2025), there are a few papers on online linear IV regressions in computer science and engineering communities: see, for example, Venkatraman et al. (2016), Chen et al. (2024), Della Vecchia and Basu (2025) and the references therein. None of these works consider the asymptotic distributions of their online linear IV estimators and statistical inference procedures, however.

includes a required first-order algorithm and an optional (but recommended) second-order algorithm that can be applied subsequently to improve efficiency. Notably, our procedure does not require an initial consistent estimator. Inference can be conducted using either a random scaling approach, which is preferred when loading the full dataset is computationally expensive since the random scaling term is updated within either the first-order or second-order algorithm, or a plug-in approach, which is preferred when full-sample access is feasible.

Furthermore, we address the challenge of instability in single-pass stochastic approximation, where each observation is accessed only once, by formally establishing asymptotic theory for multi-pass stochastic approximation with mini-batches. This greatly enhances both the stability and accuracy of the method. While Chen et al. (2025) demonstrate the empirical benefits of multi-pass strategies in linear IV regression, they do not provide formal theoretical guarantees, which we develop in this paper.

The organization of the paper is as follows. Section 2 introduces the setting and presents our proposed SLIM algorithm. It provides a high-level roadmap that outlines the procedure and offers a heuristic explanation of how and why stochastic approximation works in the context of GMM. It also includes a discussion of related methods in the literature. Section 3 develops asymptotic theory under a random-sampling framework that incorporates both stochastic approximation and sampling uncertainty. Building on this theory, Section 4 proposes random scaling methods for statistical inference. Section 5 extends our approach to efficient estimation via second-order stochastic approximation and introduces plug-in inference methods. We compare random scaling and plug-in inference methods, showing that both are asymptotically valid but suited to different computational regimes. In addition, we extend the Sargan-Hansen J-test to plug-in, debiased plug-in, and online versions tailored to stochastic learning. Section 6 presents the results of Monte Carlo experiments that illustrate our methodology using the Exact Affine Stone Index (EASI) demand model developed by Lewbel and Pendakur (2009). In the EASI system, budget shares depend on implicit utility, which is a nonlinear function of unknown model parameters and is endogenous due to its dependence on the budget shares via the Stone index. We estimate these parameters using nonlinear GMM, designing a data-generating process based on the dataset constructed by Lewbel and Pendakur (2009) and scaling up the sample size in our experiments. To evaluate both computational and inferential

performance, we compare standard full-sample nonlinear GMM with our proposed methods. When the sample size is 100,000, the number of moment conditions is 576, and the number of unknown parameters is 380, Stata's conventional GMM routine succeeds in computing the estimator and its confidence set only after 18 hours on a high-performance laptop. In contrast, our efficient stochastic approximation methods complete the same task in under 1.4 hours, demonstrating substantial gains in scalability and practical feasibility for large-scale applications. We also demonstrate that the debiased plug-in J-test delivers satisfactory finite-sample inference, and SLIM scales efficiently as the sample size increases to one million. Finally, Section 7 provides concluding remarks and outlines directions for future research. We present the proof of consistency immediately following the conclusion, as establishing the consistency of SLIM is of fundamental importance given that the algorithm does not require initialization with a consistent preliminary estimator. Appendix A develops an alternative mode of asymptotic theory and inference under a fixed-sample framework that captures the error due to stochastic approximation. Appendix B introduces a warm-start algorithm that can be used as an optional initialization step for SLIM. Appendix C offers additional discussion that complements Section 4. Appendix D contains all proofs omitted from the main text.

2 SLIM: Stochastic Learning for Nonlinear GMM

We begin by providing a roadmap for stochastic approximation in nonlinear GMM. Figure 1 outlines the recommended procedure. The roadmap emphasizes that an initial consistent estimator is not required, regardless of whether a warm start (described in Appendix B) is used. The first main step, which is required, applies a first-order approach based on U-statistics (see the current section for details). The second step, which is optional but recommended, refines this estimator using a second-order update to improve efficiency (see Section 5). The procedure branches into two alternative inference methods: the random scaling approach, described in Section 4, preferred when loading the full dataset is computationally expensive; and the plug-in approach, described in Section 5, preferred when full-sample access is feasible.

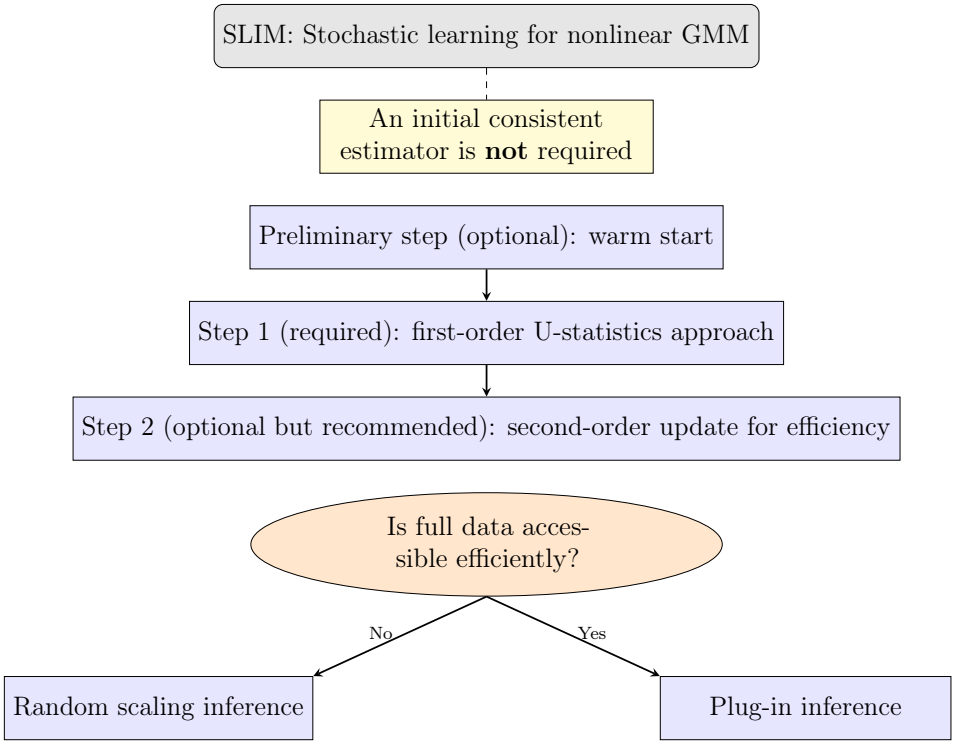

Figure 1: Outline for SLIM

2.1 Setting

In our setting, computational efficiency is as important as statistical efficiency. Thus, for the moment, we assume that the weighting matrix is fixed and positive-definite but not necessarily optimal. To reduce computational costs, we use stochastic approximation, where θ is iteratively updated using randomly selected subsamples of the data. Specifically, at each iteration, two independent mini-batches of observations are used to compute unbiased estimates of $\bar{g}_n(\theta)$ and $\bar{G}_n(\theta)$. We show that this approach significantly reduces the computational burden while remaining consistent and asymptotically normal under mild regularity conditions, even though it may sacrifice statistical efficiency compared to the fully efficient GMM estimator. We return to the issue of efficiency and optimal weighting in Section 5.

Since the weighting matrix W_n remains fixed throughout the iterations, we further simplify the optimization objective by setting W_n to the identity matrix, without loss

of generality.[2] This leads to the following simplified formulation of the estimator:

$$\hat{\theta}_n := \arg\min_{\theta \in \Theta} \bar{g}_n(\theta)' \bar{g}_n(\theta). \tag{3}$$

The following subsection details our stochastic approximation.

2.2 Algorithm

We begin by specifying the input for our stochastic approximation algorithm. Let $\theta_0^* \in \Theta$ denote the initial value of the parameter vector, which is computed based on the information in $\mathcal{F}_n := \sigma(z_{1:n})$ and remains bounded as n increases. For example, θ_0^* can be a vector of predetermined constants (e.g., zeros) or a computationally feasible estimator derived from a random subsample of $z_{1:n}$ when n is very large. A key advantage of our approach is that consistency and asymptotic normality of the final estimator do not depend on the initial estimator being consistent, although the quality of the starting value can have a non-negligible impact on finite-sample performance.

In addition, the algorithm takes as input the total number of iterations N, the learning rate schedule $\{\gamma_t : t = 1, \ldots, N\}$, and the mini-batch sizes B_G and B_g, where $B = B_G + B_g$ denotes the total mini-batch size. These quantities are predetermined before the algorithm begins. Let $\{\tilde{z}_j : j = 1, \ldots, BN\}$ denote a sequence of i.i.d. draws sampled uniformly at random from $z_{1:n}$. That is, for $j = 1$ to BN, sample $i(j)$ uniformly from $\{1, \ldots, n\}$ and let $\tilde{z}_j = z_{i(j)}$. This sequence provides the mini-batches of data used for the stochastic updates. When n is relatively large, we set $BN = n$; otherwise, we allow $BN \geq n$, which is often referred to as *multi-pass* algorithms. When n is relatively small, multi-pass algorithms can improve the finite-sample performance as they allow the algorithm to iterate over the data multiple times.

At each iteration $t \geq 1$, the parameter vector is updated using a mini-batch gradient descent step, which can be expressed as:

$$\theta_t^* = \theta_{t-1}^* - \gamma_t \tilde{G}_t(\theta_{t-1}^*)' \tilde{g}_t(\theta_{t-1}^*), \tag{4}$$

[2] It is implicitly assumed that it is computationally straightforward to compute $W_n^{1/2}$ and pre-multiply it to $g(z_i, \theta)$, so that we can re-define $g(z_i, \theta)$ to be $W_n^{1/2} g(z_i, \theta)$.

where

$$\tilde{G}_t(\theta) := \frac{1}{B_G}\sum_{i=1}^{B_G} G(\tilde{z}_{(t-1)B+i}, \theta) \quad \text{and} \quad \tilde{g}_t(\theta) := \frac{1}{B_g}\sum_{i=1}^{B_g} g(\tilde{z}_{(t-1)B+B_G+i}, \theta). \tag{5}$$

In this update, conditional on θ^*_{t-1}, the derivative matrix $\tilde{G}_t(\theta^*_{t-1})$ serves as an unbiased estimate of the full-sample derivative matrix $\bar{G}_n(\theta^*_{t-1})$, while the moment vector $\tilde{g}_t(\theta^*_{t-1})$ is an unbiased estimate of the full-sample moment vector $\bar{g}_n(\theta^*_{t-1})$. By construction, $\tilde{G}_t(\theta^*_{t-1})$ and $\tilde{g}_t(\theta^*_{t-1})$ are independent of each other conditional on θ^*_{t-1}. On one hand, the use of independent mini-batches for these two components ensures that the stochastic updates remain unbiased conditional on θ^*_{t-1}, while significantly reducing the computational burden relative to full-sample evaluations. On the other hand, mini-batches also provide variance reduction compared to stochastic approximation with $B_G = B_g = 1$. This variance reduction is crucial in finite samples, as each iteration involves a product of two noisy elements, despite both being unbiased and independent of each other.

At each iteration, the Polyak (1990)-Ruppert (1988) average is computed:

$$\bar{\theta}^*_t = \frac{1}{t}\theta^*_t + \frac{t-1}{t}\bar{\theta}^*_{t-1}$$

and our proposed estimator is $\bar{\theta}^*_N$. The averaging step helps mitigate noise in the stochastic updates. Our proposed estimator is consistent and asymptotically normal under mild regularity conditions as both n and N tend to infinity. Furthermore, we show how to conduct inference using the stochastic path $\{\theta^*_t : t = 1, \ldots, N\}$. In summary, the proposed first-order method is outlined in Algorithm 1.

3 Asymptotic Theory

We begin by stating the basic assumptions that underpin our analysis.

Assumption 1 (Uniform Sampling with Replacement)**.** *Let $\{\tilde{z}_j\}_{j=1}^{BN}$ denote a sequence of i.i.d. draws from $z_{1:n}$.*

Assumption 2 (Learning Rates)**.** *The sequence of learning rates satisfies $\gamma_t = \gamma_0 t^{-a}$ for $t \geq 1$, where $\gamma_0 > 0$ and $a \in (1/2, 1)$.*

Algorithm 1: First-Order U-statistic Approach

Input: Sample $z_{1:n} = (z_i)_{i=1}^n$; mini-batch sizes B_G and B_g; number of iterations N; schedule of learning rates $(\gamma_t)_{t=1}^N$; initial value θ_0^*.

Set $t \leftarrow 1$ and $B \leftarrow B_G + B_g$.
Initialize $\theta^* \leftarrow \theta_0^*$.
Initialize $\bar{\theta}^* \leftarrow \theta_0^*$.
while $t \leq N$ **do**
  Sample $i(1), \ldots, i(B)$ uniformly at random from $\{1, \ldots, n\}$.
  Compute $\tilde{G} = \frac{1}{B_G} \sum_{j=1}^{B_G} G(z_{i(j)}, \theta^*)$ and $\tilde{g} = \frac{1}{B_g} \sum_{j=B_G+1}^{B} g(z_{i(j)}, \theta^*)$.
  Update $\theta^* \leftarrow \theta^* - \gamma_t \tilde{G}' \tilde{g}$.
  Update $\bar{\theta}^* \leftarrow \frac{t-1}{t} \bar{\theta}^* + \frac{1}{t} \theta^*$.
  Set $t \leftarrow t + 1$.
Set $\theta_N^* \leftarrow \theta^*$ and $\bar{\theta}_N^* \leftarrow \bar{\theta}^*$.

Output: Final stochastic update θ_N^*; final estimator $\bar{\theta}_N^*$.

Assumption 3 (Smoothness of the Moment Function)**.** *There exists a set $\mathcal{Z}$ such that $z_{1:n} \subseteq \mathcal{Z}$ and, for all $z \in \mathcal{Z}$, $g(z, \theta)$ is twice continuously differentiable in $\theta \in \Theta$.*

The sequence $\{\tilde{z}_j\}_{j=1}^{BN}$ is generated by uniform sampling with replacement from $z_{1:n}$, and is partitioned into N mini-batches of size B. When $BN = c \cdot n$ for some $c > 1$, our algorithm, as described in Section 2.2, samples each data point c times on average. This framework is commonly referred to as a *multi-pass algorithm*. An alternative sampling strategy is discussed in Appendix B.

The learning-rate sequence $(\gamma_t)_{t=1}^N$ is assumed to decay at the rate $O(t^{-a})$ for some exponent $a \in (1/2, 1)$. This ensures $\sum_{t=1}^{\infty} \gamma_t = \infty$ and $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$, which are standard conditions guaranteeing convergence of the stochastic gradient-descent algorithm. We also assume the moment function $g(z, \theta)$ is twice differentiable in θ, as is customary in the nonlinear GMM setting.

We develop asymptotic theory under the i.i.d. sampling. Extensions of this framework to settings with time series or cluster dependence are left for future research.[3] To state the further regularity conditions, we introduce additional notation.

[3] In Appendix A.1, we consider an alternative asymptotic framework in which the data are treated as fixed, and the objective is to approximate the full-sample GMM estimator $\hat{\theta}_n$, defined in (3). This setting is particularly relevant when evaluating $\hat{\theta}_n$ is computationally intensive, such as when the sample size n is very large.

Define the sample Lyapunov function

$$\bar{Q}_n(\theta) := \bar{g}_n(\theta)'\bar{g}_n(\theta) - \bar{g}_n(\hat{\theta}_n)'\bar{g}_n(\hat{\theta}_n), \tag{6}$$

where $\bar{g}_n(\theta)$ and $\hat{\theta}_n$ are given in (2) and (3), respectively. The population moment function and its Jacobian are also denoted by

$$g(\theta) = \mathbb{E}[g(z_i, \theta)] \quad \text{and} \quad G(\theta) = \mathbb{E}[G(z_i, \theta)].$$

Let I_d denote the d-dimensional identity matrix. For square matrices A and B, we write $A \leq B$, $A \geq 0$, and $A > 0$ if $B - A$ is positive semidefinite, A is positive semidefinite, and positive definite, respectively. We use $\|\cdot\|$ to denote the Euclidean norm on vectors and the spectral norm on matrices.

Assumption 4 (Random Sampling and Regularity Conditions)**.** *The following conditions hold for some constants $c > 0$, $\delta > 0$, $M < \infty$, and $p \geq 1$.*

(A4.1) The observations $(z_i)_{i=1}^n \subseteq \mathcal{Z}$ are i.i.d. draws from the population distribution $\mathbb{P}$.

(A4.2) Θ is a convex set in $\mathbb{R}^d$ and $g(\theta) = 0$ if and only if $\theta_o \in \Theta$.

(A4.3) $G_o'G_o \geq cI_d$ and $\Omega_o \geq cI_d$, where $G_o := G(\theta_o)$ and $\Omega_o := \mathbb{E}[g(z_i, \theta_o)g(z_i, \theta_o)']$.

(A4.4) $\mathbb{E}[\|g(z_i, \theta_o)\|^{2p}] < \infty$, and there exists a measurable function $H : \mathcal{Z} \to \mathbb{R}$ such that $\|G(z, \theta)\| \leq H(z)$ for all $z \in \mathcal{Z}$ and $\theta \in \Theta$, and $\mathbb{E}[H(z_i)^{2p}] < \infty$.

(A4.5) The map $\theta \mapsto G(z, \theta)$ is Lipschitz for each $z \in \mathcal{Z}$. That is, $\|G(z, \theta) - G(z, \tilde{\theta})\| \leq L(z)\|\theta - \tilde{\theta}\|$ for all θ and $\tilde{\theta} \in \Theta$, where $L : \mathcal{Z} \to \mathbb{R}$ satisfies $\mathbb{E}[L(z_i)^{2p}] < \infty$.

(A4.6) For all $\|\theta - \hat{\theta}_n\| \leq \delta$, $\frac{\partial^2 \bar{Q}_n(\theta)}{\partial\theta\partial\theta'} \geq cI_d$ w.p.a.1.

(A4.7) For all $\|\theta - \hat{\theta}_n\| \geq \delta$, $\bar{Q}_n(\theta) \geq c$ and $\|\bar{G}_n(\theta)'\bar{g}_n(\theta)\|^2 \geq c$ w.p.a.1.

(A4.8) For all $\theta \in \Theta$, $\frac{\partial^2 \bar{Q}_n(\theta)}{\partial\theta\partial\theta'} \leq MI_d$ and $\frac{1}{n}\sum_{i=1}^n \|g(z_i, \theta)\|^{2p} \leq M(\bar{Q}_n(\theta)^p + 1)$ w.p.a.1.

The bulk of Assumption 4 reflects standard regularity requirements under a random-sampling framework. Condition (A4.1) assumes that the data are i.i.d. Condition (A4.2) imposes global point identification of the true parameter θ_o, while (A4.3) ensures that θ_o is locally well-separated via a rank condition on the Jacobian $G(\theta_o)$ and the covariance matrix Ω_o. Condition (A4.4) imposes integrability on $g(\cdot, \theta_o)$ and

the envelope function for the Jacobian $G(\cdot, \theta)$. Condition (A4.5) requires Lipschitz continuity of $\theta \mapsto G(z, \theta)$, with an integrable Lipschitz constant $L(z)$. Condition (A4.8) assumes the existence of a constant M such that the difference between MI_d and the Hessian of the sample Lyapunov function $\bar{Q}_n(\theta)$ is uniformly positive semidefinite, and that the ratio of $n^{-1}\sum_{i=1}^{n}\|g(z_i,\theta)\|^{2p}$ to $\bar{Q}_n(\theta)^p + 1$ is uniformly bounded by M, both with probability approaching one. This condition is used to control the variability of the stochastic errors.

It remains to discuss Conditions (A4.6) and (A4.7), which impose requirements on the behavior of the sample Lyapunov function in the regions where $\|\theta - \hat{\theta}_n\| \leq \delta$ and $\|\theta - \hat{\theta}_n\| \geq \delta$, respectively. The local lower bound in (A4.6) ensures that $\bar{Q}_n(\theta)$ is strongly convex in a neighborhood of $\hat{\theta}_n$. In turn, (A4.7) requires that both $\bar{Q}_n(\theta)$ and the squared norm of the gradient of the sample Lyapunov function, $\|\bar{G}_n(\theta)'\bar{g}_n(\theta)\|^2$, are bounded away from zero for all θ such that $\|\theta - \hat{\theta}_n\| \geq \delta$. Notably, we do not impose global convexity or compactness of the parameter space; instead, convergence is ensured by the two-region structure encoded in (A4.6) and (A4.7). These two conditions jointly serve a role analogous to global strong convexity in ensuring convergence of stochastic approximation algorithms. Condition (A4.6) provides local curvature near the solution, ensuring that stochastic iterates are attracted to the unique minimizer $\hat{\theta}_n$ eventually. Meanwhile, Condition (A4.7) prevents the algorithm from stagnating far from the solution by requiring that the gradient magnitude remains bounded away from zero outside the neighborhood. This two-region structure, consisting of local strong convexity near the optimum and a sharp separation condition away from it, allows us to establish stability and convergence without relying on global convexity or compactness. Such an approach is particularly well suited to overidentified nonlinear GMM settings.

Remark 1 (Additional Assumptions Beyond Standard GMM Theory). Conditions (A4.6), (A4.7), and (A4.8) are not part of the standard asymptotic analysis of GMM estimators. Classical econometric theory implicitly assumes that GMM estimators can be computed exactly or that any computational errors are asymptotically negligible relative to sampling variability. In our setting, we explicitly account for both statistical and computational errors, which necessitates these additional regularity conditions. These conditions are formulated in terms of the sample Lyapunov function to better separate computational errors from sampling variability. For i.i.d. data, analogous conditions can also be formulated in terms of population quantities under suitable

uniform convergence arguments via the uniform law of large numbers, though we omit these for brevity.

Remark 2 (Comparison with Forneron and Zhong (2025)). Our assumptions also differ from those in Forneron and Zhong (2025). Although their analysis does not require convexity of the GMM criterion function, it relies on full-rank conditions on $G(\theta)$, $G(\theta_1)'G(\theta_2)$, and certain integral transforms of $G(\theta)$, holding globally for all $\theta, \theta_1, \theta_2 \in \Theta$, along with other regularity requirements. As a consequence (Proposition 2 in their paper), these conditions are effectively equivalent to requiring that the population GMM objective function $Q(\theta)$ be globally lower bounded by a quadratic function and, in addition, satisfy the Polyak–Łojasiewicz inequality:

$$\left\| \frac{\partial Q(\theta)}{\partial \theta} \right\|^2 \geq \mu \left(Q(\theta) - Q(\theta_o) \right),$$

for all $\theta \in \Theta$ and some constant $\mu > 0$.

To analyze the stochastic properties of our algorithm, it is useful to decompose $(\bar{\theta}_N^* - \theta_o)$ into two parts:

$$\bar{\theta}_N^* - \theta_o = \underbrace{\left(\bar{\theta}_N^* - \hat{\theta}_n\right)}_{\text{computational error}} + \underbrace{\left(\hat{\theta}_n - \theta_o\right)}_{\text{statistical error}},$$

where the first term captures the discrepancy arising from stochastic approximation relative to the full-sample GMM estimator, and the second reflects the estimation error of the full-sample GMM estimator relative to the true parameter vector. In econometrics, the conventional analysis focuses only on the statistical error, implicitly assuming that the computational error is asymptotically negligible. Intuitively, this decomposition separates the error from using a stochastic algorithm (computational) from the inherent sampling uncertainty in estimation (statistical).

We now introduce some notation. Let $\mathbb{P}_n^*$ denote the distribution induced by the mini-batch sample $(\tilde{z}_j)_{j \geq 1}$, and let $\mathbb{E}_n^*$ denote the corresponding expectation. With this in place, we are ready to state our first theoretical result.

Theorem 1 (Consistency)**.** *Let Assumptions 1, 2, 3, and 4 hold. Then, θ_N^* and $\bar{\theta}_N^*$ are consistent for θ_o as $n \to \infty$ and $N \to \infty$. Furthermore, there exists a sequence of*

events $E_n \in \mathcal{F}_n$ *such that* $\mathbb{P}(E_n) \to 1$*, and for every* $\varepsilon > 0$*,*

$$\sup_{n\in\mathbb{N}} \mathbb{P}_n^* \left(\|\theta_N^* - \hat{\theta}_n\| \geq \varepsilon \right) \mathbb{1}_{E_n} \xrightarrow{\mathbb{P}} 0 \quad \text{as } N \to \infty.$$

In other words, the probability of the event $\{\|\theta_N^* - \hat{\theta}_n\| \geq \varepsilon\}$ converges to zero uniformly in $n \in \mathbb{N}$ as the number of iterations increases, indicating that $\|\theta_N^* - \hat{\theta}_n\|$ decays at a rate depending on N but not on the sample size n. Since $\mathbb{E}_n^* \bar{Q}_n(\theta_N^*)$ decays at rate $O(\gamma_N) = O(N^{-a})$ with probability approaching 1, the convergence rate of θ_N^* toward $\hat{\theta}_n$ is $O(N^{-a/2})$; see Lemma 4 in the appendix.

The standard parametric rate $O(N^{-1/2})$ can be achieved by averaging over the stochastic path, yielding the averaged estimator $\bar{\theta}_N^* = N^{-1} \sum_{t=1}^N \theta_t^*$ with reduced variance. This estimator also satisfies the (martingale) central limit theorem (CLT), as established in Theorem 2. Its asymptotic variance closely matches that of the full-sample GMM estimator, with an additional term arising from the uncertainty in the estimation of $G(\theta)$. To describe it precisely, let us introduce the following matrix:

$$\Sigma_o = \frac{1}{B_g} \left(G_o' \Omega_o G_o + \frac{1}{B_G} \mathbb{E}\Big[(G(z_i, \theta_o) - G(\theta_o))' \Omega_o (G(z_i, \theta_o) - G(\theta_o)) \Big] \right), \tag{7}$$

which represents $\operatorname{plim}_{n\to\infty} \bar{\Sigma}_n$, the asymptotic variance of the stochastic error, where $\bar{\Sigma}_n$ is a sample analog of Σ_o and formally defined in the appendix (see (31)). The second term reflects the variability in the mini-batch approximation of $G(\theta)$, which decays at a rate inversely proportional to the corresponding mini-batch size. The next theorem presents the CLT applied to the averaged estimator.

Theorem 2. *Let Assumptions 1, 2, 3, and 4 hold with* $p > 1$*. Then, as* $n \to \infty$*,* $N \to \infty$*, and* $N^{1-a}/n \to 0$*, it holds that*

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta_o) \\ \sqrt{N}(\bar{\theta}_N^* - \hat{\theta}_n) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} (G_o'G_o)^{-1} G_o' \Omega_o G_o (G_o'G_o)^{-1} & 0 \\ 0 & (G_o'G_o)^{-1} \Sigma_o (G_o'G_o)^{-1} \end{pmatrix} \right).$$

Theorem 2 establishes that $\sqrt{n}(\hat{\theta}_n - \theta_o)$ and $\sqrt{N}(\bar{\theta}_N^* - \hat{\theta}_n)$ are jointly asymptotically normal and asymptotically independent. While the convergence result for the full-sample estimator $\hat{\theta}_n$ is standard, the result for the averaged stochastic estimator $\bar{\theta}_N^*$ is new. It shows that $\bar{\theta}_N^*$, centered at $\hat{\theta}_n$, is asymptotically normal, achieves the $O(N^{-1/2})$ convergence rate, and has an asymptotic covariance matrix whose "meat" component

includes an additional term

$$B_G^{-1}\,\mathbb{E}\big[(G(z_i,\theta_o)-G(\theta_o))'\Omega_o(G(z_i,\theta_o)-G(\theta_o))\big], \tag{8}$$

which captures the noise introduced by approximating the Jacobian with finite mini-batches. This variance inflation diminishes as the mini-batch size B_G increases.

By leveraging global information from $(\bar{\theta}_t^*)_{t=1}^N$, it is possible to extract additional information about the asymptotic variance of the stochastic approximation estimator. This idea is formalized in the following theorem, which establishes the functional central limit theorem (FCLT) for the partial sum process of stochastic updates, generalizing Theorem 2 under a slightly stronger moment condition.

Throughout, let Z denote a standard normal vector and $W(r)$ a standard multivariate Wiener process on $[0,1]$, independent of Z. The Brownian Bridge is denoted by $\bar{W}(r) = W(r) - rW(1)$. When necessary to evade confusion, we specify their dimensions by adding a subscript, e.g., Z_d. Let $\rightsquigarrow$ denote the weak convergence in $\ell^\infty([0,1])$. Moreover, let χ^2_ℓ denote a chi-squared variable with ℓ degrees of freedom.

Theorem 3. *Let Assumptions 1, 2, 3, and 4 hold with $p > (1-a)^{-1}$. Then, as $n \to \infty$, $N \to \infty$, and $N^{1-a}/n \to 0$, it holds*

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta_o) \\ \frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_t^* - \hat{\theta}_n) \end{pmatrix} \rightsquigarrow \begin{pmatrix} (G_o'G_o)^{-1}(G_o'\Omega_o G_o)^{1/2} Z \\ (G_o'G_o)^{-1}\Sigma_o^{1/2} W(r) \end{pmatrix}.$$

Theorem 3 strengthens Theorem 2 by establishing weak convergence of the entire stochastic trajectory rather than only its terminal average. The limiting process is a multidimensional Wiener process, scaled by the same asymptotic variance as in the CLT. This result enables the development of inference procedures that utilize the full stochastic path without recomputing averages of the sample moment function or Jacobian over the entire dataset. As a result, inference can be conducted efficiently even in large-scale settings where loading the full dataset is computationally infeasible. In summary, Theorem 3 provides the theoretical foundation for the random scaling inference methods introduced in the next section.

We conclude this section by commenting on the restrictions imposed on p, a, and N by our asymptotic framework. To illustrate, consider the case $a = 0.501$, which is the value used in our Monte Carlo experiments. The moment condition $p > (1-a)^{-1}$ then requires $p > 2.004$, which is relatively mild. This reflects the fact that a slower

decay of the learning rate (i.e., smaller a) permits weaker moment conditions. At the same time, the requirement $N^{1-a}/n \to 0$ imposes a restriction on how fast the number of stochastic updates N can grow relative to the sample size n. For $a = 0.501$, this implies $N = o(n^{2.004})$, meaning that N cannot grow faster than approximately n^2. This example highlights the trade-off in choosing a: slower decay relaxes the required moment conditions but tightens the allowable growth rate of the number of iterations.

4 Random Scaling Inference

In this section, we develop methods for conducting inference on the true parameter vector θ_o. We focus on testing $\ell \leq d$ linear restrictions of the form

$$H_0 : R\theta_o = c,$$

where R is a known $(\ell \times d)$ matrix of rank ℓ, and c is a known vector in $\mathbb{R}^\ell$.

We consider two approaches to inference: one based on random scaling and the other on plug-in estimation of the asymptotic variance matrix. Random scaling inference has been used in two largely disconnected branches of the literature. In time series econometrics, it has long been employed to estimate the long-run variance, which is the sum of all autocovariances, since its introduction by Kiefer et al. (2000). More recently, Lee et al. (2022) demonstrated its usefulness for inference with SGD, after which it has been adopted in stochastic approximation methods across various estimation problems. For instance, Lee et al. (2025) applied it to quantile regression models, while Chen et al. (2025) used it for linear GMM estimation. Additional applications of random scaling can be found in Li et al. (2022); Chen et al. (2024); Du et al. (2025), among others. The plug-in method is deferred to Section 5, where it is naturally integrated with the efficient estimation procedure developed there.

4.1 Inference Based on Random Scaling

Define the $(\ell \times \ell)$ random scaling matrix by

$$V_t(R) := \frac{1}{t}\sum_{s=1}^{t}\left\{\frac{1}{\sqrt{t}}\sum_{j=1}^{s}(R\theta_j^* - R\bar{\theta}_t^*)\right\}\left\{\frac{1}{\sqrt{t}}\sum_{j=1}^{s}(R\theta_j^* - R\bar{\theta}_t^*)\right\}'. \tag{9}$$

This matrix admits the following iterative updating rule for each $t \geq 1$:

$$A_t(R) = A_{t-1}(R) + t^2 R\bar{\theta}_t^* \bar{\theta}_t^{*\prime} R', \tag{10}$$

$$b_t(R) = b_{t-1}(R) + t^2 R\bar{\theta}_t^*, \tag{11}$$

$$V_t(R) = t^{-2}\left(A_t(R) - R\bar{\theta}_t^* b_t(R)' - b_t(R)\bar{\theta}_t^{*\prime} R' + R\bar{\theta}_t^*\bar{\theta}_t^{*\prime} R' \cdot \frac{t(t+1)(2t+1)}{6}\right), \tag{12}$$

with initial values $A_0 = 0_{\ell\times\ell}$ and $b_0 = 0_\ell$. The Wald statistic based on $V_N(R)$ is asymptotically pivotal, as established by the FCLT in Theorem 3.

Computational advantage: The iterative structure in (10)–(12) enables inference based on random scaling to be implemented efficiently. Crucially, it avoids computing or storing full $d \times d$ matrices and instead requires updates and storage only at the reduced $\ell \times \ell$ level. This leads to substantial computational gains, particularly when the number of restrictions ℓ is much smaller than the parameter dimension d. For instance, in many applications we are interested in a single linear combination of the parameter vector, so that $\ell = 1$, making the gain especially pronounced.

4.2 Asymptotic Regimes

Since we can decompose $(\bar{\theta}_N^* - \theta_o)$ as the sum of $(\bar{\theta}_N^* - \hat{\theta}_n)$ and $(\hat{\theta}_n - \theta_o)$, the relative rates at which n and N diverge affects asymptotics. We consider four regimes: two intermediate cases and two polar extremes. The leading case is discussed below, and three other cases are in Appendix C.

4.2.1 Intermediate Case I. $N \asymp n$ and $B_G < \infty$

We begin with the scenario where N and n tend to infinity at the same rate. Define the Wald statistic as

$$T_{n,N} := \underbrace{(1 + n/(NB_g))^{-1}}_{\text{deflating factor}} \cdot \underbrace{n(R\bar{\theta}_N^* - c)'(B_g V_N(R))^{-1}(R\bar{\theta}_N^* - c)}_{\text{Wald stat. w. random scaling}} \tag{13}$$

$$= (n^{-1} + (NB_g)^{-1})^{-1}(R\bar{\theta}_N^* - c)'(B_g V_N(R))^{-1}(R\bar{\theta}_N^* - c). \tag{14}$$

Here, the scaling factor $(n^{-1} + (NB_g)^{-1})^{-1}$ accounts for the combined uncertainty from both stochastic approximation and random sampling. From the decomposition

$$R\bar{\theta}_N^* - c = R(\bar{\theta}_N^* - \hat{\theta}_n) + R(\hat{\theta}_n - \theta_0),$$

under $H_0 : R\theta_o = c$, and Theorem 2, the asymptotic variance of $R\bar{\theta}_N^* - c$ is given by

$$\begin{aligned} &N^{-1}R(G_o'G_o)^{-1}\Sigma_o(G_o'G_o)^{-1}R' + n^{-1}R(G_o'G_o)^{-1}G_o'\Omega_o G_o(G_o'G_o)^{-1}R' \\ &\leq (n^{-1} + (NB_g)^{-1})B_g R(G_o'G_o)^{-1}\Sigma_o(G_o'G_o)^{-1}R' \end{aligned}$$

because $B_g\Sigma_o \geq G_o'\Omega_o G_o$. Hence, we obtain the following corollary as a direct implication of Theorem 3.

Corollary 1. *Let the conditions of Theorem 3 hold. As $n \to \infty$, $N \to \infty$, and $N \asymp n$, it follows that under $H_0 : R\theta_o = c$,*

$$T_{n,N} \leqslant_1 Z' \left(\int_0^1 \bar{W}(r)\bar{W}(r)'dr \right)^{-1} Z,$$

where $\leqslant_1$ denotes first-order stochastic dominance.

In Corollary 1, we have the following distributional equivalence:

$$Z' \left(\int_0^1 \bar{W}(r)\bar{W}(r)'dr \right)^{-1} Z \overset{d}{=} W(1)' \left(\int_0^1 \bar{W}(r)\bar{W}(r)'dr \right)^{-1} W(1), \tag{15}$$

which follows from the fact that $V_N(R)$ is asymptotically a functional of the process $\bar{W}(r) = W(r) - rW(1)$, up to an unknown variance matrix. Since $W(1)$ is projected out, $\mathrm{Cov}(\bar{W}(r), W(1)) = 0$ and thus the process $\bar{W}(r)$ is independent of $W(1)$ due to the Gaussianity .

4.2.2 Universality of Asymptotic Validity in Inference via Random Scaling

Recall that the test statistic $T_{n,N}$ defined in (13) takes the form

$$T_{n,N} = \left(n^{-1} + (NB_g)^{-1}\right)^{-1} (R\bar{\theta}_N^* - c)'(B_g V_N(R))^{-1}(R\bar{\theta}_N^* - c).$$

The analysis across all asymptotic regimes considered in Section 4.2.1 and Appendix C shows that asymptotically valid inference is possible under any divergence rates of n

and N, by employing the following decision rule:

$$\text{Reject } H_0 : R\theta_o = c \text{ if } T_{n,N} > \text{cv},$$

where cv is a critical value that depends only on the rank of R and the nominal significance level. This general validity stems from the scaling factor

$$\tau_n := \left(n^{-1} + (NB_g)^{-1}\right)^{-1},$$

which adapts automatically to different asymptotic regimes. When $N/n \to 0$, we have $\tau_n/(NB_g) \to 1$; when $n/N \to 0$, we have $\tau_n/n \to 1$. As a result, the procedure accommodates both sources of uncertainty—stochastic approximation and random sampling—without requiring explicit knowledge of their relative magnitudes. If inference precision is a priority, it is advisable to increase both NB_g and B_G as much as the available computational budget allows.

4.3 Summary and Practical Implementation

We conclude this section with a summary of the random scaling inference procedure and practical guidance for its implementation. The test statistic $T_{n,N}$ enables valid inference across a wide range of asymptotic regimes by adapting to the relative growth of n and N. Its implementation is computationally efficient due to recursive formulas for $V_N(R)$, which require only $\ell \times \ell$ operations and avoid computing full $d \times d$ matrices. The method also bypasses the need to estimate the asymptotic variance explicitly, making it particularly attractive for large-scale problems.

For practical use, we recommend choosing mini-batch sizes B_G and B_g to balance statistical precision and computational cost. Increasing B_G improves efficiency but raises computational burden. Run Algorithm 1 for a sufficiently large number of iterations N, typically with $N \geq C \cdot n$ for some constant $C > 1$, unless n is extremely large. Inference can be conducted by adding the random scaling inference component to Algorithm 1. This procedure provides a unified and robust method for inference across different asymptotic regimes. The limiting distribution in (15) is pivotal and depends only on the number of restrictions ℓ. When $\ell = 1$, critical values $\text{cv}_{1-\alpha/2}$ are given in Abadir and Paruolo (1997, Table I). For instance, $\text{cv}_{1-\alpha/2} = 6.747$ when $\alpha = 0.05$, and $\text{cv}_{1-\alpha/2} = 5.323$ when $\alpha = 0.10$. The corresponding $(1-\alpha)$ confidence

interval for $R\theta_o$ is given by

$$\left[R\bar{\theta}_N^* \pm \mathrm{cv}_{1-\alpha/2} \cdot \sqrt{n^{-1} + (NB_g)^{-1}} \cdot (B_g V_N(R))^{1/2}\right].$$

For general $\ell > 1$, critical values are available from Kiefer et al. (2000, Table II).

5 Efficient Estimation

Although the stochastic approximation steps in (4) yield asymptotically normal estimators, they are generally not efficient. The potential gains from improving efficiency can be substantial, especially when the sample size n is not exceedingly large. To address this limitation, we propose optional (but recommended) refinement steps that incorporate second-order information.

Let $\bar{\theta}_N^*$ denote the averaged estimator after the N-th iteration. Compute:

$$\Phi_n := \frac{1}{n}\sum_{i=1}^{n} G\big(z_i, \bar{\theta}_N^*\big), \quad \text{and} \quad W_{\mathrm{MB}} := \left(\frac{B_g}{M_{\mathrm{MB}}}\sum_{\ell=1}^{M_{\mathrm{MB}}} \tilde{g}_\ell\big(\bar{\theta}_N^*\big)\,\tilde{g}_\ell\big(\bar{\theta}_N^*\big)'\right)^{\dagger}, \tag{16}$$

where W_{MB} is based on M_{MB} randomly sampled mini-batches, $\dagger$ denotes the generalized inverse, and M_{MB} is a tuning parameter that diverges as $n \to \infty$.

Let $B_{G,t}$ be an increasing and diverging sequence after the N-th iteration; for example, $B_{G,t} = B_{G,0} + \log(t-N)\cdot 1(t > N)$, where $B_{G,0}$ is the mini-batch size used in the first step. Define $B_t := \sum_{s=1}^{t}(B_{G,s} + B_g)$ as the total number of mini-batch draws up to iteration t. Choose a predetermined value $T > 0$ (e.g., $T = C \cdot n$ for some constant $C > 0$), and sample $\{\tilde{z}_j\}_{j=B(N+1)}^{B_T}$ uniformly at random from the original dataset $\{z_1, \ldots, z_n\}$. For $t = N+1$ to $t = T$, sequentially update:

$$\theta_t^* = \theta_{t-1}^* - \gamma_t \left(\Phi_n' W_{\mathrm{MB}} \Phi_n\right)^{\dagger} \tilde{G}_t(\theta_{t-1}^*)'\, W_{\mathrm{MB}}\, \tilde{g}_t(\theta_{t-1}^*), \tag{17}$$

$$\bar{\theta}_t^* = \frac{t-N-1}{t-N}\bar{\theta}_{t-1}^* + \frac{1}{t-N}\theta_t^*, \tag{18}$$

where $\tilde{G}_t(\theta)$ and $\tilde{g}_t(\theta)$ are re-defined as

$$\tilde{G}_t(\theta) := \frac{1}{B_{G,t}}\sum_{i=1}^{B_{G,t}} G(\tilde{z}_{B_{t-1}+i}, \theta), \quad \text{and} \quad \tilde{g}_t(\theta) := \frac{1}{B_g}\sum_{i=1}^{B_g} g(\tilde{z}_{B_{t-1}+B_{G,t}+i}, \theta), \tag{19}$$

replacing the definitions in (5). This procedure replaces lines 7 and 8 in Algorithm 1 with (17) and (18), respectively.

It is important to emphasize that Φ_n and W_{MB} are computed only once and reused throughout the refinement step. Since $\bar{\theta}_N^*$ is already a consistent estimator, there is no need to recompute $(\Phi_n' W_{\text{MB}} \Phi_n)^\dagger$ or W_{MB}. However, updating $\tilde{G}_t(\theta)$ and $\tilde{g}_t(\theta)$ at each iteration remains essential, as it leverages the Lyapunov-function-based convergence guarantees of the algorithm.

Choice of Φ_n and W_{MB} Alternative choices for Φ_n and W_{MB} are possible, as long as they satisfy the consistency conditions:

$$(\Phi_n' W_{\text{MB}} \Phi_n)^{-1} \xrightarrow{\mathbb{P}} (G_o' \Omega_o^{-1} G_o)^{-1}, \quad \text{and} \quad W_{\text{MB}} \xrightarrow{\mathbb{P}} \Omega_o^{-1}.$$

When the sample size n is relatively small, a natural alternative to W_{MB} is the full-sample analog:

$$W_n = \left(\frac{1}{n} \sum_{i=1}^{n} g(z_i, \bar{\theta}_N^*) g(z_i, \bar{\theta}_N^*)' \right)^\dagger.$$

By contrast, Φ_n is based on the full sample by default, since it does not involve matrix inversion and its mini-batch counterpart tends to yield similar results due to averaging. However, when n is too large for full-sample computation, one can approximate Φ_n using either of the following alternatives: (i) a simple average over a randomly selected subsample of the data, or (ii) an average over multiple randomly drawn mini-batches. Both approaches offer computational savings and yield similar results due to averaging.

Define the modified Lyapunov function,

$$\bar{Q}_{n,W}(\theta) := \bar{g}_n(\theta)' W_{\text{MB}} \bar{g}_n(\theta) - \min_{\theta \in \Theta} \bar{g}_n(\theta)' W_{\text{MB}} \bar{g}_n(\theta)$$

and the corresponding GMM estimator,

$$\hat{\theta}_{n,W} = \arg\min_{\theta \in \Theta} \bar{g}_n(\theta)' W_{\text{MB}} \bar{g}_n(\theta).$$

We are now ready to state the additional assumptions required for this section.

Assumption 5. *The following holds for some $c > 0$, $\delta > 0$, $M < \infty$, and $p \geq 1$.*

(A5.1) $\frac{\partial^2 \bar{Q}_{n,W}(\theta)}{\partial \theta \partial \theta'} \geq c I_d$ *for all* $\|\theta - \hat{\theta}_{n,W}\| \leq \delta$ *w.p.a.*1.

(A5.2) $\bar{Q}_{n,W}(\theta) \geq c$ *and* $\|\frac{\partial}{\partial\theta}\bar{Q}_{n,W}(\theta)\|^2 \geq c$ *for all* $\|\theta - \hat{\theta}_{n,W}\| \geq \delta$ *w.p.a.*1.

(A5.3) $\frac{\partial^2 \bar{Q}_{n,W}(\theta)}{\partial\theta\partial\theta'} \leq MI_d$ *and* $\frac{1}{n}\sum_{i=1}^{n}\|g(z_i,\theta)\|^{2p} \leq M(\bar{Q}_{n,W}(\theta)^p + 1)$ *for all* $\theta \in \Theta$ *w.p.a.*1.

(A5.4) *As* $t \to \infty$, $B_{G,t} \to B_G \equiv B_{G,\infty}$, *where either* $B_G \in \mathbb{N}$ *or* $B_G = \infty$.

(A5.5) *The iteration numbers* (N, T) *satisfy* $N = O(T - N)$, *with* $(T-N)^{1-a}/n \to 0$, $N^{1-a}/n \to 0$, *and* $N^a/n = O(1)$.

Conditions (A5.1)–(A5.3) mirror Conditions (A4.6)–(A4.8), but are formulated for the weighted objective $\bar{Q}_{n,W}(\theta)$ in place of $\bar{Q}_n(\theta)$. Condition (A5.4) allows for a sequence $B_{G,t}$ that either converges to a finite constant or diverges. Condition (A5.5) governs the joint growth of N and T relative to the sample size n. To interpret this condition, suppose $a = 0.501$, as used in our Monte Carlo experiments. The requirement $(T-N)^{1-a}/n \to 0$ implies that $T - N$ must grow slower than $n^{1/(1-a)}$, or approximately $n^{2.004}$. Likewise, the conditions $N^{1-a}/n \to 0$ and $N^a/n = O(1)$ are satisfied when $N = O(n^{1/a})$, or approximately $O(n^{1.996})$. Therefore, both conditions are satisfied as long as long as N and $T - N$ grow at a polynomial rate in n, with exponents slightly below 2, ensuring the validity of the refinement procedure in large samples.

Denote the additional variance component by

$$V_G = (G_o'\Omega_o^{-1}G_o)^{-1}\mathbb{E}\left[(G(z_i,\theta_o) - G(\theta_o))'\,\Omega_o^{-1}\,(G(z_i,\theta_o) - G(\theta_o))\right](G_o'\Omega_o^{-1}G_o)^{-1},$$

which reflects the variance of the mini-batch estimator $\tilde{G}_t(\theta^*_{t-1})$ as $\theta^*_{t-1} \to \theta_o$.[4]

Theorem 4 (Second-Order Algorithm: CLT and FCLT)**.** *(i) Let Assumptions 1, 2, 3, 4 and 5 hold with* $p > 1$. *Then, as* n, N, M_{MB}, *and* $T - N$ *all tend to infinity, it holds*

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_{n,W} - \theta_o) \\ \sqrt{T-N}(\bar{\theta}^*_T - \hat{\theta}_{n,W}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} (G_o'\Omega_o^{-1}G_o)^{-1} & 0 \\ 0 & \frac{1}{B_g}\left((G_o'\Omega_o^{-1}G_o)^{-1} + \frac{1}{B_G}V_G\right) \end{pmatrix}\right).$$

(ii) Let Assumptions 1, 2, 3, 4 and 5 with $p > (1-a)^{-1}$. *Then, as* n, N, M_{MB}, *and*

[4]Note that V_G/B_G differs from (8) because we now employ the optimal weighting matrix.

$T - N$ *all tend to infinity, it holds*

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_{n,W} - \theta_o) \\ \frac{1}{\sqrt{T-N}} \sum_{t=N+1}^{N+\lfloor (T-N)r \rfloor} (\theta_t^* - \hat{\theta}_{n,W}) \end{pmatrix} \rightsquigarrow \begin{pmatrix} (G_o' \Omega_o^{-1} G_o)^{-1/2} Z \\ \frac{1}{B_g^{1/2}} \left((G_o' \Omega_o^{-1} G_o)^{-1} + \frac{1}{B_G} V_G \right)^{1/2} W(r) \end{pmatrix}.$$

Theorem 4 establishes the CLT and FCLT for the efficient second-order algorithm, extending the results in Theorems 2 and 3. The limiting distributions capture the efficiency gains achieved through the second-order update rule in (17). In particular, the variance of the asymptotic distribution for $\sqrt{(T-N)B_g}(\bar{\theta}_T^* - \hat{\theta}_{n,W})$ closely approximates that of the efficient GMM estimator, up to an additional variance term V_G/B_G, which arises from the stochastic approximation $\tilde{G}_t(\theta)$ to $\bar{G}_n(\theta)$. This additional variance vanishes in the limit as $B_{G,t} \to B_G = \infty$, in which case $\bar{\theta}_T^*$ attains the same asymptotic efficiency as $\hat{\theta}_{n,W}$, up to a known scalar factor.

Random Scaling Inference Just as Theorem 3 provides a foundation for random scaling inference, Theorem 4(ii) implies that the algorithm described in Section 4 remains applicable with minor modifications. Specifically, we begin updating the random scaling matrix V_t at the start of the refinement stage and modify the test statistic to

$$(n^{-1} + ((T-N)B_g)^{-1})^{-1} (R\bar{\theta}_T^* - c)' (B_g V_T(R))^{-1} (R\bar{\theta}_T^* - c),$$

while continuing to use the same critical values as before.

When $R\theta_o$ is scalar (i.e., $\ell = 1$), the $(1-\alpha)$-level random scaling confidence interval for $R\theta_o$ becomes

$$\left[R\bar{\theta}_T^* \pm \mathrm{cv}_{1-\alpha/2} \sqrt{n^{-1} + ((T-N)B_g)^{-1}} \, (B_g V_T(R))^{1/2} \right], \tag{20}$$

where $\mathrm{cv}_{1-\alpha/2}$ is the critical value defined in Section 4.

Plug-In Inference We now turn to an important special case in which $B_{G,t} \to B_G = \infty$. In this setting, the refinement procedure achieves near-efficiency and allows for plug-in inference based on consistently estimated quantities. This leads to the following corollary.

Corollary 2 (Second-Order Algorithm: Plug-in Inference)**.** *Assume the conditions in*

Theorem 4(i) hold with $B_G = \infty$. Then, under $H_0 : R\theta_o = c$,

$$\frac{(R\bar{\theta}_T^* - c)' \left(R(G_o'\Omega_o^{-1}G_o)^{-1}R'\right)^{-1} (R\bar{\theta}_T^* - c)}{n^{-1} + ((T-N)B_g)^{-1}} \xrightarrow{d} \chi_\ell^2. \tag{21}$$

The corollary provides a Wald-type test statistic based on $\bar{\theta}_T^*$, using a plug-in estimate of the asymptotic variance implied by Theorem 4(i). Since plug-in inference is most appropriate when the sample size n is relatively modest, we estimate G_o and Ω_o^{-1} using full-sample quantities evaluated at $\bar{\theta}_T^*$. Specifically, define

$$\Phi_n(\bar{\theta}_T^*) := \frac{1}{n}\sum_{i=1}^n G\big(z_i, \bar{\theta}_T^*\big), \quad W_n(\bar{\theta}_T^*) := \left(\frac{1}{n}\sum_{i=1}^n g(z_i, \bar{\theta}_T^*)g(z_i, \bar{\theta}_T^*)'\right)^\dagger.$$

Then, a feasible version of the Wald statistic in (21) is

$$\frac{(R\bar{\theta}_T^* - c)' \left(R\left(\Phi_n(\bar{\theta}_T^*)'W_n(\bar{\theta}_T^*)\Phi_n(\bar{\theta}_T^*)\right)^{-1}R'\right)^{-1} (R\bar{\theta}_T^* - c)}{n^{-1} + ((T-N)B_g)^{-1}}, \tag{22}$$

and the null $H_0 : R\theta_o = c$ is rejected if this statistic exceeds the χ_ℓ^2 critical value.

As a counterpart to (20), a $(1-\alpha)$ plug-in confidence interval for the scalar case of $R\theta_o$ is given by

$$\left[R\bar{\theta}_T^* \pm z_{1-\alpha/2}\sqrt{n^{-1} + ((T-N)B_g)^{-1}}\left(R\left(\Phi_n(\bar{\theta}_T^*)W_n(\bar{\theta}_T^*)\Phi_n(\bar{\theta}_T^*)'\right)^{-1}R'\right)^{1/2}\right], \tag{23}$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution.

5.1 Comparison of Estimation and Inference Methods

Both random scaling and plug-in approaches provide asymptotically valid inference for the refined GMM estimator $\bar{\theta}_T^*$, but each offers distinct practical advantages. Random scaling avoids computing full-sample quantities such as the sample moment function and Jacobian, making it especially useful when the dataset is too large to fit in memory. By contrast, the plug-in method relies on full-sample estimates of the Jacobian and the optimal GMM weighting matrix, which can deliver more accurate and tighter confidence intervals when n is moderate and full-sample access is feasible. The choice between the two methods should depend on computational constraints and

the precision required.

To sharpen the discussion, it is useful to distinguish three scenarios. In the first, a *strictly online* setting possibly involving real-time decision-making, each data batch is observed only once and cannot be revisited. This also covers cases in which privacy concerns prevent storage of the full dataset. In such situations, only a single-pass stochastic algorithm is feasible since neither a consistent initial estimator nor multi-pass processing is available. For inference, random scaling remains viable, as it permits continuous updating without revisiting past data, while plug-in methods are infeasible because they require recomputation of full-sample statistics after estimation.

The second scenario involves *sequential streaming* of very large datasets. Here, the data cannot be loaded into memory all at once and must be processed in parts. Estimation again requires a stochastic algorithm due to memory constraints, but multi-pass processing is now possible. For inference, both random scaling and plug-in methods can be applied, although the plug-in approach entails more demanding memory usage.

Table 1: Estimation and inference methods across scenarios

Scenario	Estimation	Inference
Online (real-time, no storage)	Single-pass stochastic only	Random scaling only
Streaming (too large for memory)	Stochastic; multi-pass possible	Both feasible; plug-in memory-heavy
In-memory but intensive	Stochastic preferred	Both feasible; differences minor

In the third scenario, the dataset fits into memory but batch algorithms that process the entire dataset in each iteration are *computationally burdensome*. In this case, stochastic algorithms are preferred for estimation because of their computational efficiency. Both random scaling and plug-in inference methods are feasible, and the computational difference between them may be negligible since the plug-in method involves only sample averages and matrix inversions rather than repeated iterations over the data. Our numerical experiments in Section 6 correspond to this third setting.

In summary, random scaling is indispensable in strictly online environments, advantageous for sequential streaming, and competitive in memory-intensive settings, while the plug-in approach is attractive when full-sample access is feasible and tighter

confidence intervals are desired. Table 1 provides a compact overview.

5.2 Sargan–Hansen Overidentification Tests

We propose Sargan–Hansen specification tests based on the refined GMM estimator $\bar{\theta}_T^*$. The first is the plug-in J-statistic,

$$J := n\,\bar{g}_n(\bar{\theta}_T^*)' W_{\mathrm{MB}} \bar{g}_n(\bar{\theta}_T^*).$$

Because $\bar{\theta}_T^*$ does not solve the first-order condition $\Phi_n(\theta)' W_n \bar{g}_n(\theta) = 0$ of the full-sample GMM criterion $\bar{g}_n(\theta)' W_n \bar{g}_n(\theta)$, we debias the plug-in moment condition before constructing the test. Let $\bar{\Phi} = \Phi_n(\bar{\theta}_T^*)$, $\bar{W} = W_n(\bar{\theta}_T^*)$, and $\bar{g} = \bar{g}_n(\bar{\theta}_T^*)$. Define

$$\bar{g}_n^D = \Big(I - \bar{W}^{1/2}\bar{\Phi}(\bar{\Phi}'\bar{W}\bar{\Phi})^{-1}\bar{\Phi}'\bar{W}^{1/2}\Big)\bar{W}^{1/2}\bar{g},$$

which reduces to $\bar{W}^{1/2}\bar{g}$ if and only if $\bar{\theta}_T^*$ coincides with the full-sample GMM estimate. The associated debiased J-statistic is

$$J_{\mathrm{D}} := n\,\bar{g}_n^{D\prime}\bar{g}_n^D = n\,\bar{g}'\big(\bar{W} - \bar{W}\bar{\Phi}(\bar{\Phi}'\bar{W}\bar{\Phi})^{-1}\bar{\Phi}'\bar{W}\big)\bar{g}.$$

Finally, we consider an online version. Starting at iteration $t = N+1$, update the averaged moment function recursively as

$$\bar{g}_t^* = \frac{t-N-1}{t-N}\,\bar{g}_{t-1}^* + \frac{1}{t-N}\,\tilde{g}_t(\theta_{t-1}^*), \qquad t = N+1, \ldots, T,$$

with $\bar{g}_N^* = 0$. The corresponding online J-statistic is

$$J^* := \Big(n^{-1} + ((T-N)B_g)^{-1}\Big)^{-1}\,\bar{g}_T^{*\prime} W_{\mathrm{MB}} \bar{g}_T^*.$$

The next theorem establishes the asymptotic distributions of J, J_{D}, and J^*.

Theorem 5 (Plug-in, Debiased Plug-in, and Online J-Tests)**.** *Let $d_g > d$, where d_g denotes the dimension of the moment function, and $\chi^2_{d_g-d}$ and χ^2_d be independent chi-squares. Suppose further that $\frac{n}{(T-N)B_g} \to \tau < \infty$ as $n \to \infty$. Then, under the same conditions as in Theorem 4(i),*

$$J \xrightarrow{d} \chi^2_{d_g-d} + \tau\chi^2_d, \quad J_D \xrightarrow{d} \chi^2_{d_g-d}, \quad \text{and} \quad J^* \xrightarrow{d} \chi^2_{d_g-d}.$$

Theorem 5 shows that the plug-in statistic converges to a mixture of two independent chi-squared distributions, with weights determined by the limit of $n/\{(T-N)B_g\}$. The debiased version eliminates this extra variability and converges to the standard $\chi^2_{d_g-d}$. It also shows that the online statistic also converges to $\chi^2_{d_g-d}$. Equivalently, $n\,\bar{g}_T^{*\prime} W_{\mathrm{MB}} \bar{g}_T^* \xrightarrow{d} (1+\tau)\chi^2_{d_g-d}$.

Theorem 5 highlights the key differences between these statistics. Under the null, the classical full-sample Sargan–Hansen test has limiting distribution $\chi^2_{d_g-d}$. By contrast, the plug-in statistic J includes an additional $\tau\chi^2_d$ term, reflecting the variability introduced by using $\bar{\theta}_T^*$ instead of the efficient full-sample GMM estimator $\hat{\theta}_{n,W}$. The debiased version corrects for this source of error, while the online statistic J^* absorbs it through stochastic averaging and converges to a scaled chi-squared distribution.

When τ is small (e.g. $n \ll (T-N)B_g$), all three tests are well approximated by $\chi^2_{d_g-d}$. When τ is non-negligible, the additional variability takes the form $\tau\chi^2_d$ for the plug-in statistic and $\tau\chi^2_{d_g-d}$ for the online statistic, which may lead to differences in size and power. In practice, the online version is preferable when full-sample computation is difficult, while the debiased plug-in test is preferred otherwise.

6 Monte Carlo Experiments

In this section, we report the results of Monte Carlo experiments. First, we discuss the rule-of-thumb methods for selecting the learning rates and mini-batch sizes.

6.1 Learning Rates and Mini-Batch Sizes in Applications

We describe hyperparameter choices for both the warm-start and main stages. In the warm-start stage, limited prior information makes learning rate selection difficult. For Algorithm 2 in the appendix, we set $\gamma_e = \gamma_{0,\mathrm{ws}} e^{-a}$, with $a = 0.501$. We explore several candidate pairs $(\gamma_{0,\mathrm{ws}}, B_{\mathrm{ws}})$, adopting a linear scaling rule where $\gamma_{0,\mathrm{ws}}$ increases linearly with B_{ws}.

For the main stage, we recommend $\gamma_t = \gamma_0 t^{-a}$ with the same exponent. To determine γ_0, we compute $\tilde{G}_t(\bar{\theta}^*_{\mathrm{ws}})$ using mini-batches of size B_{ws} and the warm-start estimate $\bar{\theta}^*_{\mathrm{ws}}$. If n is moderate, we use all $\lfloor n/B_{\mathrm{ws}} \rfloor$ batches; otherwise, we sample a large random subset. Let $\mathbb{T}$ denote the selected batch indices.

Next, in the main algorithm, suppose we set the mini-batch size to $B_{\mathrm{main}} = B_g = B_G$,

which may differ from B_{ws}. To specify γ_0, define

$$\Psi_0 \;:=\; \text{median}\Big\{ \big\| \tilde{G}_t(\bar{\theta}^*_{\text{ws}})' \tilde{G}_t(\bar{\theta}^*_{\text{ws}}) \big\|_2 : t \in \mathbb{T} \Big\},$$

where $\|\cdot\|_2$ denotes the spectral norm. We then propose the following for γ_0:

$$\gamma_0 = \frac{1}{s_0 \Psi_0} \cdot \frac{B_{\text{main}}}{B_{\text{ws}}}, \tag{24}$$

where $s_0 > 0$ is a tuning constant. This choice controls early-stage instability and scales γ_0 appropriately with the batch size. Since γ_0 is one of the most critical hyperparameters, we strongly recommend tuning s_0 based on performance, as shown in our Monte Carlo results.[5]

6.2 EASI GMM

Lewbel and Pendakur (2009) introduced the Exact Affine Stone Index (EASI) demand model. Specifically, the implicit Marshallian budget shares ($\mathbf{w} \in \mathbb{R}^J$) for J goods are given by:

$$\mathbf{w} = \sum_{r=0}^{5} \mathbf{b}_r y^r + \mathbf{C}\boldsymbol{z} + \mathbf{D}\boldsymbol{z}y + \sum_{l=0}^{L} z_l \mathbf{A}_l \mathbf{p} + \mathbf{B}\mathbf{p}y + \boldsymbol{\varepsilon}, \tag{25}$$

where $y \in \mathbb{R}$ is the implicit utility defined as:

$$y = \frac{x - \mathbf{p}'\mathbf{w} + \frac{1}{2}\sum_{l=0}^{L} z_l \mathbf{p}'\mathbf{A}_l\mathbf{p}}{1 - \frac{1}{2}\mathbf{p}'\mathbf{B}\mathbf{p}}, \tag{26}$$

$\boldsymbol{z} \equiv (z_1, \ldots, z_L)' \in \mathbb{R}^L$ is a vector of demographic characteristics with $z_0 = 1$, $\mathbf{p} \in \mathbb{R}^J$ denotes a vector of log prices, $x \in \mathbb{R}$ is the logarithm of nominal total expenditures, and $\boldsymbol{\varepsilon} \in \mathbb{R}^J$ is a vector of unobserved preference shocks. Unknown parameter vectors and matrices in (25) are: $\{\mathbf{b}_r : r = 0, \ldots, 5\}$, $\mathbf{C}$, $\mathbf{D}$, $\{\mathbf{A}_l : l = 0, \ldots, L\}$, and $\mathbf{B}$.

The EASI demand system is nonlinear in its parameters because y depends on $\{\mathbf{A}_l : l = 0, \ldots, L\}$ and $\mathbf{B}$. Furthermore, y can act as an endogenous right-hand-side variable since the budget shares influence y through the Stone index, $\mathbf{p}'\mathbf{w}$. Unknown

[5] In a related context, Bengio (2012, p. 447) emphasized the importance of tuning the initial learning rate, noting: "If there is only time to optimize one hyper-parameter and one uses stochastic gradient descent, then this is the hyper-parameter that is worth tuning."

parameters in the EASI demand functions can be estimated using nonlinear GMM, provided that a sufficient number of valid instruments are available.

Their empirical analysis examines annual expenditures across $J = 9$ categories: food at home, food out, rent, household operation, household furnishing, clothing, transportation operation, recreation, and personal care. The dataset incorporates price information spanning 12 years across 4 Canadian regions, resulting in 48 distinct price vectors. The model includes $L = 5$ demographic characteristics: age minus 40, binary indicators for male, non-ownership of a car, and receipt of social assistance, along with a linear time trend (calendar year minus 1986). Their dataset comprises $n = 4,847$ observations of single-person households renting their homes, aged 25 to 64, with positive expenditures on rent, recreation, and transportation. For further details on the estimation sample, refer to their Section II.B and Table 1.

As is standard practice, Lewbel and Pendakur (2009) excluded the last demand equation (personal care) to satisfy the adding-up constraint for the budget shares and estimated eight share equations. Consequently, only $J - 1 = 8$ log prices, expressed relative to the omitted category, need to be included. This results in a total of $[6 + 2L + (L + 2)(J - 1)](J - 1) = 576$ unknown parameters, which will be further reduced to 380 if Slutsky symmetry is imposed by assuming the symmetry of $\{\mathbf{A}_l : l = 0, \ldots, L\}$, and $\mathbf{B}$.

For computational implementation of nonlinear GMM, we regard the Stata code developed by Pendakur (2015) as the benchmark, which utilizes Stata's GMM command. We use the same instruments as in Pendakur (2015):

$$\boldsymbol{q} = [1, x, \ldots, x^5, \boldsymbol{p}', \boldsymbol{z}', \boldsymbol{z}'x, \boldsymbol{p}'x, \boldsymbol{p}'z_1, \ldots, \boldsymbol{p}'z_L]', \tag{27}$$

where $\boldsymbol{p}$ (slightly abusing notation) represents the $J - 1$ vector of log prices relative to the omitted category. The rationale for these instruments is that x is exogenous but correlated with the endogenous variable y. As noted by Pendakur (2015), the original instruments proposed in Lewbel and Pendakur (2009) are more complex and less transparent, as they require preliminary estimation steps. The model is just-identified when symmetry is not imposed and overidentified under the symmetry restrictions. In the remainder of the paper, we focus on the overidentified case.

6.3 Empirical Results

As in Lewbel and Pendakur (2009), we begin by describing the estimated Engel curves. They focused on the estimated budget shares evaluated at $\mathbf{p} = \mathbf{0}_J$, $\boldsymbol{z} = \mathbf{0}_L$, and $\boldsymbol{\varepsilon} = \mathbf{0}_J$, where $\mathbf{0}_M$ denotes the zero vector of dimension M:

$$x \mapsto \widehat{\mathbf{w}}(x) = \sum_{r=0}^{5} \widehat{\mathbf{b}}_r x^r,$$

where $\{\widehat{\mathbf{b}}_r\}_{r=0}^5$ are the GMM estimators of $\{\mathbf{b}_r\}_{r=0}^5$. As a function of x, it corresponds to the Engel curves for the reference individual facing the log prices of $\mathbf{p} = \mathbf{0}_J$. Here, the reference individual is defined as a 40-year-old, car-owning female in 1986 who did not receive social assistance (i.e., $\boldsymbol{z} = \mathbf{0}_L$) with $\boldsymbol{\varepsilon} = \mathbf{0}_J$.

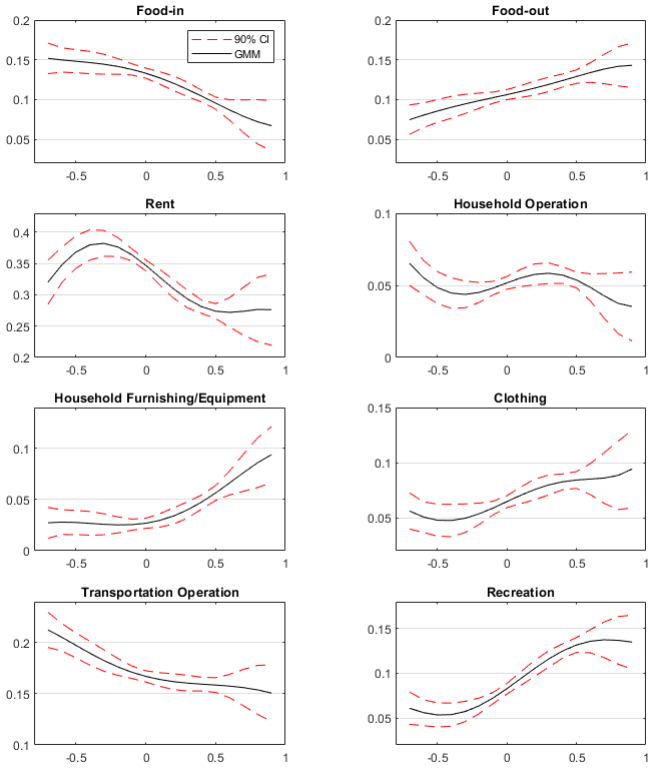

Figure 2: Estimated Engel Curves

Figure 2 shows the estimated Engel curves along with 90% pointwise confidence intervals. The domain of the curves is restricted to [-0.7, 0.9], which is well within the

support of x. The estimated Engel curves presented in Figure 2 differ slightly from those in Lewbel and Pendakur (2009), as the latter are based on nonlinear 3SLS and use a different set of instruments as explained in Section 6.2. For instance, the slope of the house operation Engel curve becomes negative when the nominal expenditure, x, is large, whereas their estimate remains nearly a positive linear segment in that range. We also find that the confidence intervals based on the GMM tend to be wider than those of the 3SLS particularly when x is large.

We now turn to the estimated price effects. Instead of going over all the empirical results reported in Lewbel and Pendakur (2009), we focus on one noteworthy case: Lewbel and Pendakur (2009, p. 850) remarked, for the reference individual who faces $\mathbf{p} = \mathbf{0}_J$ at median expenditure where $x = 0$, that "a rent price increase of 10 percent would be associated with a budget share 0.63 percentage points higher when expenditure is raised to equate utility with that in the initial situation." This remark is based on the estimate of the relevant component in $\boldsymbol{A}_0$. In other words, the own-price compensated semi-elasticity for the rent budget share is 0.063. Using the same instruments as in Pendakur (2015), we find that now the point estimate is 0.053 and its 90% confidence interval is [0.010, 0.096].

In terms of computation, it took 27 minutes to obtain all the empirical results reported above using a laptop computer with an Apple M1 Pro chip (8-core CPU, 14-core GPU), 16GB of RAM, and macOS Ventura 13.5. As it can be computed in reasonable time with $n = 4,847$, the sample size used in Lewbel and Pendakur (2009), we carry out experiments based on this empirical example, while increasing the scale from $n = 20,000$ to $n = 10^6$.

6.4 The Data Generating Process in the Experiments

Because of the simultaneity of $\mathbf{w}$ and y, it is challenging to simulate $\mathbf{w}$ while allowing for non-zero higher-order polynomial terms of y. As the main goal of the experiment is to compare the standard GMM with our stochastic approximation approach in terms of inference precision and computational cost in a well-identified setting, we simplify the data generating process in the experiments as follows.

We first set the true values of parameters at the GMM estimates reported in the previous subsection. Obtain the sample variance of the regression residuals, denoted by $\widehat{\sigma}_j^2$, where $j = 1, \ldots, J-1$. Compute the sample average of observed budget shares,

denoted by $\overline{\boldsymbol{w}}_{\text{sample}}$. Normalize the scale of the demographic characteristics, $\boldsymbol{z}_i$, by dividing each component by its maximum absolute value.

Simulating observations, we first randomly draw $\{(x_i, \mathbf{p}_i, \boldsymbol{z}_i), i = 1, \ldots, n\}$ with replacement from the original dataset, with $\boldsymbol{z}_i$ normalized as described above, where n can be larger than the original sample size. Then, generate y_i by replacing $\mathbf{w}$ in (26) with $\overline{\boldsymbol{w}}_{\text{sample}}$:

$$y_i = \frac{x_i - \mathbf{p}_i' \overline{\boldsymbol{w}}_{\text{sample}} + \frac{1}{2} \sum_{l=0}^{L} z_{l,i} \mathbf{p}_i' \widehat{\mathbf{A}}_l \mathbf{p}_i}{1 - \frac{1}{2} \mathbf{p}_i' \widehat{\mathbf{B}} \mathbf{p}_i}, \tag{28}$$

where $\widehat{\mathbf{A}}_l$ and $\widehat{\mathbf{B}}$ are the GMM estimates obtained in the preliminary step. Finally, for each $i = 1, \ldots, n$, generate $\mathbf{w}_i$ by

$$\mathbf{w}_i = \sum_{r=0}^{5} \widehat{\mathbf{b}}_r y_i^r + \widehat{\mathbf{C}} \boldsymbol{z}_i + \widehat{\boldsymbol{D}} \boldsymbol{z}_i y_i + \sum_{l=0}^{L} z_{l,i} \widehat{\boldsymbol{A}}_l \mathbf{p}_i + \widehat{\boldsymbol{B}} \mathbf{p}_i y_i + \boldsymbol{\varepsilon}_i, \tag{29}$$

where the jth component of $\boldsymbol{\varepsilon}_i$ is randomly drawn from $N\left(0, \hat{\sigma}_j^2\right)$ and the components of $\boldsymbol{\varepsilon}_i$ are mutually independent. By design, y is exogenous; however, GMM using the instruments given in (27) is still valid. One caveat is that the simulation design does not guarantee that each element of $\mathbf{w}_i$ is bounded between 0 and 1; however, it is of little concern because the moment conditions we consider do not impose the boundedness requirement either.

6.5 Simulation Results: Precision and Computational Cost

We employ two different types of sample sizes: We first consider $n \in \{$20,000, 50,000, 100,000$\}$, which reflect scales that can be encountered in contemporary demand estimation scenarios. We then use $n =$ 1,000,000, an extremely large sample size, for which full-sample GMM is practically infeasible due to computational limitations. The number of simulation replications is set to 500. Our simulations are conducted on the High-Performance Computing (HPC) cluster at the University of Connecticut. The cluster features a range of CPU architectures, including Intel Skylake, AMD EPYC 64-core, and AMD EPYC 128-core processors, with compute nodes offering between 34 and 126 cores and between 187 GB and 503 GB of RAM. Computation time is measured separately using 5 replications on a laptop computer equipped with an Apple

M1 Pro chip (8-core CPU, 14-core GPU), 16GB of RAM, and macOS Ventura 13.5.

We begin by conducting simulations with $n \in \{20{,}000,\ 50{,}000,\ 100{,}000\}$. In this setting, we estimate the model using our second-order stochastic approximation approach. For the comparative purpose, we also consider the Stata GMM by Pendakur (2015). Since simulations for our method are conducted in MATLAB, to minimize simulation discrepancies between two platforms, simulated datasets are first generated and then same datasets are used in both.

Let θ denote the vectorization of model parameters $\{\mathbf{b}_r : r = 0, \ldots, 5\}$, $\mathbf{C}$, $\mathbf{D}$, $\{\mathbf{A}_l : l = 0, \ldots, L\}$, and $\mathbf{B}$ under the Slutsky symmetry restrictions. Our procedure follows the steps outlined below. First, we apply the warm-start algorithm described in Algorithm 2 with $\theta^*_{0,\text{ws}} = \mathbf{0}_{380}$. We set $\gamma_{0,\text{ws}} = 0.1$ and $\alpha = 0.501$. Second, we use our stochastic approximation, based on Algorithm 1, to obtain θ^*_N and $\bar{\theta}^*_N$. In both steps, we employ the weighting matrix,

$$W_{2SLS} = I_{J-1} \otimes \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{q}_i \boldsymbol{q}_i' \right)^{-1},$$

assuming the independence of unobserved preference shocks across equations. The initial step size γ_0 is selected based on the rule-of-thumb method proposed in Section 6.1. We use $s_0 = 5$ and 8 for $n = 20,000$, which yield average values of $\gamma_0 = 0.082$ and 0.051, respectively. For $n = 50,000$ and $100,000$, we use $s_0 = 3$ and 8, resulting in average values of $\gamma_0 = 0.136$ and 0.051, respectively. Finally, we apply the refinement step for efficient estimation, as proposed in Section 5. We use the mini-batch based optimal weight in (16), with $M_{\text{MB}} = 50,000$, under the same independence restriction.

The batch sizes are set to $B_G = B_g = 512$ in both the warm-start and mini-batch stochastic approximation steps. In the refinement step, we set $B_g = 512$ and $B_{G,t} = B_{G,0} + \log(t - N)$ with $B_{G,0} = 512$, so that $B_{G,t}$ diverges as T increases, ensuring asymptotic efficiency. We begin updating the efficient average estimator at the start of the refinement step: $\bar{\theta}^*_t = \frac{t-N-1}{t-N}\bar{\theta}^*_{t-1} + \frac{1}{t-N}\theta^*_t$ for $t = N+1, ..., T$. For inference based on stochastic approximation, we consider the two methods described in Section 5: random scaling (RS) inference and plug-in (PI) inference (see (20) and (23) for the corresponding confidence interval formulas).

We first focus on estimating the Engel curves. We use the root integrated mean

squared error (RIMSE) as a precision measure, that is defined as

$$\sqrt{\sum_{j=1}^{J-1}\int_{\underline{x}}^{\overline{x}}\mathbb{E}\left(\sum_{r=0}^{5}\left(\widehat{b}_{j,r}-b_{j,r}\right)x^{r}\right)^{2}dx},$$

where $b_{j,r}$ is the jth element of $\mathbf{b}_r$, $[\underline{x},\overline{x}]=[-0.7,0.9]$ and the integral is approximated by partitioning the domain into an interval width of 0.1.

Table 2: Simulation Results for Engel Curves

	Full-sample GMM	SLIM			
	(1)	(2)	(3)	(4)	(5)
Sample Size			(N,T)		
(n)		(20,000,	40,000)	(20,000,	60,000)
			(s_0, E_{ws})		
20,000		$(5,25)$	$(8,25)$	$(5,25)$	$(8,25)$
50,000		$(3,\ 4)$	$(8,\ 4)$	$(3,\ 4)$	$(8,\ 4)$
100,000		$(3,\ 1)$	$(8,\ 1)$	$(3,\ 1)$	$(8,\ 1)$
Panel A. Precision (Root Integrated Mean Squared Error)					
20,000	0.024	0.025	0.025	0.024	0.024
50,000	0.015	0.016	0.017	0.016	0.016
100,000	NA	0.012	0.013	0.011	0.012
Panel B. Computational Cost (Hours)					
20,000	1.93	1.09	1.06	1.31	1.30
50,000	7.84	1.05	1.06	1.33	1.33
100,000	18.93*	1.06	1.07	1.35	1.35

$*$ This exceeds our computational budget.

Table 2 reports the results. In column (1), we present simulation results from Stata (full-sample) GMM. Columns (2)-(5) report results from our SLIM method, using varying numbers of iterations and step sizes. The computational cost represents the average computation time in hours. We fix the number of iterations at (N,T)=(20,000, 40,000) and (20,000, 60,000) for all sample sizes. The number of epochs in the warm-

start step is set to $E_{\text{ws}} = 25$, 4, and 1 for $n = 20{,}000$, 50,000, and 100,000, respectively, which yield approximately 37,000 to 38,000 iterations ($= \lfloor n/B_{\text{ws}} \rfloor \times (\lfloor n/B_{\text{ws}} \rfloor) - 1) \times E_{\text{ws}}$). Under this design, the computation time for our method remains within 1 to 1.5 hours, regardless of the sample size.

The table shows that our SLIM demonstrates a substantial gain in computational efficiency compared to the full-sample GMM, while maintaining comparable estimation precision. Our approach tends to yield accurate estimates within 1.05 to 1.35 hours across all sample sizes considered. In contrast, the computation time for the full-sample GMM increases substantially with sample size. For instance, when $n = 50{,}000$, the full-sample GMM takes 7.84 hours, while our method yields slightly larger RIMSEs in only 1.05 to 1.33 hours. The computational advantage becomes even more pronounced when $n = 100{,}000$. Our method produces accurate estimates within 1.06 to 1.35 hours, whereas the full-sample GMM exceeds 18 hours, surpassing our computational budget. These results highlight that the computational gain of our procedure becomes increasingly significant when working with large datasets.

We next examine the finite sample performance based on the price effect on the rent budget share for the reference individual with $\mathbf{p} = \mathbf{0}_J$ and $x = 0$, which corresponds to the third diagonal element of $\boldsymbol{A}_0$, namely $\boldsymbol{A}_{0,33} = 0.0533374$.

Table 3 summarizes the simulation results. It first reports the root mean square error (RMSE), bias and standard deviation (SD) to evaluate estimation accuracy. We observe similar patterns to those in the Engel curves estimation in Table 2. Our SLIM produces accurate estimates, comparable to the full-sample GMM estimates in terms of RMSE, but with significantly shorter computation time. As in the Engel curve estimation, the full-sample GMM exceeds the time limit when $n = 100{,}000$, while our method delivers accurate estimates in slightly over one hour.

In this simulation, we also assess the accuracy of statistical inference. The table reports the empirical coverage probabilities at the 0.95 nominal level and the average lengths of the associated confidence intervals. For our approach, we consider the RS inference as well as the PI inference. Overall, both procedures tend to yield accurate empirical coverage rates, which are comparable to the full-sample GMM inference, particularly when $s_0 = 3$ and 5. Finally, the table shows that the average confidence interval lengths of our PI method are comparable to those of the full-sample GMM.

We then examine how the SLIM estimator and its associated random scaling confidence intervals behave as the number of iterations T increases. Figure 3 shows

Table 3: Simulation Results for Price Effect on Rent Budget Share

	Full-Sample GMM	SLIM							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Sample Size		(N,T)							
(n)		(20,000, 40,000)				(20,000, 60,000)			
		(s_0, E_{ws})							
20,000		$(5, 25)$		$(8, 25)$		$(5, 25)$		$(8, 25)$	
50,000		$(3, 4)$		$(8, 4)$		$(3, 4)$		$(8, 4)$	
100,000		$(3, 1)$		$(8, 1)$		$(3, 1)$		$(8, 1)$	
		Panel A. Bias							
20,000	0.001	-0.006		-0.006		-0.006		-0.006	
50,000	0.000	-0.001		-0.002		-0.001		-0.002	
100,000	NA	0.000		-0.001		0.000		-0.001	
		Panel B. SD							
20,000	0.019	0.016		0.016		0.016		0.016	
50,000	0.013	0.012		0.012		0.012		0.012	
100,000	NA	0.009		0.009		0.009		0.009	
		Panel C. RMSE							
20,000	0.019	0.017		0.017		0.017		0.017	
50,000	0.013	0.012		0.012		0.012		0.012	
100,000	NA	0.009		0.009		0.009		0.009	
		Panel D. Coverage Probability (Nominal Level: 0.95)							
		RS	PI	RS	PI	RS	PI	RS	PI
20,000	0.962	0.946	0.988	0.934	0.986	0.970	0.990	0.960	0.990
50,000	0.948	0.946	0.956	0.910	0.962	0.972	0.958	0.954	0.962
100,000	NA	0.958	0.960	0.934	0.962	0.964	0.962	0.948	0.960
		Panel E. CI Length (Nominal Level: 0.95)							
		RS	PI	RS	PI	RS	PI	RS	PI
20,000	0.077	0.138	0.078	0.142	0.078	0.138	0.078	0.144	0.078
50,000	0.049	0.076	0.049	0.072	0.049	0.076	0.049	0.075	0.049
100,000	NA	0.053	0.035	0.051	0.035	0.056	0.035	0.056	0.035
		Panel F. Computational Cost (Hours)							
		RS	PI	RS	PI	RS	PI	RS	PI
20,000	1.93	1.05	1.04	1.03	1.05	1.33	1.29	1.36	1.30
50,000	7.84	1.06	1.06	1.07	1.06	1.34	1.36	1.34	1.34
100,000	18.93*	1.07	1.09	1.07	1.08	1.38	1.38	1.36	1.37

$*$ This exceeds our computational budget.

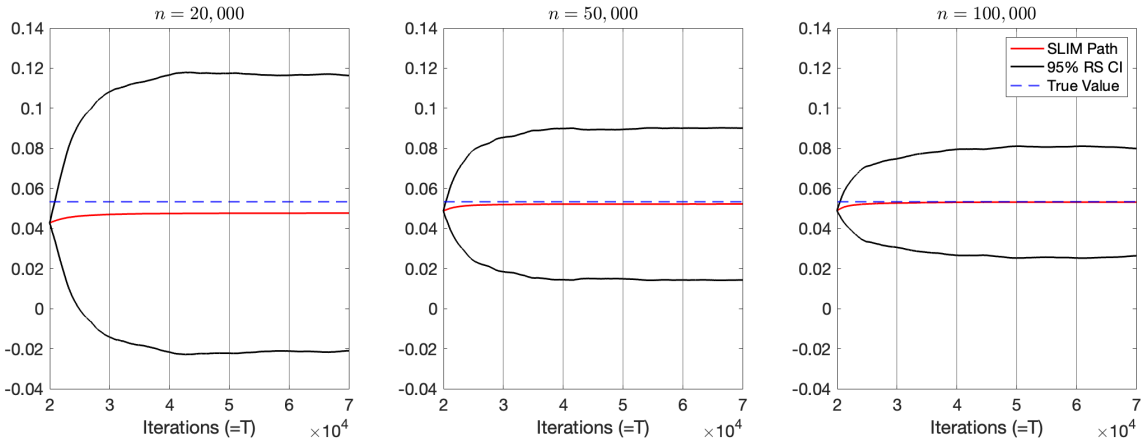

Figure 3: SLIM Path and Random Scaling Confidence Intervals over Iterations for the Price Effect on Rent Budget Share

the average SLIM path and the corresponding 95% random scaling confidence intervals over T, computed by averaging across simulation replications. We set $(s_0, E_{\text{ws}}) = (5, 25), (3, 4)$, and $(3, 1)$ for n = 20,000, 50,000, and 100,000, respectively. The confidence intervals begin to stabilize around T = 40,000 for all sample sizes.

We next evaluate the performance of the SLIM based Sargan–Hansen overidentification tests developed in Section 5.2. For this experiment, we generate instruments based on

$$x - c\,\mathbf{p}'\mathbf{w}$$

in place of x in (27), where $c \in \{0, 1, 2\}$ controls the strength of endogeneity. When $c = 0$, the instruments are exogenous, while larger values of c correspond to stronger endogeneity. We consider the proposed debiased overidentification test, as well as the plug-in and online overidentification tests. For comparison, we also employ the standard overidentification test based on the full-sample GMM.

Table 4 reports the empirical rejection probabilities at the 0.05 nominal level. We observe that our debiased overidentification test achieves accurate empirical size, while the plug-in and online overidentification tests rarely reject the null hypothesis when $c = 0$. The power of the debiased test increases with both the degree of endogeneity and the sample size. The plug-in and online tests fail to detect the endogeneity of instruments for $c = 1, 2$ when n = 20,000, 50,000, while they show good power properties when n = 100,000 and $c = 2$. The standard overidentification test performs

comparable to our debiased test, but requires much longer computation time as the sample size grows. As in the previous simulation experiments, the computation time exceeds our budget when n =100,000, preventing us from simulating the empirical rejection probabilities.

Table 4: Empirical Rejection Probabilities of Sargan-Hansen Overidentification Tests

Sample Size	Full-Sample		SLIM	
(n)	GMM	Debiased	Plug-in	Online
		$c = 0$		
20,000	0.044	0.044	0.018	0.018
50,000	0.048	0.044	0.000	0.000
100,000	NA*	0.052	0.002	0.000
		$c = 1$		
20,000	0.090	0.088	0.000	0.000
50,000	0.198	0.198	0.000	0.000
100,000	NA*	0.483	0.023	0.025
		$c = 2$		
20,000	0.304	0.303	0.000	0.000
50,000	0.906	0.905	0.000	0.000
100,000	NA*	1.000	0.825	0.833

$(s_0, E_{\text{ws}}) = (5, 25), (3, 4)$, and $(3, 1)$ for n = 20,000, 50,000, and 100,000, respectively, and $(N, T) = (20,000,\ 40,000)$

$*$ This exceeds our computational budget.

Lastly, we examine the performance of our approach using an extremely large sample size of n = 1,000,000. In the setting, the full-sample GMM is infeasible due to computational limitations. We evaluate the performance of various versions of our stochastic approximation approach. We first employ the efficient second-order SLIM, as in the previous simulation. We also consider two versions of first-order SLIM methods: one based on Algorithm 1 and one with the refined weight (that is, omitting $(\Phi_n' W_{\text{MB}} \Phi_n)^\dagger$ in (17)). Given that the sample size is extremely large, we use a random sub-sample of observations to implement the warm-start algorithm and to construct the initial weight for all methods, as well as Φ_n for the second-order

stochastic approximation. For the first-order stochastic approximation without the refinement step, we do not update the weighting matrix with the optimal weight; thus, it corresponds to one-step (possibly inefficient) GMM estimation. For inference, we use the RS method, which is computationally efficient.

Table 5: Simulation Results when n = 1,000,000

	First-order SLIM (2SLS Weight)		First-order SLIM (Refined Weight)		Second-order SLIM	
	(1)	(2)	(3)	(4)	(5)	(6)
N	300,000	500,000	20,000	20,000	20,000	20,000
T			120,000	140,000	40,000	60,000
n_{ws}	150,000	150,000	100,000	100,000	100,000	100,000
n_{w}	150,000	150,000	100,000	100,000	100,000	100,000
$(\gamma_{0,\text{ws}}, s_0)$	(0.4, 1)	(0.4, 1)	(0.1, 3)	(0.1, 3)	(0.1, 3)	(0.1, 3)
Panel A. Engel Curves						
RIMSE	0.020	0.020	0.031	0.031	0.006	0.005
Panel B. Price Effect on Rent Budget Share						
Bias	-0.001	-0.001	-0.001	-0.002	0.000	0.001
SD	0.005	0.005	0.004	0.004	0.003	0.003
RMSE	0.005	0.005	0.004	0.004	0.003	0.003
Coverage Probability	0.866	0.924	0.936	0.944	0.934	0.944
CI Length	0.024	0.035	0.026	0.028	0.018	0.018
Panel C. Computational Cost (Hours)						
	4.98	7.74	2.09	2.38	1.11	1.38

$E_{\text{ws}} = 1$ all methods. n_{ws} is the number of observations used in the warm-start stage. n_{w} denotes the number of observations used for the initial weight and Φ_n. "Coverage Probability" and "CI Length" in Panel B are based on 95% confidence.

The results are reported in Table 5. We first examine the performance in estimating the Engel curves. As shown in the table, the first-order SLIM approach without refinement produces accurate estimates with $N = 300,000$ iterations, requiring slightly less than 5 hours of computation time. Moreover, the second-order approach provides additional refinement, yielding more accurate estimates within 1.11-1.38 hours. For the

price effect on the rent budget share, the first-order SLIM without refinement exhibits over-rejection with $N = 300{,}000$, while more reliable inference is achieved with $N = 500{,}000$. The first-order approximation with refinement yields more accurate inference within 2.09-2.38 hours. Further improvement, along with reduced computation time, is achieved using the second-order SLIM approach.

7 Conclusions

This paper has proposed SLIM, a stochastic approximation method for nonlinear generalized method of moments that does not require a consistent initial estimator. The approach exploits U-statistics for both the moment vector and its derivative matrices and incorporates multi-pass mini-batch updates to improve stability and accuracy. Monte Carlo experiments show that SLIM can solve large-scale nonlinear GMM problems with hundreds of parameters and samples of size $n = 100{,}000$ or more in under two hours, whereas full-sample GMM requires an order of magnitude more computation time. These results demonstrate that SLIM provides a practical and scalable tool for econometric applications that are otherwise computationally infeasible with standard methods.

Several extensions merit further study. One promising direction is to extend SLIM to GMM with nonsmooth moment functions, following Chen and Liao (2015), which would broaden its applicability to models with quantile restrictions and other non-differentiable structures. Another is to examine the case of model misspecification, for example by embedding stochastic approximation into the framework of Hansen and Lee (2021), in order to assess the robustness of SLIM. It would also be valuable to develop principled strategies for tuning hyperparameters such as mini-batch sizes, stopping rules, and learning rates, as well as to explore applications in high-dimensional settings where the number of parameters grows with the sample size. We leave these topics to future research.

Appendix: Proof of Consistency

As emphasized in the introduction and earlier sections, SLIM delivers a consistent estimator without the need to initialize the algorithm with a consistent preliminary estimator. Establishing the consistency of SLIM is therefore of fundamental importance.

The proof of Theorem 1 is provided below. The auxiliary lemmas required for this proof, together with the proofs of the remaining theorems, are presented in Appendix D.

Proof of Theorem 1. We write $\theta_N^* - \theta_o = (\theta_N^* - \hat{\theta}_n) + (\hat{\theta}_n - \theta_o)$. The latter term satisfies $\hat{\theta}_n - \theta_o \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$ by Lemma 1. Thus, it suffices to verify that $\theta_N^* - \hat{\theta}_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ and $N \to \infty$. The same applies to $\bar{\theta}_N^*$. By Lemma 3, we can find an increasing deterministic sequence $(\tilde{T}_k)_{k \geq 1}$ such that, for all $k \in \mathbb{N}$,

$$\sup_{n \in \mathbb{N}} \mathbb{P}_n^*(\tau^*_{k^{-2}\underline{c}} > \tilde{T}_k) \mathbb{1}_{E_n} < \frac{1}{k},$$

where $\underline{c} > 0$ is given in Lemma 2 and $\tau^*_{k^{-2}\underline{c}} = \inf\{t \geq 0 : \sup_{s \geq t} \bar{Q}_n(\theta_t^*) \leq k^{-2}\underline{c}\}$. Define the event $\mathcal{E}_{k,n} := \{\tau^*_{k^{-2}\underline{c}} \leq \tilde{T}_k\} \cap E_n$, where E_n is the event on which all statements in Lemma 2 hold and $\mathbb{P}(E_n) \to 1$. By construction of $\mathcal{E}_{k,n}$, it holds that

$$\forall t \geq \tilde{T}_k : \quad \underline{c}(\|\theta_t^* - \hat{\theta}_n\|^2 \wedge \delta^2) \leq \bar{Q}_n(\theta_t^*) \leq k^{-2}\underline{c} \quad \Longrightarrow \quad \|\theta_t^* - \hat{\theta}_n\| \leq \frac{1}{k}$$

on the event $\mathcal{E}_{k,n}$ for $k > 1/\delta$, where $\delta > 0$ is given in Assumption 4. Since $\mathbb{P}(\mathcal{E}_{k,n}^c) = \mathbb{E}[\mathbb{P}_n^*(\tau^*_{k^{-2}\underline{c}} > \tilde{T}_k)\mathbb{1}_{E_n}] + \mathbb{P}(E_n^c) \leq \frac{1}{k} + \mathbb{P}(E_n^c)$, we find that $\limsup_{n \to \infty} \mathbb{P}(\mathcal{E}_{k,n}^c) \leq 1/k$ for all $k > 1/\delta$.

We first show $\|\theta_N^* - \hat{\theta}_n\| \xrightarrow{\mathbb{P}} 0$. Observe that, for all $\varepsilon > 0$ and $k \in \mathbb{N}$ such that $1/k < (\varepsilon \wedge \delta)$, it holds

$$\mathbb{P}\left(\|\theta_N^* - \hat{\theta}_n\| > \varepsilon\right) \leq \mathbb{P}\left(\|\theta_N^* - \hat{\theta}_n\| > 1/k, \mathcal{E}_{k,n}\right) + \mathbb{P}(\mathcal{E}_{k,n}^c) \leq \mathbb{1}\{N < \tilde{T}_k\} + \mathbb{P}(\mathcal{E}_{k,n}^c).$$

As $n \to \infty$ and $N \to \infty$, this implies that $\limsup_{n,N \to \infty} \mathbb{P}(\|\theta_N^* - \hat{\theta}_n\| > \varepsilon) \leq 1/k$. Since $1/k$ can be chosen arbitrarily close to 0 for any given $\varepsilon > 0$, this shows $\lim_{n,N \to \infty} \mathbb{P}(\|\theta_N^* - \hat{\theta}_n\| > \varepsilon) = 0$, establishing $\theta_N^* - \theta_o \xrightarrow{\mathbb{P}} 0$.

Next, we show $\|\bar{\theta}_N^* - \hat{\theta}_n\| \xrightarrow{\mathbb{P}} 0$. Let $\varepsilon > 0$ be given, and pick $k \in \mathbb{N}$ such that $1/k < (\varepsilon \wedge \delta)$. We first write $\bar{\theta}_N^* - \hat{\theta}_n = \frac{1}{N}\sum_{t=1}^{\tilde{T}_k}(\theta_t^* - \hat{\theta}_n) + \frac{1}{N}\sum_{t=\tilde{T}_k+1}^{N}(\theta_t^* - \hat{\theta}_n)$. Since $\|\theta_t^* - \hat{\theta}_n\| < 1/k < \varepsilon$ for all $t \geq \tilde{T}_k$ on $\mathcal{E}_{k,n}$, we have, by the triangle inequality,

$$\begin{aligned}
\mathbb{P}\left(\|\bar{\theta}_N^* - \hat{\theta}_n\| \geq 2\varepsilon\right) &\leq \mathbb{P}\left(\{\|\bar{\theta}_N^* - \hat{\theta}_n\| \geq 2\varepsilon\} \cap \mathcal{E}_{k,n}\right) + \mathbb{P}(\mathcal{E}_{k,n}^c) \\
&\leq \mathbb{P}\left(\frac{1}{N}\sum_{t=1}^{\tilde{T}_k} \|\theta_t^* - \hat{\theta}_n\| \geq \varepsilon\right) + \mathbb{P}(\mathcal{E}_{k,n}^c). \qquad (30)
\end{aligned}$$

We now show that $\lim_{n,N\to\infty}\mathbb{P}(N^{-1}\sum_{t=1}^{\tilde{T}_k}\|\theta_t^*-\hat{\theta}_n\|\geq\varepsilon)=0$ for all k. Since $\tilde{T}_k$ is a fixed integer, it suffices to establish that $N^{-1}\|\theta_t^*-\hat{\theta}_n\|\xrightarrow{\mathbb{P}}0$ for all $t\geq 0$. We proceed by induction on $t\geq 0$. Observe that the statement is true for $t=0$, because $\theta_0^*=O_{\mathbb{P}}(1)$ by assumption, and $\hat{\theta}_n\xrightarrow{\text{a.s.}}\theta_o$ by Lemma 1. Therefore, for the induction step, it suffices to show $N^{-1}\|\theta_t^*-\theta_{t-1}^*\|=N^{-1}\gamma_t\|\tilde{G}_t(\theta_{t-1}^*)'\tilde{g}_t(\theta_{t-1}^*)\|\xrightarrow{\mathbb{P}}0$ for all $t\geq 1$.

In the proof of (D.9) in Lemma 2, we find that, on the event E_n,

$$\mathbb{E}_n^*[\|\tilde{G}_t(\theta)\|^2]\leq C,\quad \mathbb{E}_n^*[\|\tilde{g}_t(\theta)\|^2]\leq C(\bar{Q}_n(\theta)+1),\quad \forall\theta\in\Theta.$$

Let $\eta>0$ and pick $K>0$ such that $\sup_{n\in\mathbb{N}}\mathbb{P}_n^*(T_K^*<\infty)\mathbb{1}_{E_n}<\eta$ by Lemma 3, where $T_K^*=\inf\{t\geq 0:\bar{Q}_n(\theta_t)^*\geq K\}$. Let $\tilde{\mathcal{E}}_{t,n}:=\{T_K^*\geq t\}\cap E_n$. Then, $\mathbb{P}(\tilde{\mathcal{E}}_{t,n}^c)\leq\mathbb{E}[\mathbb{P}_n^*(T_K^*<\infty)\mathbb{1}_{E_n}]+\mathbb{P}(E_n^c)$, and therefore

$$\begin{aligned}
&\mathbb{P}\left(\|\tilde{G}_t(\theta_{t-1}^*)'\tilde{g}_t(\theta_{t-1}^*)\|\geq N\varepsilon\right)\\
&\leq\mathbb{P}\left(\|\tilde{G}_t(\theta_{t-1}^*)'\tilde{g}_t(\theta_{t-1}^*)\|\geq N\varepsilon,\tilde{\mathcal{E}}_{t,n}\right)+\mathbb{P}(\tilde{\mathcal{E}}_{t,n}^c)\\
&\leq\mathbb{E}\left[\mathbb{P}_n^*\left(\|\tilde{G}_t(\theta_{t-1}^*)'\tilde{g}_t(\theta_{t-1}^*)\|\geq N\varepsilon|\mathcal{F}_{n,t-1}^*\right)\mathbb{1}_{\tilde{\mathcal{E}}_{t,n}}\right]+\mathbb{P}(E_n^c)+\mathbb{E}[\mathbb{P}_n^*(T_K^*<\infty)\mathbb{1}_{E_n}]\\
&\leq\frac{C}{N^2\varepsilon^2}\mathbb{E}\left[(Q_{t-1}^*+1)\mathbb{1}\left\{T_K^*\geq t\right\}\right]+\mathbb{P}(E_n^c)+\eta\leq\frac{C(K+1)}{N^2\varepsilon^2}+\mathbb{P}(E_n^c)+\eta.
\end{aligned}$$

Since $\eta>0$ is arbitrary, we obtain $\lim_{n,N\to\infty}\mathbb{P}(N^{-1}\|\tilde{G}_t(\theta_{t-1}^*)'\tilde{g}_t(\theta_{t-1}^*)\|\geq\varepsilon)=0$ for all $t\geq 1$ and $\varepsilon>0$. This establishes that $N^{-1}\|\theta_t^*-\hat{\theta}_n\|\xrightarrow{\mathbb{P}}0$ for all $t\geq 0$, and hence, $\lim_{n,N\to\infty}\mathbb{P}(N^{-1}\sum_{t=1}^{\tilde{T}_k}\|\theta_t^*-\hat{\theta}_n\|\geq\varepsilon)=0$ for all $k\geq 1$. This, in turn, implies that

$$\limsup_{n,N\to\infty}\mathbb{P}\left(\|\bar{\theta}_N^*-\hat{\theta}_n\|\geq 2\varepsilon\right)\leq\frac{1}{k}$$

for all $1/k<(\varepsilon\wedge\delta)$ by (30). Letting $k\to\infty$ completes the proof. □

References

Abadir, K. M. and P. Paruolo (1997). Two mixed normal densities from cointegration analysis. *Econometrica 65*(3), 671–680.

Andrews, D. W. K. (1997). A stopping rule for the computation of generalized method of moments estimators. *Econometrica 65*(4), 913–931.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In G. Montavon, G. B. Orr, and K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade: Second Edition*, pp. 437–478. Berlin, Heidelberg: Springer Berlin Heidelberg.

Benveniste, A., M. Métivier, and P. Priouret (2012). *Adaptive algorithms and stochastic approximations*, Volume 22. Springer Science & Business Media.

Bottou, L., F. E. Curtis, and J. Nocedal (2018). Optimization methods for large-scale machine learning. *SIAM Review 60*(2), 223–311.

Chen, X., Z. Lai, H. Li, and Y. Zhang (2024). Online statistical inference for stochastic optimization via Kiefer-Wolfowitz methods. *Journal of the American Statistical Association 119*(548), 2972–2982.

Chen, X., S. Lee, Y. Liao, M. H. Seo, Y. Shin, and M. Song (2025). SGMM: Stochastic approximation to generalized method of moments. *Journal of Financial Econometrics 23*(1), nbad027.

Chen, X. and Z. Liao (2015). Sieve semiparametric two-step GMM under weak dependence. *Journal of Econometrics 189*(1), 163–186.

Chen, X., A. Roy, Y. Hu, and K. Balasubramanian (2024). Stochastic optimization algorithms for instrumental variable regression with streaming data. In *Advances in Neural Information Processing Systems*.

Chung, K. L. (2000). *A Course in Probability Theory*. Elsevier.

Della Vecchia, R. and D. Basu (2025). Stochastic online instrumental variable regression: Regrets for endogeneity and bandit feedback. *The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*.

Du, X., W. Z. ahd Wei Biao Wu, and S. Na (2025). Online statistical inference of constrained stochastic optimization via random scaling. available at https://doi.org/10.48550/arXiv.2505.18327.

Forneron, J.-J. and L. Zhong (2025). Convexity not required: Estimation of smooth moment condition models. available at https://arxiv.org/abs/2304.14386v2.

Gadat, S. and F. Panloup (2023). Optimal non-asymptotic analysis of the Ruppert-Polyak averaging stochastic algorithm. *Stochastic Processes and their Applications 156*, 312–348.

Hall, P. and C. C. Heyde (1980). *Martingale Limit Theory and Its Application*. Academic Press, Boston,.

Hansen, B. E. and S. Lee (2021). Inference for iterated GMM under misspecification. *Econometrica 89*(3), 1419–1447.

Hansen, P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Kiefer, N. M., T. J. Vogelsang, and H. Bunzel (2000). Simple robust testing of regression hypotheses. *Econometrica 68*(3), 695–714.

Lee, S., Y. Liao, M. H. Seo, and Y. Shin (2022). Fast and robust online inference with stochastic gradient descent via random scaling. *Proceedings of the AAAI Conference on Artificial Intelligence 36*(7), 7381–7389.

Lee, S., Y. Liao, M. H. Seo, and Y. Shin (2025). Fast inference for quantile regression with tens of millions of observations. *Journal of Econometrics 249*, 105673.

Lee, S. and S. Ng (2020). An econometric perspective on algorithmic subsampling. *Annual Review of Economics 12*(1), 45–80.

Lee, S. and S. Ng (2022). Least squares estimation using sketched data with heteroskedastic errors. *Proceedings of the 39th International Conference on Machine Learning*.

Leung, M. F., K. W. Chan, and X. Shao (2025). Online generalized method of moments for time series. available at https://doi.org/10.48550/arXiv.2502.00751.

Lewbel, A. and K. Pendakur (2009). Tricks with Hicks: The EASI demand system. *American Economic Review 99*(3), 827–63.

Li, X., J. Liang, X. Chang, and Z. Zhang (2022). Statistical estimation and online inference via local SGD. In P.-L. Loh and M. Raginsky (Eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, Volume 178 of *Proceedings of Machine Learning Research*, pp. 1613–1661. PMLR.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Pendakur, K. (2015). EASI GMM moment evaluator code for Stata. available at www.sfu.ca/~pendakur.

Polyak, B. T. (1990). New method of stochastic approximation type. *Automation and Remote Control 51*(7), 937–946.

Polyak, B. T. and A. B. Juditsky (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization 30*(4), 838–855.

Robbins, H. and S. Monro (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics 22*(3), 400 – 407.

Robinson, P. M. (1988). The stochastic difference between econometric statistics. *Econometrica 56*(3), 531–548.

Ruppert, D. (1988). Efficient estimations from a slowly convergent Robbins–Monro process. Technical Report 781, Cornell University Operations Research and Industrial Engineering. available at https://ecommons.cornell.edu/bitstream/handle/1813/8664/TR000781.pdf?sequence=1.

Venkatraman, A., W. Sun, M. Hebert, J. Bagnell, and B. Boots (2016). Online instrumental variable regression with applications to online linear system identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 30.

A Approximation to Full-Sample GMM Estimate

In Appendix A, we focus on the first-order SLIM algorithm when the sample is treated as fixed. The goal is to approximate the full-sample GMM estimator

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \bar{g}_n(\theta)'\bar{g}_n(\theta),$$

which can be computationally burdensome when the sample size n is large. Unlike in the random sampling setting, we disregard the uncertainty in $\hat{\theta}_n$ and focus on quantifying the error of the stochastic approximation estimator, remaining agnostic about the nature of the data, whether it is cross-sectional, time series, or clustered.

A.1 Asymptotic Theory When Data are Treated as Fixed

Since the data are treated as fixed, we directly impose the regularity conditions on $\bar{g}_n(\cdot)$, $\bar{G}_n(\cdot)$, and the related moments. Recall that

$$\bar{G}_n := \frac{1}{n}\sum_{i=1}^{n} G(z_i, \hat{\theta}_n) \quad \text{and} \quad \bar{\Omega}_n := \frac{1}{n}\sum_{i=1}^{n} g(z_i, \hat{\theta}_n)g(z_i, \hat{\theta}_n)'.$$

Define the following quantities:

$$\bar{Q}_n(\theta) := \frac{1}{2}\bar{g}_n(\theta)'\bar{g}_n(\theta) - \frac{1}{2}\min_{\theta\in\Theta} \bar{g}_n(\theta)'\bar{g}_n(\theta),$$
$$\bar{\mathcal{H}}_n := \frac{\partial^2 \bar{Q}_n(\hat{\theta}_n)}{\partial\theta\partial\theta'} = \bar{G}_n'\bar{G}_n + \sum_{j=1}^{d_g} \bar{g}_{nj}(\hat{\theta}_n)\frac{\partial^2}{\partial\theta\partial\theta'}\bar{g}_{nj}(\hat{\theta}_n),$$

where $\bar{g}_{nj}(\theta)$ denotes the jth element of $\bar{g}_n(\theta)$ for $j = 1, \ldots, d_g$, and

$$\bar{\Sigma}_n := \frac{1}{B_g}\left(\bar{G}_n'\bar{\Omega}_n\bar{G}_n + \frac{1}{B_G}\cdot\frac{1}{n}\sum_{i=1}^{n}(G(z_i, \hat{\theta}_n) - \bar{G}_n)'\bar{\Omega}_n(G(z_i, \hat{\theta}_n) - \bar{G}_n)\right). \tag{31}$$

The function $\theta \mapsto \bar{Q}_n(\theta)$ serves as the Lyapunov function. All sample moments that do not depend on θ are evaluated at the full-sample GMM estimator $\hat{\theta}_n$. In particular, $\bar{G}_n$, $\bar{\Omega}_n$, $\bar{\mathcal{H}}_n$, and $\bar{\Sigma}_n$ are all computed at $\hat{\theta}_n$ in the definitions above.

Assumption 6 (Data are Fixed)**.** *The data $z_{1:n} = (z_i)_{i=1}^n$ are fixed. Moreover, the following holds for some finite constants $c > 0$, $\delta > 0$, and $M < \infty$:*

(A6.1) $\|\bar{G}_n(\theta)\| \leq M$ *for all* $\theta \in \Theta$;

(A6.2) $\|\bar{G}_n(\theta) - \bar{G}_n(\tilde{\theta})\| \leq M\|\theta - \tilde{\theta}\|$ *for all* θ *and* $\tilde{\theta} \in \Theta$;

(A6.3) For each $j = 1, \ldots, d_g$, $\|\frac{\partial^2 \bar{g}_{nj}}{\partial\theta\partial\theta'}(\theta) - \frac{\partial^2 \bar{g}_{nj}}{\partial\theta\partial\theta'}(\tilde{\theta})\| \leq M\|\theta - \tilde{\theta}\|$ *for all* θ *and* $\tilde{\theta}$ *in* $\{\theta \in \Theta : \|\theta - \hat{\theta}_n\| \leq \delta\}$;

(A6.4) $\frac{\partial^2 \bar{Q}_n(\theta)}{\partial\theta\partial\theta'} \geq c$ *for all* $\|\theta - \hat{\theta}_n\| \leq \delta$;

(A6.5) $\bar{Q}_n(\theta) \geq c$ *and* $\|\bar{G}_n(\theta)'\bar{g}_n(\theta)\|^2 \geq c$ *for all* $\|\theta - \hat{\theta}_n\| \geq \delta$;

(A6.6) $\frac{\partial^2 \bar{Q}_n(\theta)}{\partial\theta\partial\theta'} \leq M$ *for all* $\theta \in \Theta$;

Assumption 6 presents regularity conditions required in the fixed-sample asymptotic framework. Conditions (A6.4) through (A6.6) serve as deterministic analogues of (A4.6) – (A4.8) under random sampling. In (A6.4), the Hessian of the GMM criterion $\theta \mapsto \bar{g}_n(\theta)'\bar{g}_n(\theta)$ is assumed to be positive definite in the δ-neighborhood of its minimum $\hat{\theta}_n$, ensuring that $\hat{\theta}_n$ is locally well-separated. Combined with (A6.5), $\hat{\theta}_n$ must be the unique critical point and the global minimizer of the GMM objective function. Boundedness and smoothness conditions on the Jacobian and Hessian of the moment function are given in (A6.1) through (A6.3). We additionally impose Lipschitz continuity of the Hessian to ensure that the Hessian of $\bar{Q}_n(\theta)$ varies continuously with θ in the δ-neighborhood of $\hat{\theta}_n$. This condition was previously omitted under random sampling, where its effect was asymptotically negligible when multiplied by $\bar{g}_n(\hat{\theta}_n) = o_{\mathrm{P}}(1)$. Together, the conditions in Assumption 6 form a deterministic counterpart to Assumption 4, enabling us to establish similar convergence guarantees in settings where the data are assumed to be fixed.

Recall that in the main text, we introduced the following notation: $\mathbb{P}_n^*$ denotes the distribution induced by the mini-batch sample $(\tilde{z}_i)_{i\geq 1}$, and we write $\mathbb{E}_n^*$ for the corresponding expectation. For each $\theta \in \Theta$, define the stochastic error at θ by

$$\xi_t(\theta) = \tilde{G}_t(\theta)'\tilde{g}_t(\theta) - \bar{G}_n(\theta)'\bar{g}_n(\theta), \tag{32}$$

where $\tilde{G}_t(\theta)$ and $\tilde{g}_t(\theta)$ are defined in (5). By definition, $\mathbb{E}_n^*[\xi_t(\theta)|\tilde{z}_{1:B(t-1)}] = 0$ holds for all $t \geq 1$ and $\theta \in \Theta$, characterizing the martingale difference property of the stochastic errors. We note that $\bar{\Sigma}_n = \lim_{t\to\infty} \mathbb{E}_n^*[\xi_t(\theta_{t-1}^*)\xi_t(\theta_{t-1}^*)'|\mathcal{F}_{n,t-1}^*]$ corresponds to the steady-state variance of the stochastic error as θ_{t-1}^* tends to $\hat{\theta}_n$. We present additional assumptions on the behavior of the stochastic errors as θ evolves.

Assumption 7 (Stochastic Errors)**.** *The following hold for some finite constants* $M < \infty$ *and* $p > (1-a)^{-1}$*, where* $a \in (1/2, 1)$ *denotes the learning rate exponent defined in Assumption 2:*

(A7.1) $\mathbb{E}_n^*[\|\xi_t(\theta) - \xi_t(\hat{\theta}_n)\|^2] \leq C\|\theta - \hat{\theta}_n\|^2$ *for all* $\theta \in \Theta$;

(A7.2) $\mathbb{E}_n^*[\|\xi_t(\theta)\|^{2p}] \leq M(\bar{Q}_n(\theta)^p + 1)$ *for all* $\theta \in \Theta$.

(A7.1) requires that the second moment of the stochastic error depends smoothly on θ. This condition ensures that the conditional variance of $\xi_t(\theta_{t-1}^*)$ converges to $\bar{\Sigma}_n$ as θ_{t-1}^* approaches $\hat{\theta}_n$. Additionally, we require that the $2p$th moment of $\xi_t(\theta)$ grows at most on the order of $\bar{Q}_n(\theta)^p + 1$ for all $\theta \in \Theta$. This condition ensures the stability of our stochastic approximation algorithm, preventing θ_t^* from diverging due to explosive variability.

The following theorem establishes the consistency and the functional central limit theorem (FCLT) for the averaged stochastic approximation estimator when the data are treated as fixed.

Theorem 6. *Let Assumptions 1, 2, 3, 6, and 7 hold. Then, both* θ_N^* *and* $\bar{\theta}_N^*$ *converge* $\mathbb{P}_n^*$*-almost surely to* $\hat{\theta}_n$ *as* $N \to \infty$*. Moreover, as* $N \to \infty$*,*

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_t^* - \hat{\theta}_n) \rightsquigarrow^* (\bar{\mathcal{H}}_n^{-1} \bar{\Sigma}_n \bar{\mathcal{H}}_n^{-1})^{1/2} W(r).$$

The FCLT in Theorem 6 states that the normalized partial sum process of the stochastic iterates, recentered around its limit $\hat{\theta}_n$, converges weakly to a Wiener process whose variance is given by the sandwich formula $\bar{\mathcal{H}}_n^{-1} \bar{\Sigma}_n \bar{\mathcal{H}}_n^{-1}$. That is, $\bar{\mathcal{H}}_n^{-1}$ and $\bar{\Sigma}_n$ play the roles of the "bread" and "meat," respectively. As a special case of this FCLT, we obtain the asymptotic normality of the averaged estimator $\bar{\theta}_N^*$ (in other words, this is achieved by setting $r = 1$). This FCLT can be used for online inference on $\hat{\theta}_n$ as detailed in Section A.2.

A.2 Inference with Fixed Data

As in Section A.1, it is assumed that data are fixed and the full-sample GMM estimate $\hat{\theta}_n$ is computationally infeasible. Hence, we focus on inference about $\hat{\theta}_n$ via random scaling.

Consider testing $\ell \leq d$ linear restrictions

$$H_0^* : R\hat{\theta}_n = c,$$

where again R is an $(\ell \times d)$-dimensional known matrix of rank ℓ and c is an ℓ-dimensional known vector. Recall the definition of $V_t(R)$ in (9) and their iterative updating rule in (10)-(12). As in the setting with random data, the conventional Wald test based on $V_N(R)$ becomes asymptotically pivotal, due to FCLT established in Theorem 6. Formally,

Corollary 3. *Let the conditions assumed in Theorem 6 hold. Suppose that* $H_0^* : R\hat{\theta}_n = c$ *holds with* $\text{rank}(R) = \ell$. *Then, as* $N \to \infty$,

$$NB_g \left(R\bar{\theta}_N^* - c\right)' \left(B_g V_N(R)\right)^{-1} \left(R\bar{\theta}_N^* - c\right) \rightsquigarrow^* W(1)' \left(\int_0^1 \bar{W}(r)\bar{W}(r)'dr\right)^{-1} W(1).$$

When increasing n is costly, Corollary 3 provides a method for conducting inference on $\hat{\theta}_n$, offering uncertainty quantification for the stochastic approximation via random scaling.

B Warm-Start Algorithm

In this section, we describe a warm-start algorithm to effectively initialize the method introduced in Section 2.2. While our framework does not require a consistent initial estimator, a well-designed warm-start phase can substantially improve practical performance. This is achieved by reshuffling the data at the start of each epoch and running the algorithm over multiple epochs during initialization, thereby reducing redundant sampling that may hinder convergence. Unlike the main algorithm in (4), which simultaneously updates the moments and their derivatives using two independent mini-batches, the warm-start procedure adopts a sequential updating scheme: the inner loop updates the mini-batch for the moments while holding the derivative batch fixed, and the outer loop updates the derivative batch. The parameter vector is updated at every step. This nested-loop structure promotes greater stability during initialization and improves finite-sample performance in subsequent optimization. Moreover, the learning rate is updated at the epoch level rather than at each iteration. The full warm-start procedure, which combines random reshuffling with sequential updates, is

detailed in Algorithm 2.

Algorithm 2: Warm-Start Algorithm for SLIM

Input: Sample $z_{1:n} = (z_i)_{i=1}^n$; mini-batch size B_{ws}; number of epochs E_{ws}; learning rate schedule $(\gamma_e)_{e=1}^{E_{\text{ws}}}$; initial value $\theta_{0,\text{ws}}^*$ for a warm start.

Set $t \leftarrow 1$.
Initialize $\theta^* \leftarrow \theta_{0,\text{ws}}^*$ and $\bar{\theta}^* \leftarrow \theta_{0,\text{ws}}^*$.
for *epoch* $e = 1, \ldots, E_{\text{ws}}$ **do**
  Shuffle $\{1, \ldots, n\}$ into a permutation $\{m(1), \ldots, m(n)\}$ such that $m(i) \neq m(i')$ for $i \neq i'$.
  Set $\tilde{z}_{1:n} \leftarrow (z_{m(i)})_{i=1}^n$.
  for $j = 1, \ldots, \lfloor n/B_{\text{ws}} \rfloor$ **do**
    for $k = 1, \ldots, \lfloor n/B_{\text{ws}} \rfloor$, $k \neq j$ **do**
      Compute $\tilde{G}_j \leftarrow \frac{1}{B_{\text{ws}}} \sum_{i=(j-1)B_{\text{ws}}+1}^{jB_{\text{ws}}} G(\tilde{z}_i, \theta^*)$.
      Compute $\tilde{g}_k \leftarrow \frac{1}{B_{\text{ws}}} \sum_{i=(k-1)B_{\text{ws}}+1}^{kB_{\text{ws}}} g(\tilde{z}_i, \theta^*)$.
      Update $\theta^* \leftarrow \theta^* - \gamma_e \tilde{G}_j' \tilde{g}_k$.
      Update $\bar{\theta}^* \leftarrow \frac{t-1}{t}\bar{\theta}^* + \frac{1}{t}\theta^*$.
      Increment $t \leftarrow t + 1$.

Output: Final estimator $\bar{\theta}^*$, to be used as the initial value θ_0^* in Algorithm 1

C Alternative Asymptotic Regimes

Recall that we have written $(\bar{\theta}_N^* - \theta_o)$ as the sum of $(\bar{\theta}_N^* - \hat{\theta}_n)$ and $(\hat{\theta}_n - \theta_o)$. In Section 4.2.1, we have considered the case such that $N \asymp n$ and $B_G < \infty$. In this section, we consider three alternative regimes.

C.0.1 Intermediate Case II. $n \asymp N$ and $B_G \to \infty$

The conservativeness of the inference procedure in the previous case can be mitigated by allowing the batch size B_G to grow with the sample size. Specifically, note that $B_g \Sigma_o \to G_o' \Omega_o G_o$ as $B_G \to \infty$, thereby tightening the upper bound on the asymptotic variance. However, increasing B_G substantially raises the computational burden. In particular, the regime where $B_G \to \infty$ closely resembles second-order stochastic approximation methods. In summary, there is a trade-off: larger values of B_G improve statistical precision but come at the cost of higher computational complexity. We consider the implications of increasing B_G in Section 5.

C.0.2 Polar Case I. Very large n such that $N/n \to 0$

Consider the polar case where $N/n \to 0$. Then, by Theorem 2,

$$\begin{aligned}\sqrt{N}(\bar{\theta}_N^* - \theta_o) &= \sqrt{N}(\bar{\theta}_N^* - \hat{\theta}_n) + \left(\frac{N}{n}\right)^{1/2} \sqrt{n}(\hat{\theta}_n - \theta_o) \\ &= \sqrt{N}(\bar{\theta}_N^* - \hat{\theta}_n) + o_p(1),\end{aligned}$$

where the $o_p(1)$ term is with respect to the joint distribution of $(\hat{\theta}_n, \bar{\theta}_N^*)$. It follows that the limiting distribution of $\sqrt{N}(\bar{\theta}_N^* - \theta_o)$ is asymptotically equivalent to that of $\sqrt{N}(\bar{\theta}_N^* - \hat{\theta}_n)$. In this regime, inference can therefore proceed as if the data were fixed. That is, when the sample size is sufficiently large so that the stochastic approximation error dominates the sampling variability, the asymptotic distribution becomes insensitive to whether we center at the full-sample GMM estimator or at the true parameter vector. Similar points were made in the context of random sketching in linear models (see, e.g., Lee and Ng, 2020, 2022).

C.0.3 Polar Case II. Very large N such that $N/n \to \infty$

Consider the opposite polar case where $N/n \to \infty$. Then, by Theorem 2,

$$\begin{aligned}\sqrt{n}(\bar{\theta}_N^* - \theta_o) &= \left(\frac{n}{N}\right)^{1/2} \sqrt{N}(\bar{\theta}_N^* - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta_o) \\ &= \sqrt{n}(\hat{\theta}_n - \theta_o) + o_p(1).\end{aligned}$$

In this regime, the error due to stochastic approximation becomes asymptotically negligible, and the limiting distribution is governed entirely by the sampling variation in $\hat{\theta}_n$. This case corresponds closely to the conventional mode of inference in econometrics, where n is not assumed to be very large. In such settings, computational error is typically ignored, and inference is based solely on sampling variability. Since the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_o)$ is well understood, standard inference procedures can be applied by substituting consistent sample analogs for the unknown population quantities. This approach remains computationally feasible when n is moderate. Alternatively, inference may still be conducted using the random scaling approach. In this case, the Wald statistic simplifies to

$$n(R\bar{\theta}_N^* - c)'(B_g V_N(R))^{-1}(R\bar{\theta}_N^* - c).$$

D Lemmas and Proofs of the Main Results

Appendix D develops a sequence of lemmas for the main results in the paper. We focus on the random-sampling setting where the sample consists of n i.i.d. draws from a population distribution with n growing to infinity.

D.1 Setup

Notation. We adopt the following notation. Generic absolute positive constants are denoted by $\underline{c}$ and C, which may differ from line to line. Absolute constants are allowed to depend on (fixed) quantities appearing in Assumption 8, but not on the sample size n, the time index t, or the iteration numbers, $N^\star$ and T. We write $\mathbb{P}$ for the distribution of the sample $(z_i)_{i\geq 1}$ with the corresponding expectation denoted by $\mathbb{E}$. The conditional law and expectation of $(\tilde{z}_i)_{i\geq 1}$ are denoted by $\mathbb{P}_n^\star$ and $\mathbb{E}_n^\star$, respectively, given the sample $(z_i)_{i=1}^n$ of size n. We write $a \vee b$ (resp. $a \wedge b$) for $\max\{a, b\}$ (resp. $\min\{a, b\}$). Henceforth, we use the abbreviation w.p.a.1 for *with probability approaching one*.

Algorithm. We first set out a unified framework that encompasses both the first-order and second-order algorithms in the paper. Let $z_{1:n} = (z_i)_{i=1}^n$ denote the sample of size n. Let $(B_{G,t})_{t\geq 1}$ denote a sequence of possibly time-varying mini-batch sizes used to approximate the Jacobian of the moment function. Let $B_g \geq 1$ be a fixed mini-batch size used for the moment function. Let $B_t = \sum_{s=1}^t (B_{G,s} + B_g)$ denote the total number of mini-batch draws up to iteration t. Let $(\tilde{z}_i)_{i=1}^{B_T}$ denote mini-batch draws up to iteration T, each of which is sampled uniformly at random from $z_{1:n}$. Define the mini-batch approximations to the Jacobian and the moment function as follows:

$$\tilde{G}_t(\theta) = \frac{1}{B_{G,t}} \sum_{j=1}^{B_{G,t}} G(\tilde{z}_{j+B_{t-1}}, \theta), \quad \tilde{g}_t(\theta) = \frac{1}{B_g} \sum_{j=1}^{B_g} g(\tilde{z}_{j+B_{t-1}+B_{G,t}}, \theta).$$

We write P_n for a positive definite pre-conditioning matrix and W_n for a positive definite weighting matrix, both of which are allowed to be random and vary with n.

We use the following iterative algorithm in general form. Let $\theta_0^\star \in \Theta$ denote an

initial value. The stochastic iterates $(\theta_t^\star)_{t=0}^T$ evolve according to the updating rule,

$$\theta_t^\star = \theta_{t-1}^\star - \gamma_t^\star P_n \tilde{G}_t(\theta_{t-1}^\star)' W_n \tilde{g}_t(\theta_{t-1}^\star) \quad \text{for } t \geq 1, \tag{33}$$

where $(\gamma_t^\star)_{t\geq 1}$ denotes the learning rate (step size) schedule as specified below. Let $\gamma_0 > 0$ and $a \in (1/2, 1)$ be the initial learning rate and the learning rate exponent, respectively. The learning rate is now defined by $\gamma_t^\star = \gamma_0((t + N^\star) \vee 1)^{-a}, \quad t \geq 1$, where $N^\star \geq 0$ is an integer representing the number of first-order iterations in the SLIM algorithm. When $N^\star = 0$, $\gamma_t^\star$ reduces to the standard learning rate form $\gamma_0 t^{-a}$, which corresponds to no previous learning steps. In this case, the choice of $P_n = I_d$ and $W_n = I_{d_g}$ yields the first-order algorithm. Otherwise, $N^\star > 0$ indicates that the algorithm has already advanced $N^\star$ iterations. The second-order refinement step in Section 5 then takes the form (33), where $N^\star$ denotes the iteration number of the first-order algorithm, and the resulting stochastic iterate $\theta_{N^\star}^*$ is relabeled as $\theta_0^\star$.

After T iterations, where T is chosen by the user, the algorithm yields the averaged stochastic approximation estimator $\bar{\theta}_T^\star = \frac{1}{T}\sum_{t=1}^T \theta_t^\star$, which can also be updated recursively via Algorithm 1. The approximand is the full-sample GMM estimator $\hat{\theta}_{n,W}$ under the weighting matrix W_n:

$$\hat{\theta}_{n,W} = \arg\min_{\theta\in\Theta} \bar{g}_n(\theta)' W_n \bar{g}_n(\theta).$$

The sample Lyapunov function is accordingly defined as

$$\bar{Q}_{n,W}(\theta) = \frac{1}{2}\left(\bar{g}_n(\theta)' W_n \bar{g}_n(\theta) - \bar{g}_n(\hat{\theta}_{n,W})' W_n \bar{g}_n(\hat{\theta}_{n,W})\right) \geq 0.$$

Additionally, we define the following moment quantities:

$$\begin{aligned}
\bar{\Omega}_n(\theta) &= \frac{1}{n}\sum_{i=1}^n g(z_i,\theta)g(z_i,\theta)', \quad \tilde{\Omega}_n(\theta) = \bar{\Omega}_n(\theta) - \bar{g}_n(\theta)\bar{g}_n(\theta)', \quad \theta \in \Theta, \\
\bar{\Omega}_n &= \bar{\Omega}_n(\hat{\theta}_{n,W}), \quad \tilde{\Omega}_n = \tilde{\Omega}_n(\hat{\theta}_{n,W}), \quad \bar{G}_n = \bar{G}_n(\hat{\theta}_{n,W}), \\
\bar{\mathcal{H}}_n &= \frac{\partial^2 \bar{Q}_{n,W}(\hat{\theta}_{n,W})}{\partial\theta\partial\theta'} = \bar{G}_n' W_n \bar{G}_n + \sum_{j=1}^{d_g} \bar{g}_{nj}(\hat{\theta}_{n,W})' W_n \frac{\partial^2 \bar{g}_{nj}(\hat{\theta}_{n,W})}{\partial\theta\partial\theta'}.
\end{aligned}$$

Let $\xi_t(\theta)$ denote the stochastic approximation error at iteration t,

$$\xi_t(\theta) = \tilde{G}_t(\theta)' W_n \tilde{g}_t(\theta) - \bar{G}_n(\theta)' W_n \bar{g}_n(\theta), \quad \theta \in \Theta,$$

so that (33) can be rewritten as

$$\theta_t^\star = \theta_{t-1}^\star - \gamma_t^\star P_n \bar{G}_n(\theta_{t-1}^\star)' W_n \bar{g}_n(\theta_{t-1}^\star) - \gamma_t^\star P_n \xi_t(\theta_{t-1}^\star).$$

Let $\Sigma_{n,t}(\theta)$ denote the covariance matrix of $\xi_t(\theta)$,

$$\Sigma_{n,t}(\theta) = \frac{1}{B_g} \left(\begin{array}{l} \bar{G}_n(\theta)' W_n \tilde{\Omega}_n(\theta) W_n \bar{G}_n(\theta) \\ \quad + \dfrac{1}{n B_{G,t}} \displaystyle\sum_{i=1}^{n} (G(z_i,\theta) - \bar{G}_n(\theta))' W_n \tilde{\Omega}_n(\theta) W_n (G(z_i,\theta) - \bar{G}_n(\theta)) \end{array} \right).$$

and write $\Sigma_{n,t} = \Sigma_{n,t}(\hat{\theta}_{n,W})$. The asymptotic covariance matrix is then given by

$$\Sigma_t = \operatorname*{plim}_{n\to\infty} \Sigma_{n,t} = \frac{1}{B_g} \left(G_o' W \Omega_o W G_o + \frac{1}{B_{G,t}} \mathbb{E}\left[(G(z_i,\theta_o) - G_o)' W \Omega_o W (G(z_i,\theta_o) - G_o) \right] \right).$$

The time-stationary covariance, $\Sigma_n = \operatorname{plim}_{t\to\infty} \Sigma_{n,t}$, is defined by

$$\frac{1}{B_g} \left(\bar{G}_n' W_n \tilde{\Omega}_n W_n \bar{G}_n + \frac{1}{n B_G} \sum_{i=1}^{n} \left[(G(z_i,\hat{\theta}_{n,W}) - \bar{G}_n)' W_n \tilde{\Omega}_n W_n (G(z_i,\hat{\theta}_{n,W}) - \bar{G}_n) \right] \right).$$

Finally, let Σ_o denote the asymptotic variance as both $n \to \infty$ and $t \to \infty$:

$$\lim_{t\to\infty} \Sigma_t = \operatorname*{plim}_{n\to\infty} \Sigma_n = \frac{1}{B_g} \left(G_o' W \Omega_o W G_o + \frac{1}{B_G} \mathbb{E}[(G(z_i,\theta_o) - G_o)' W \Omega_o W (G(z_i,\theta_o) - G_o)] \right).$$

To describe the dynamic properties of the stochastic iterates, we define the natural filtration as the sequence of sigma-algebras generated by the mini-batch sequence and the initial information:

$$\mathcal{F}_{n,t}^\star = \sigma((\tilde{z}_i)_{1 \le i \le B_{t-1}}, \mathcal{F}_{n,0}^\star), \quad t \ge 1,$$

where $\mathcal{F}_{n,0}^\star$ denotes the sigma-algebra of observables at the outset of the algorithm, including the sample $z_{1:n}$.

We collectively state the necessary assumptions in the following.

Assumption 8. *The following holds.*

(A8.1) $\theta_0^\star$ *satisfies the following bounds for* $M < \infty$*:*

$$\|\theta_0^\star\| \leq M, \quad \bar{Q}_{n,W}(\theta_0^\star) \leq M\gamma_0^\star.$$

(A8.2) $P_n \xrightarrow{\mathbb{P}} P$ *and* $W_n \xrightarrow{\mathbb{P}} W$*, where* $P > 0$ *and* $W > 0$ *are non-random matrices;*

(A8.3) As $t \to \infty$*,* $B_{G,t} \to B_G$ *where either* $B_G \in \mathbb{N}$ *or* $B_G = \infty$*;*

(A8.4) The following holds for some constants $c > 0$*,* $\delta > 0$*,* $M < \infty$*, and* $p \geq 1$ *:*

- *(A8.4.1)* $(z_i)_{i=1}^n \subseteq \mathcal{Z}$ *are i.i.d. draws from the population distribution* $\mathbb{P}$*;*
- *(A8.4.2) there exists a unique* θ_o *that solves* $g(\theta) = 0$*;*
- *(A8.4.3)* $G_o' G_o \geq cI_d$ *and* $\Omega_o \geq cI_{d_g}$*;*
- *(A8.4.4) It is assumed that* $\mathbb{E}[\|g(z_i, \theta_o)\|^{2p}] < \infty$*. Moreover, there exists a measurable function* $H : \mathcal{Z} \to \mathbb{R}$ *such that* $\|G(z, \theta)\| \leq H(z)$ *for all* $z \in \mathcal{Z}$ *and* $\theta \in \Theta$*, and* $\mathbb{E}[H(z_i)^{2p}] < \infty$*;*
- *(A8.4.5)* $\|G(z, \theta) - G(z, \tilde{\theta})\| \leq L(z)\|\theta - \tilde{\theta}\|$ *holds for any* $z \in \mathcal{Z}$ *and for all* θ *and* $\tilde{\theta}$ *in* Θ*, where* $L(\cdot)$ *satisfies* $\mathbb{E}[L(z_i)^{2p}] < \infty$*;*
- *(A8.4.6)* $\frac{\partial^2 \bar{Q}_{n,W}(\theta)}{\partial\theta\partial\theta'} \geq cI_d$ *for all* $\|\theta - \hat{\theta}_{n,W}\| \leq \delta$ *w.p.a.*1*;*
- *(A8.4.7)* $\bar{Q}_{n,W}(\theta) \geq c$ *and* $\|\frac{\partial \bar{Q}_{n,W}(\theta)}{\partial\theta}\|^2 \geq c$ *for all* $\|\theta - \hat{\theta}_{n,W}\| \geq \delta$ *w.p.a.*1*;*
- *(A8.4.8)* $\frac{\partial^2 \bar{Q}_{n,W}(\theta)}{\partial\theta\partial\theta'} \leq MI_d$ *and* $\frac{1}{n}\sum_{i=1}^n \|g(z_i, \theta)\|^{2p} \leq M(\bar{Q}_{n,W}(\theta)^p + 1)$ *for all* $\theta \in \Theta$ *w.p.a.*1*.*

D.2 Key Lemmas

We begin with the standard large-sample properties of the full-sample GMM estimator.

Lemma 1. *Under Assumption 8,* $\hat{\theta}_{n,W} \xrightarrow{\mathbb{P}} \theta_o$ *and,*

$$\sqrt{n}(\hat{\theta}_{n,W} - \theta_o) \xrightarrow{d} \mathcal{N}(0, (G_o' W G_o)^{-1} G_o' W \Omega_o W G_o (G_o' W G_o)^{-1}).$$

Next, we collect useful asymptotic bounds for the random quantities in our analysis, implied by Assumption 8.

Lemma 2. *Let Assumption 8 hold. Then, the following holds w.p.a.*1*:*

$$\underline{c} I_d \leq P_n \leq C I_d, \tag{D.1}$$
$$\underline{c} I_{d_g} \leq W_n \leq C I_{d_g}, \tag{D.2}$$
$$\underline{c} I_d \leq \bar{G}_n' W_n \bar{G}_n \leq C I_d, \tag{D.3}$$
$$\underline{c} I_d \leq \bar{\mathcal{H}}_n \leq C I_d, \tag{D.4}$$
$$\|\bar{G}_n(\theta)\| \leq C \text{ for all } \theta \in \Theta, \tag{D.5}$$
$$\|\bar{G}_n(\theta) - \bar{G}_n(\tilde{\theta})\| \leq C\|\theta - \tilde{\theta}\| \quad \text{for all } \theta, \tilde{\theta} \in \Theta, \tag{D.6}$$
$$\underline{c}(\|\theta - \hat{\theta}_{n,W}\|^2 \wedge \delta^2) \leq \bar{Q}_{n,W}(\theta) \quad \text{for all } \theta \in \Theta, \tag{D.7}$$
$$\underline{c}(\bar{Q}_{n,W}(\theta) \wedge 1) \leq \left\| \frac{\partial \bar{Q}_{n,W}(\theta)}{\partial \theta} \right\|^2 \quad \text{for all } \theta \in \Theta, \tag{D.8}$$
$$\mathbb{E}_n^\star[\|\xi_t(\theta) - \xi_t(\hat{\theta}_{n,W})\|^2] \leq C\|\theta - \hat{\theta}_{n,W}\|^2 \quad \text{for all } \theta \in \Theta \text{ and } t \geq 1, \tag{D.9}$$
$$\mathbb{E}_n^\star[\|\xi_t(\theta)\|^{2p}] \leq C(\bar{Q}_{n,W}(\theta)^p + 1) \quad \text{for all } \theta \in \Theta \text{ and } t \geq 1. \tag{D.10}$$

For each $n \in \mathbb{N}$, define E_n as the event on which (D.1)–(D.10) hold. Then, Assumption 8 and Lemma 2 ensure that $\mathbb{P}(E_n) \to 1$. Note that E_n depends on $z_{1:n}$ and is thus measurable with respect to $\sigma(z_{1:n})$.

Relative to the natural filtration $(\mathcal{F}_{n,t}^\star)_{t \geq 0}$, we define a stopping time

$$T_K^\star = \inf\left\{t \geq 0 : \bar{Q}_{n,W}(\theta_t^\star) \geq K\right\} \quad \text{for } K > 0,$$

where we set $\inf \varnothing \equiv \infty$. We also define a tail-measurable time

$$\tau_\varepsilon^\star = \inf\left\{ s \geq 0 : \sup_{t \geq s} \bar{Q}_{n,W}(\theta_t^\star) < \varepsilon \right\} \quad \text{for } \varepsilon > 0. \tag{34}$$

The next lemma concerns the concentration and tightness of these random times conditional on the event E_n.

Lemma 3. *Let Assumption 8 hold. Then the following hold.*

(i) Let $n \in \mathbb{N}$ be fixed. As $T \to \infty$, $\|\theta_T^\star - \hat{\theta}_{n,W}\| \xrightarrow{\mathbb{P}^*} 0$ on the event E_n.

(ii) For any $\varepsilon > 0$, there exists an absolute constant $K < \infty$ independent of $N^\star$ such

that, on the event E_n, uniformly in $N^\star \geq 0$,

$$\sup_{n\in\mathbb{N}} \mathbb{P}_n^\star \left(T_K^\star < \infty\right) < \varepsilon.$$

(iii) For any $\varepsilon > 0$, there exists an absolute constant $\tilde{T} \in \mathbb{N}$ independent of $N^\star$ such that, on the event E_n, uniformly in $N^\star \geq 0$,

$$\sup_{n\in\mathbb{N}} \mathbb{P}_n^\star \left(\tau_\varepsilon^\star > \tilde{T}\right) < \varepsilon.$$

The next lemma establishes an explicit convergence rate of the Lyapunov function evaluated at stochastic iterates.

Lemma 4. *Let Assumption 8 hold. For any $K > 0$, there exists an absolute constant $C_K > 0$ such that, on the event E_n,*

$$\mathbb{E}_n^\star[\bar{Q}_{n,W}(\theta_t^\star)\mathbb{1}\left\{T_K^\star \geq t\right\}] \leq C_K \gamma_t^\star$$

for all $n \in \mathbb{N}$, $N^\star \geq 0$, and $t \geq 1$.

To approximate the distribution of the stochastic approximation estimator, we introduce a first-order difference sequence $(\theta_t^\circ)_{t\geq 0}$, defined as follows:

$$\theta_t^\circ = \theta_{t-1}^\circ - \gamma_t^\star P_n \bar{\mathcal{H}}_n(\theta_{t-1}^\circ - \hat{\theta}_{n,W}) - \gamma_t^\star P_n \xi_t(\theta_{t-1}^\star), \quad \forall t \geq 1, \tag{35}$$

with the same initial value $\theta_0^\circ = \theta_0^\star$ as that of $(\theta_t^\star)_{t\geq 0}$.

The following lemma shows that the partial sum trajectory $(\sum_{s=1}^t \theta_s^\star)_{1\leq t\leq T}$ of the stochastic iterates $(\theta_t^\star)_{t\geq 1}$ can be uniformly approximated by the partial sum trajectory of $(\theta_t^\circ)_{t\geq 1}$.

Lemma 5. *Let Assumption 8 hold. As $n \to \infty$, $T \to \infty$, and $T^{1-a}/n \to 0$, it holds, for every $\varepsilon > 0$,*

$$\sup_{N^\star \geq 0} \mathbb{P}\left[\mathbb{P}_n^\star\left(\sup_{1\leq t\leq T}\left\|\sum_{s=1}^t(\theta_s^\star - \theta_s^\circ)\right\| > \sqrt{T}\varepsilon\right) \geq \varepsilon\right] \to 0.$$

The next two lemmas establish that, under suitable moment conditions and growth conditions on the pair $(N^\star, T)$ relative to n and each other, the CLT and FCLT apply to the approximating sequence, thereby enabling distributional approximation of the stochastic approximation estimator.

Lemma 6. *Let Assumption 8 hold for $p > (1-a)^{-1}$. Then, as $n \to \infty$, $T \to \infty$, $N^\star/T = O(1)$, it holds*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} (\theta_t^\circ - \hat{\theta}_{n,W}) \rightsquigarrow \left((G_o' W G_o)^{-1} \Sigma_o (G_o' W G_o)^{-1} \right)^{1/2} W(r).$$

Lemma 7. *Let Assumption 8 hold for $p > 1$. Then, as $n \to \infty$, $T \to \infty$, $N^\star/T = O(1)$, it holds*

$$\sqrt{T}(\bar{\theta}_T^\circ - \hat{\theta}_{n,W}) \xrightarrow{d} \mathcal{N}\left(0, (G_o' W G_o)^{-1} \Sigma_o (G_o' W G_o)^{-1}\right),$$

where the limiting random variable is independent of $(z_i)_{i \geq 1}$.

Other Technical Lemmas.

Lemma 8. *Suppose that X_n, A_n, C_n, and D_n are finite, non-negative random variables, adapted to the filtration $\{\mathcal{F}_n\}_{n=0}^\infty$, which satisfy*

$$\mathbb{E}[X_{n+1} | \mathcal{F}_n] \leq (1 + A_n) X_n + C_n - D_n.$$

Then, on the event $\{\sum_{n=1}^\infty A_n < \infty,\ \sum_{n=1}^\infty C_n < \infty\}$, we have

$$\sum_{n=1}^\infty D_n < \infty \quad \text{and} \quad X_n \to X$$

almost surely for some random variable X.

Lemma 9. *Let $\gamma_t = \gamma_0 t^{-a}$ for $t \geq 1$. Assume that $(v_n)_{n \geq 0}$ is a real sequence that satisfies for all $n \geq n_0$ ($n_0 \geq 0$) and for a given $\mu \in (0, 1/\gamma_{n_0+1})$:*

$$v_{n+1} \leq (1 - \mu \gamma_{n+1}) v_n + C \gamma_{n+1},$$

then for all $n \geq n_0$,

$$v_n \leq \max\{C, 1\}(v_{n_0} + \mu^{-1}).$$

For a d-dimensional matrix A, denote its eigenvalues by $\lambda_i(A)$, $i = 1, \ldots, d$, counted with multiplicity. We write $\mu_{\max}(A)$ and $\mu_{\min}(A)$ for $\max_{1\le i\le d} \mathrm{Re}(\lambda_i(A))$ and $\min_{1\le i\le d} \mathrm{Re}(\lambda_i(A))$, respectively.

Lemma 10. *Let A be a d-dimensional matrix such that $\mu_{\min}(A) > 0$. Then, there exists a constant $M_1(A, a, \gamma_0) < \infty$ that depend continuously on A, a, and γ_0, but not on $N^\star$, such that*

$$\|\alpha_s^t(A)\| \le M_1(A, a, \gamma_0)$$

for all $0 \le s \le t$ and $t \ge 1$, where $\alpha_s^t(A)$ is defined as[6]

$$\alpha_s^t(A) = \gamma_s^\star \sum_{i=s}^{t} \prod_{k=s+1}^{i} (I_d - \gamma_k^\star A).$$

Moreover, there exists a constant $M_2(A, a, \gamma_0) < \infty$ that depends continuously on A, a, and γ_0, but not on $N^\star$, such that, for all $t \ge 1$ and $N^\star \ge 0$,

$$\sum_{s=1}^{t} \|\alpha_s^t(A) - A^{-1}\| \le M_2(A, a, \gamma_0)(T + N^\star)^a.$$

Lemmas for the Fixed-Sample Case.

Lemma 11. *Let $z_{1:n}$ and $n \in \mathbb{N}$ be fixed. Let $\varepsilon > 0$ be arbitrary. Under Assumptions 6 and 7, as $N \to \infty$, it holds*

$$\mathbb{P}_n^* \left(\sup_{1\le t\le N} \left\| \sum_{s=1}^{t} (\theta_s^* - \theta_s^\circ) \right\| > \sqrt{N}\varepsilon \right) \to 0.$$

Lemma 12. *Let $z_{1:n}$ and $n \in \mathbb{N}$ be fixed. Under Assumptions 6 and 7, as $N \to \infty$,*

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_s^\circ - \hat{\theta}_n) \rightsquigarrow \left(\bar{\mathcal{H}}_n^{-1} \Sigma_n \bar{\mathcal{H}}_n^{-1}\right)^{1/2} W(r).$$

E Proofs of Main Theorems

Theorems 2–3 concern the stochastic approximation estimator produced by our first-order SLIM algorithm, in which $N^\star = 0$ and both P_n and W_n are chosen to be

[6] We follow the convention that $\prod_{k=s+1}^{s} \cdot$ is an identity matrix.

identity matrices. In the proof of Theorems 2–3, we use N to denote the number of iterations. We suppress W in the subscripts of $\hat{\theta}_{n,W}$ and $\bar{Q}_{n,W}(\cdot)$ and write $\hat{\theta}_n$ and $\bar{Q}_n(\theta)$ instead since $W_n = I_{d_g}$. Let $Q_t^* = \bar{Q}_n(\theta_t^*)$. We also use the following definitions:

$$T_K^* = \inf\{t \geq 0 : Q_t^* \geq K\}, \quad \text{and} \quad \tau_\varepsilon^* = \inf\left\{t \geq 0 : \sup_{s \geq t} Q_s^* \leq \varepsilon\right\}.$$

In contrast, the second-order SLIM algorithm, as used in Theorems 4 and 5, continues from $N^\star = N$ iterations in the preliminary first-order step.

E.1 Proof of Theorem 2

We begin by writing

$$\sqrt{N}(\bar{\theta}_N^* - \hat{\theta}_n) = \sqrt{N}(\bar{\theta}_N^\circ - \hat{\theta}_n) + \sqrt{N}(\bar{\theta}_N^* - \bar{\theta}_N^\circ) =: J_1 + J_2.$$

For the first term, apply Lemma 7 to get $J_1 \xrightarrow{d} ((G_o'G_o)^{-1}\Sigma_o(G_o'G_o)^{-1})^{1/2}Z \quad \text{as} \quad n \to \infty, N \to \infty$. The second term converges in probability to zero by Lemma 5 since

$$\|J_2\| \leq \frac{1}{\sqrt{N}} \sup_{1 \leq t \leq N} \left\| \sum_{s=1}^{t} (\theta_s^* - \theta_s^\circ) \right\| \xrightarrow{\mathbb{P}} 0$$

under $n \to \infty$, $N \to \infty$, and $N^{1-a}/n \to 0$. Putting these together, we have

$$\sqrt{N}(\bar{\theta}_N^* - \hat{\theta}_n) \xrightarrow{d} ((G_o'G_o)^{-1}\Sigma_o(G_o'G_o)^{-1})^{1/2}Z$$

as $n \to \infty$, $N \to \infty$, and $N^{1-a}/n \to 0$. Finally, note that $\sqrt{n}(\hat{\theta}_n - \theta_o) \xrightarrow{d} \mathcal{N}(0, (G_o'G_o)^{-1}G_o'\Omega_o G_o(G_o'G_o)^{-1})$ by Lemma 1. Since $\sqrt{n}(\hat{\theta}_n - \theta_o)$ depends on $z_{1:n}$ and is asymptotically independent of $\sqrt{N}(\bar{\theta}_N^* - \hat{\theta}_n)$, the proof is complete.

E.2 Proof of Theorem 3

We begin by writing

$$I(r) = \frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_t^* - \hat{\theta}_n) = \frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_t^\circ - \hat{\theta}_n) + \frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_t^* - \theta_t^\circ)$$
$$=: J_1(r) + J_2(r), \quad r \in [0, 1].$$

Lemma 6 then shows $J_1(\cdot) \rightsquigarrow ((G_o'G_o)^{-1}\Sigma_o(G_o'G_o)^{-1})^{1/2} W(\cdot)$ as $n \to \infty$ and $N \to \infty$. To address the remainder term, we show that $\|J_2\|_\infty := \sup_{0\le r\le 1} \|J_2(r)\|$ converges in probability to zero. For every $\varepsilon > 0$, $n \in \mathbb{N}$, and $N \in \mathbb{N}$, note that

$$\mathbb{P}\left(\|J_2\|_\infty > \varepsilon | z_{1:n}\right) \le \mathbb{1}_{E_n^c} + \mathbb{P}_n^*\left(\|J_2\|_\infty > \varepsilon\right) \mathbb{1}_{E_n}.$$

Letting $n \to \infty$ and $N \to \infty$ at the rate $N^{1-a}/n \to 0$, Lemma 5 implies

$$\mathbb{1}_{E_n^c} + \mathbb{P}_n^*\left(\|J_2\|_\infty > \varepsilon\right) \mathbb{1}_{E_n} \xrightarrow{\mathbb{P}} 0,$$

thereby establishing $\|J_2\|_\infty \xrightarrow{\mathbb{P}} 0$. Putting these pieces together, we obtain

$$I(\cdot) \rightsquigarrow \left((G_o'G_o)^{-1}\Sigma_o(G_o'G_o)^{-1}\right)^{1/2} W(\cdot).$$

Lemma 1 states that $\sqrt{n}(\hat{\theta}_n - \theta_o) \xrightarrow{d} \mathcal{N}(0, (G_o'G_o)^{-1}G_o'\Omega_o G_o(G_o'G_o)^{-1})$. Since $\sqrt{n}(\hat{\theta}_n - \theta_o)$ and $I(\cdot)$ are asymptotically independent of each other, their joint asymptotic distribution is given by

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta_o) \\ I(r) \end{pmatrix} \rightsquigarrow \begin{pmatrix} ((G_o'G_o)^{-1}G_o'\Omega_o G_o(G_o'G_o)^{-1})^{1/2} Z \\ ((G_o'G_o)^{-1}\Sigma_o(G_o'G_o)^{-1})^{1/2} W(r) \end{pmatrix}.$$

E.3 Proof of Theorem 4

We proceed by verifying Assumptions (A8.1) and (A8.2) under the assumptions in Theorem 4.

Proof of (A8.1). First of all, the initial value satisfies $\theta_N^* = O_{\mathrm{P}}(1)$ because θ_N^* is consistent for θ_o by Theorem 1. Thus, we can find $M < \infty$ such that $\|\theta_N^*\| \le M$ with probability greater than $1 - \varepsilon$ for any given $\varepsilon > 0$. On the other hand, Lemma 3 implies that, for any given $\varepsilon > 0$, there exists $K > 0$ such that

$$\sup_{n\in\mathbb{N}} \mathbb{P}_n^*(T_K^* < \infty)\mathbb{1}_{E_n} < \varepsilon.$$

Moreover, Lemma 4 implies that, for such a $K > 0$, there exists $C_K < \infty$ such that

$$\mathbb{E}_n^*[\bar{Q}_n(\theta_N^*)\mathbb{1}\{T_K^* \ge N\}]\mathbb{1}_{E_n} \le C_K \gamma_0 N^{-a}, \quad \forall n, N \in \mathbb{N}.$$

Let $\mathcal{E}_{N,n} := \{T_K^* \geq N\} \cap E_n$ so that

$$\sup_{N\in\mathbb{N}} \mathbb{P}(\mathcal{E}_{N,n}^c) \leq \mathbb{E}[\mathbb{P}_n^*(T_K^* < \infty)\mathbb{1}_{E_n}] + \mathbb{P}(E_n^c) \leq \varepsilon + \mathbb{P}(E_n^c), \quad \forall n \in \mathbb{N}.$$

We choose n_0 such that $\mathbb{P}(E_n^c) < \varepsilon$ for all $n \geq n_0$, and hence $\sup_{N\in\mathbb{N}} \mathbb{P}(\mathcal{E}_{N,n}^c) < 2\varepsilon$ for all $n \geq n_0$. Since

$$\mathrm{E}[\bar{Q}_n(\theta_N^*)|\mathcal{E}_{N,n}] = \mathbb{E}\left[\mathbb{E}_n^*[\bar{Q}_n(\theta_N^*)\mathbb{1}\{T_K^* \geq N\}]\mathbb{1}_{E_n}\right]/\mathbb{P}(\mathcal{E}_{N,n}) \leq C_K\gamma_0 N^{-a},$$

we can find an absolute constant $C_1 = C_1(K,\varepsilon) < \infty$ such that

$$\mathbb{P}\left(\bar{Q}_n(\theta_N^*) \leq C_1\gamma_0 N^{-a}, \mathcal{E}_{N,n}\right) \geq 1 - 3\varepsilon \quad \text{for all } n \geq n_0 \text{ and } N \in \mathbb{N}.$$

To complete the proof, we now show that $\bar{Q}_n(\theta_N^*) \leq C_1\gamma_0 N^{-a}$ implies that $\bar{Q}_{n,W}(\theta_N^*) \leq C(\gamma_0 N^{-a} + n^{-1})$ with probability $\geq 1-\varepsilon$, thereby suggesting that Assumption (A8.1) is fulfilled with probability at least $1-4\varepsilon$ for any given ε and all sufficiently large n. As $N \to \infty$, on the event $\mathcal{E}_{N,n} \subseteq E_n$, we have

$$\underline{c}(\|\theta_N^* - \hat{\theta}_n\|^2 \wedge 1) \leq \bar{Q}_n(\theta_N^*) \leq C_1\gamma_0 N^{-a} < \underline{c},$$

and hence $\|\theta_N^* - \hat{\theta}_n\|^2 \leq (C_1/\underline{c})\gamma_0 N^{-a}$. Since $\sqrt{n}\|\hat{\theta}_n - \hat{\theta}_{n,W}\| = O_{\mathbb{P}}(1)$,

$$\mathbb{P}\left(\|\hat{\theta}_n - \hat{\theta}_{n,W}\|^2 \leq C_2/n\right) \geq 1 - \varepsilon$$

for some absolute constant $C_2 = C_2(\varepsilon)$. Since $\bar{Q}_{n,W}(\theta) \leq C\|\theta - \hat{\theta}_{n,W}\|^2$ for all $\theta \in \Theta$, we have

$$\bar{Q}_{n,W}(\theta_N^*) \leq C\|\theta_N^* - \hat{\theta}_{n,W}\|^2 \leq C(\|\theta_N^* - \hat{\theta}_n\|^2 + \|\hat{\theta}_n - \hat{\theta}_{n,W}\|^2) \leq C(\gamma_0 N^{-a} + n^{-1})$$

on the event $\mathcal{E}_{N,n} \cap \{\|\hat{\theta}_n - \hat{\theta}_{n,W}\|^2 \leq C_2/n\}$ that has probability at least $1-4\varepsilon$. This verifies $\bar{Q}_{n,W}(\theta_N^*) \leq C(\gamma_0 N^{-a} + n^{-1}) \leq C\gamma_0 N^{-a}$ with arbitrarily high probability under $N^a/n = O(1)$.

Proof of (A8.2). The consistency of Φ_n and W_{MB} in (16) as $n \to \infty$, $M_{\mathrm{MB}} \to \infty$, and $N \to \infty$ follows from the standard argument on plug-in estimation, and thus will be omitted for brevity. By continuous mapping theorem, it is evident that

$P_n = (\Phi_n' W_{\text{MB}} \Phi_n)^\dagger \xrightarrow{\mathbb{P}} P = (G_o' \Omega_o^{-1} G_o)^{-1}$.

Proof of Part (i): Let us write

$$\sqrt{T-N}(\bar{\theta}_T^* - \hat{\theta}_{n,W}) = J_1 + J_2,$$

where

$$J_1 = \sqrt{T-N}(\bar{\theta}_T^\circ - \hat{\theta}_{n,W}) = \frac{1}{\sqrt{T-N}} \sum_{t=N+1}^{T} (\theta_t^\circ - \hat{\theta}_{n,W}),$$
$$J_2 = \sqrt{T-N}(\bar{\theta}_T^* - \bar{\theta}_T^\circ) = \frac{1}{\sqrt{T-N}} \sum_{t=N+1}^{T} (\theta_t^* - \theta_t^\circ).$$

Here, the approximating sequence $(\theta_t^\circ)_{t=N}^T$ is defined as in (35), except that approximation starts at $t = N$ with $\theta_N^\circ = \theta_N^*$. Since $N/(T-N) = O(1)$ under the given assumptions, Lemma 7 applies to J_1, yielding

$$J_1 \xrightarrow{d} \mathcal{N}(0, (G_o' \Omega_o^{-1} G_o)^{-1} \Sigma_o (G_o' \Omega_o^{-1} G_o)^{-1}),$$

where the limiting normal random variable is independent of $(z_i)_{i \geq 1}$, and

$$\Sigma_o = \frac{1}{B_g} \left(G_o' \Omega_o^{-1} G_o + \frac{1}{B_G} \mathbb{E} \left[(G(z_i, \theta_o) - G_o)' \Omega_o^{-1} (G(z_i, \theta_o) - G_o) \right] \right).$$

The same argument as in the proof of Theorem 2 implies $J_2 \xrightarrow{\mathbb{P}} 0$ under $(T-N)^{1-a}/n \to 0$. Putting J_1 and J_2 together, we obtain

$$\sqrt{T-N}(\bar{\theta}_T^* - \hat{\theta}_{n,W}) \xrightarrow{d} \mathcal{N}(0, (G_o' \Omega_o^{-1} G_o)^{-1} \Sigma_o (G_o' \Omega_o^{-1} G_o)^{-1}).$$

Since $\sqrt{n}(\hat{\theta}_{n,W} - \theta_o) \xrightarrow{d} \mathcal{N}(0, (G_o' \Omega_o^{-1} G_o)^{-1} G_o' \Omega_o G_o (G_o' \Omega_o^{-1} G_o)^{-1})$ by Lemma 1 and is asymptotically independent of J_1, we conclude

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_{n,W} - \theta_o) \\ \sqrt{T-N}(\bar{\theta}_T^* - \hat{\theta}_{n,W}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} ((G_o' \Omega_o^{-1} G_o)^{-1} G_o' \Omega_o G_o (G_o' \Omega_o^{-1} G_o)^{-1})^{1/2} Z \\ ((G_o' \Omega_o^{-1} G_o)^{-1} \Sigma_o (G_o' \Omega_o^{-1} G_o)^{-1})^{1/2} W \end{pmatrix}.$$

Part (ii): Start with the decomposition

$$\begin{aligned} I(r) &= \frac{1}{\sqrt{T-N}} \sum_{t=N+1}^{N+\lfloor (T-N)r \rfloor} (\theta_t^* - \hat{\theta}_{n,W}) \\ &= \frac{1}{\sqrt{T-N}} \sum_{t=1}^{\lfloor (T-N)r \rfloor} (\theta_{t+N}^\circ - \hat{\theta}_{n,W}) + \frac{1}{\sqrt{T-N}} \sum_{t=1}^{\lfloor (T-N)r \rfloor} (\theta_{t+N}^* - \theta_{t+N}^\circ) \\ &=: J_1(r) + J_2(r), \quad r \in [0,1]. \end{aligned}$$

By Lemma 6, under $p > (1-a)^{-1}$, we have

$$J_1(\cdot) \rightsquigarrow \left((G_o' \Omega_o^{-1} G_o)^{-1} \Sigma_o (G_o' \Omega_o^{-1} G_o)^{-1}\right)^{1/2} W(\cdot),$$

By the same argument as in the proof of Theorem 3, $\|J_2\|_\infty \xrightarrow{\mathbb{P}} 0$. This establishes $I(\cdot) \rightsquigarrow ((G_o' \Omega_o^{-1} G_o)^{-1} \Sigma_o (G_o' \Omega_o^{-1} G_o)^{-1})^{1/2} W(\cdot)$. Since $W(\cdot)$ is independent of $(z_i)_{i \geq 1}$ and $\hat{\theta}_{n,W} \in \sigma(z_{1:n})$, the proof is complete.

E.4 Proof of Theorem 5

Part (i): We derive the limiting distribution of the plug-in and debiased J test statistics. Let $\hat{W}$ denote a generic consistent estimator for Ω_o^{-1}. We begin by writing the J statistic as

$$J = n \bar{g}_n(\bar{\theta}_T^*)' \hat{W} \bar{g}_n(\bar{\theta}_T^*) = (\sqrt{n} \Omega_o^{-1/2} \bar{g}_n(\bar{\theta}_T^*))' \Omega_o^{-1/2} \hat{W} \Omega_o^{-1/2} (\sqrt{n} \Omega_o^{-1/2} \bar{g}_n(\bar{\theta}_T^*)).$$

We shall derive the asymptotic distribution of $\sqrt{n} \Omega_o^{-1/2} \bar{g}_n(\bar{\theta}_T^*)$. By the second-order Taylor expansion of $\bar{g}_n(\cdot)$ around $\hat{\theta}_{n,W}$, we have

$$\begin{aligned} \sqrt{n} \Omega_o^{-1/2} \bar{g}_n(\bar{\theta}_T^*) = \sqrt{n} \Omega_o^{-1/2} \bar{g}_n(\hat{\theta}_{n,W}) + \sqrt{\frac{n}{B_g(T-N)}} \Omega_o^{-1/2} \bar{G}_n(\hat{\theta}_{n,W}) \cdot \sqrt{B_g(T-N)} (\bar{\theta}_T^* - \hat{\theta}_{n,W}) \\ + O_{\mathbb{P}}(\sqrt{n}(T-N)^{-1}). \end{aligned}$$

Let $\Pi_o := \Omega_o^{-1/2} G_o (G_o' \Omega_o^{-1} G_o)^{-1} G_o' \Omega_o^{-1/2}$, an orthogonal projection matrix of rank d. Since $\hat{\theta}_{n,W}$ is an efficient estimator, it is well-known that the first term tends to $\mathcal{N}(0, I_{d_g} - \Pi_o)$ as in the standard J test. The second term converges in distribution to $\mathcal{N}(0, \tau \Pi_o)$ by the CLT established in Theorem 4. Since the second term is asymptoti-

cally independent of $(z_i)_{i\geq 1}$ by Lemma 7, the first two terms converge in distribution to

$$\mathcal{N}(0, I_{d_g} - (1-\tau)\Pi_o).$$

The last terms tends to zero as $\sqrt{n}/(T-N) \to 0$. Putting it all together, we conclude

$$J \xrightarrow{d} \|\mathcal{N}(0, I_{d_g} - (1-\tau)\Pi_o)\|^2 \sim \chi^2_{d_g - d} + \tau\chi^2_d.$$

The debiased J test statistic takes the form

$$\begin{aligned} J^D = (\sqrt{n}\Omega_o^{-1/2}\bar{g}_n(\bar{\theta}_T^*))'\Omega_o^{-1/2}\hat{W}^{1/2}(I_{d_g} - \hat{W}^{1/2}\bar{\Phi}_n(\bar{\Phi}_n'\hat{W}\bar{\Phi}_n)^{-1}\bar{\Phi}_n'\hat{W}^{1/2}) \\ \times \hat{W}^{1/2}\Omega_o^{-1/2}(\sqrt{n}\Omega_o^{-1/2}\bar{g}_n(\bar{\theta}_T^*)), \end{aligned}$$

where $\bar{\Phi}_n$ denotes a consistent estimator for G_o. From the result above, it follows that

$$J^D \xrightarrow{d} \mathcal{N}(0, I_{d_g} - (1-\tau)\Pi_o)'(I_{d_g} - \Pi_o)\mathcal{N}(0, I_{d_g} - (1-\tau)\Pi_o) \sim \chi^2_{d_g - d}.$$

Part (ii): Observe that $\bar{g}_T^* = \frac{1}{T-N}\sum_{t=N+1}^T \tilde{g}_t(\theta_{t-1}^*)$. We decompose $\bar{g}_T^*$ as follows:

$$\begin{aligned} \sqrt{n}\Omega_o^{-1/2}\bar{g}_T^* &= \frac{\sqrt{n}}{\sqrt{B_g(T-N)}} \cdot \frac{\sqrt{B_g}}{\sqrt{T-N}}\Omega_o^{-1/2}\sum_{t=N+1}^T [\tilde{g}_t(\theta_{t-1}^*) - \bar{g}_n(\theta_{t-1}^*)] \\ &\quad + \frac{\sqrt{n}}{\sqrt{B_g(T-N)}} \cdot \frac{\sqrt{B_g}}{\sqrt{T-N}}\Omega_o^{-1/2}\sum_{t=N+1}^T [\bar{g}_n(\theta_{t-1}^*) - \bar{g}_n(\hat{\theta}_{n,W})] \\ &\quad + \sqrt{n}\Omega_o^{-1/2}\bar{g}_n(\hat{\theta}_{n,W}) \\ &=: \frac{\sqrt{n}}{\sqrt{B_g(T-N)}} I_1 + \frac{\sqrt{n}}{\sqrt{B_g(T-N)}} I_2 + I_3. \end{aligned}$$

It is well-known that $I_3 \xrightarrow{d} \mathcal{N}(0, I_{d_g} - \Pi_o)$. Next, by the first-order Taylor expansion of $\theta \mapsto \bar{g}_n(\theta)$ around $\bar{\theta}_{n,W}$, we observe

$$\bar{g}_n(\theta_{t-1}^*) - \bar{g}_n(\bar{\theta}_{n,W}) = \bar{G}_n(\bar{\theta}_{n,W})(\theta_{t-1}^* - \bar{\theta}_{n,W}) + \mathrm{err}_{n,t-1},$$

where, on the event E_n, $\|\mathrm{err}_{n,t-1}\| \leq C\|\theta^*_{t-1} - \bar{\theta}_{n,W}\|^2$. This allows us to write

$$\left\| I_2 - \Omega_o^{-1/2}\bar{G}_n \underbrace{\frac{\sqrt{B_g}}{\sqrt{T-N}} \sum_{t=N+1}^{T} (\theta^*_{t-1} - \bar{\theta}_{n,W})}_{=I_4} \right\| \leq \frac{C}{\sqrt{T-N}} \sum_{t=N+1}^{T} \|\theta^*_{t-1} - \bar{\theta}_{n,W}\|^2$$

conditional on the event E_n. By the same argument as in the proof of Lemma 5, it holds

$$\frac{1}{\sqrt{T-N}} \sum_{t=N+1}^{T} \|\theta^*_{t-1} - \bar{\theta}_{n,W}\|^2 = o_{\mathbb{P}}(1).$$

Turning to the approximation of I_4, we note that, by Lemma 5,

$$I_4 = \frac{\sqrt{B_g}}{\sqrt{T-N}} \sum_{t=N+1}^{T} (\theta^\circ_{t-1} - \bar{\theta}_{n,W}) + o_{\mathbb{P}}(1).$$

Using the approximation from the proof of Lemma 6, we obtain

$$I_4 = -\frac{\sqrt{B_g}}{\sqrt{T-N}} \bar{\mathcal{H}}_n^{-1} \sum_{t=N+1}^{T} \xi_t(\theta^*_{t-1}) + o_{\mathbb{P}}(1).$$

Let $\zeta_t = (\tilde{g}_t(\theta^*_{t-1}) - \bar{g}_n(\theta^*_{t-1}), \xi_t(\theta^*_{t-1}))$, $t \geq 1$, so that $(\zeta_t)_{t\geq 1}$ forms a martingale difference sequence. Analogously to the proof of Lemma 6, one can verify the conditional Lindeberg conditions

$$\frac{1}{T-N} \sum_{t=N+1}^{T} \mathbb{E}^*_n \left[\|\zeta_t\|^2 \mathbb{1}\left\{ \|\zeta_t\| > \sqrt{T-N}\delta \right\} \middle| \mathcal{F}^*_{n,t-1} \right] \xrightarrow{\mathbb{P}} 0, \quad \forall \delta > 0,$$

$$\frac{1}{T-N} \sum_{t=N+1}^{N+\lfloor (T-N)r \rfloor} \mathbb{E}^*_n \left[\zeta_t \zeta_t' | \mathcal{F}^*_{n,t-1} \right] \xrightarrow{\mathbb{P}} r\Sigma_\zeta, \quad \forall r \in (0,1),$$

where

$$\Sigma_\zeta = B_g^{-1} \begin{bmatrix} \Omega_o & G_o \\ G_o' & G_o'\Omega_o^{-1}G_o \end{bmatrix},$$

whose details are omitted for brevity of the proof. This implies that, by martingale

CLT (Hall and Heyde (1980)),

$$(I_1, I_4) \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} I_{d_g} & -\Omega_o^{-1/2} G_o (G_o' \Omega_o^{-1} G_o)^{-1} \\ -(G_o' \Omega_o^{-1} G_o)^{-1} G_o' \Omega_o^{-1/2} & (G_o' \Omega_o^{-1} G_o)^{-1} \end{bmatrix}\right).$$

Putting these together,

$$I_1 + I_2 = I_1 + \Omega_o^{-1/2} \bar{G}_n I_4 + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}\left(0, I_{d_g} - \Pi_o\right).$$

Since the limiting random variable is asymptotically independent of $(z_i)_{i \geq 1}$, we obtain

$$\begin{aligned} \sqrt{n}\Omega_o^{-1/2} \bar{g}_T^* = \frac{\sqrt{n}}{\sqrt{B_g(T-N)}}(I_1 + I_2) + I_3 &\xrightarrow{d} \mathcal{N}(0, \tau(I_{d_g} - \Pi_o)) + \mathcal{N}(0, I_{d_g} - \Pi_o) \\ &\sim \mathcal{N}(0, (1+\tau)(I_{d_g} - \Pi_o)). \end{aligned}$$

This implies $(\sqrt{n}\Omega_o^{-1/2} \bar{g}_T^*)' \Omega_o^{1/2} \hat{W} \Omega_o^{1/2} (\sqrt{n}\Omega_o^{-1/2} \bar{g}_T^*) \xrightarrow{d} (1+\tau)\chi^2_{d_g - d}$.

E.5 Proof of Theorem 6

We first establish consistency. Note that the event E_n holds under Assumptions 6 and 7. The assertion that

$$\mathbb{P}_n^*\left(\theta_N^* \to \hat{\theta}_n \quad \text{as} \quad N \to \infty\right) = 1$$

then follows from Lemma 3(i). As a result, the averaged estimator $\bar{\theta}_N^* = \frac{1}{N}\sum_{t=1}^N \theta_t^*$ also converges to the same limit $\mathbb{P}^*$-almost surely. We now establish the FCLT. We write

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_t^* - \hat{\theta}_n) = \frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_t^* - \theta_t^\circ) + \frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_t^\circ - \hat{\theta}_n).$$

For the first term, Lemma 11 implies that

$$\sup_{r \in [0,1]} \left\| \frac{1}{\sqrt{N}} \sum_{t=1}^{\lfloor Nr \rfloor} (\theta_t^* - \theta_t^\circ) \right\| = \frac{1}{\sqrt{N}} \sup_{1 \leq t \leq N} \left\| \sum_{s=1}^{t} (\theta_s^* - \theta_s^\circ) \right\| \xrightarrow{\mathbb{P}^*} 0.$$

For the second term, Lemma 12 establishes the FCLT. Putting these together completes the proof.

F Proof of Lemmas

F.1 Proof of Lemma 1

By Assumption (A8.4), w.p.a.1, it holds

$$\hat{\theta}_{n,W} \equiv \arg\min_{\theta\in\theta} \bar{g}_n(\theta)' W_n \bar{g}_n(\theta) = \arg\min_{\theta:\|\theta\|\le\delta} \bar{g}_n(\theta)' W_n \bar{g}_n(\theta)$$

since $\bar{g}_n(\theta)' W_n \bar{g}_n(\theta) \ge c$ outside $\{\theta \in \Theta : \|\theta - \theta_o\| \le \delta\}$ and

$$\bar{g}_n(\hat{\theta}_{n,W}) W_n \bar{g}_n(\hat{\theta}_{n,W}) \le \bar{g}_n(\theta_o) W_n \bar{g}_n(\theta_o) \le C\|\bar{g}_n(\theta_o)\|^2 \xrightarrow{\mathbb{P}} 0,$$

where we use (D.2) to bound $W_n \le C I_{d_g}$. The standard ULLN (uniform law of large numbers) applies under Assumption (A8.4), implying that $\sup_{\|\theta\|\le\delta} \|\bar{g}_n(\theta) - g(\theta)\| \xrightarrow{\text{a.s.}} 0$ and hence

$$\begin{aligned}
&\sup_{\|\theta\|\le\delta} |\bar{g}_n(\theta)' W_n \bar{g}_n(\theta) - g(\theta)' W g(\theta)| \\
&\le C \sup_{\|\theta\|\le\delta} \|\bar{g}_n(\theta) - g(\theta)\|^2 + \sup_{\|\theta\|\le\delta} \|g(\theta)\|^2 \|W_n - W\| \xrightarrow{\mathbb{P}} 0.
\end{aligned}$$

The strong consistency of $\hat{\theta}_n$ then follows from Theorem 2.1 in Newey and McFadden (1994), and asymptotic normality also follows from their Theorem 3.2.

F.2 Proof of Lemma 2

Proof of (D.1)*:* This is an immediate consequence of $P_n \xrightarrow{\mathbb{P}} P > 0$ in Assumption (A8.2).

Proof of (D.2)*:* This is an immediate consequence of $W_n \xrightarrow{\mathbb{P}} W > 0$ in Assumption (A8.2).

Proof of (D.3)*:* It suffices to establish that $\bar{G}_n' W_n \bar{G}_n \xrightarrow{\mathbb{P}} G_o' W G_o > 0$. Let $B := \{\theta \in \Theta : \|\theta - \theta_o\| \le 1\}$ be a closed unit ball around θ_o. By Lemma 1, it follows $\hat{\theta}_{n,W} \in B$ w.p.a.1. By the ULLN, we have $\sup_{\theta\in B} \|\bar{G}_n(\theta) - G(\theta)\| \xrightarrow{\text{a.s.}} 0$. Since $W_n \le C I_{d_g}$ and $\|\bar{G}_n\| \le C$ w.p.a.1 by (D.2) and (D.5), which can be verified under Assumption 8, it

follows that

$$\begin{aligned}
&\|\bar{G}_n' W \bar{G}_n - G_o' W G_o\| \\
&\leq C\|\bar{G}_n - G_o\|^2 + C\|W_n - W\| \\
&\leq C \sup_{\theta \in B} \|\bar{G}_n(\theta) - G(\theta)\|^2 \mathbb{1}\left\{\hat{\theta}_{n,W} \in B\right\} + C\mathbb{1}\left\{\hat{\theta}_{n,W} \notin B\right\} + C\|W_n - W\| \xrightarrow{\mathbb{P}} 0.
\end{aligned}$$

Proof of (D.4)*:* It suffices to establish that $\bar{\mathcal{H}}_n \xrightarrow{\mathbb{P}} G_o' W G_o$. In light of the proof of (D.3), this should follow from

$$\bar{\mathcal{H}}_n - \bar{G}_n' W_n \bar{G}_n = \sum_{j=1}^{d_g} \bar{g}_{nj}(\hat{\theta}_{n,W}) W_n \frac{\partial^2 \bar{g}_{nj}(\hat{\theta}_{n,W})}{\partial\theta\partial\theta'} \xrightarrow{\mathbb{P}} 0.$$

It is straightforward to verify this since $\|\frac{\partial^2 \bar{g}_{nj}(\hat{\theta}_{n,W})}{\partial\theta\partial\theta'}\| \leq C$ for $j = 1, \ldots, d_g$, by (D.6), $W_n \leq C$ by (D.2), and

$$\|\bar{g}_{nj}(\hat{\theta}_{n,W})\|^2 \leq \underline{c}^{-1} \bar{g}_n(\hat{\theta}_{n,W})' W_n \bar{g}_n(\hat{\theta}_{n,W}) \leq \underline{c}^{-1} \bar{g}_n(\theta_o)' W_n \bar{g}_n(\theta_o) \leq \frac{C}{\underline{c}} \|\bar{g}_n(\theta_o)\|^2 \xrightarrow{\mathbb{P}} 0,$$

for each $j = 1, \ldots, d_g$, where we use (D.2).

Proof of (D.5)*:* By assumption (A8.4.4) and the SLLN (strong law of large numbers),

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} \|\bar{G}_n(\theta)\| \leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(z_i) = \mathbb{E}[H(z_i)] < \infty \quad \text{a.s.}$$

Proof of (D.6)*:* This follows from the fact that

$$\sup_{\theta\neq\tilde{\theta}} \frac{\|\bar{G}_n(\theta) - \bar{G}_n(\tilde{\theta})\|}{\|\theta - \tilde{\theta}\|} \leq \frac{1}{n} \sum_{i=1}^{n} L(z_i) \xrightarrow{\text{a.s.}} \mathbb{E}[L(z_i)] < \infty$$

where we use the SLLN. This also implies

$$\left\| \frac{\partial^2 \bar{g}_{nj}(\hat{\theta}_{n,W})}{\partial\theta\partial\theta'} \right\| \leq C \quad \text{w.p.a.1} \quad \text{for } j = 1, \ldots, d_g.$$

Proof of (D.7)*:* We begin by writing

$$\bar{Q}_{n,W}(\theta) = \frac{1}{2}\left(\bar{g}_n(\theta)'W_n\bar{g}_n(\theta) - \bar{g}_n(\hat{\theta}_{n,W})'W_n\bar{g}_n(\hat{\theta}_{n,W})\right).$$

We divide into two cases. If θ satisfies $\|\theta - \hat{\theta}_{n,W}\| \leq \delta$, then by Taylor's theorem, there exists $s \in (0,1)$ such that, for $\theta_s = s\theta + (1-s)\hat{\theta}_{n,W}$,

$$\bar{Q}_{n,W}(\theta) = \frac{1}{2}(\theta - \hat{\theta}_{n,W})'\frac{\partial^2 \bar{Q}_{n,W}(\theta_s)}{\partial\theta\partial\theta'}(\theta - \hat{\theta}_{n,W}). \tag{36}$$

By (A8.4), it follows

$$\bar{Q}_{n,W}(\theta) \geq c/2\|\theta - \hat{\theta}_{n,W}\|^2 \geq c/2\min\{\|\theta - \hat{\theta}_{n,W}\|^2, \delta^2\}$$

w.p.a.1.

Otherwise, if $\|\theta - \hat{\theta}_{n,W}\| \geq \delta$, we have $\bar{Q}_{n,W}(\theta) \geq c$, thereby implying that

$$\bar{Q}_{n,W}(\theta) \geq c \geq \underline{c}\min\{\|\theta - \hat{\theta}_{n,W}\|^2, \delta^2\}$$

w.p.a.1, where $\underline{c} = \min\{c/2, c/\delta^2\}$. This completes the proof of (D.7).

Proof of (D.8)*:* We divide into two cases. Suppose first that θ falls into the region $\mathcal{R} := \{\theta \in \Theta : \|\theta - \hat{\theta}_{n,W}\| \leq \varepsilon\}$, where $0 < \varepsilon < \delta$ is a small number to be determined later. By (D.5) and (D.6), we find that

$$\|\bar{G}_n(\theta) - \bar{G}_n\| \leq C\|\theta - \hat{\theta}_{n,W}\| \leq C\varepsilon \tag{37}$$

and

$$\|\bar{G}_n(\theta)'\bar{G}_n(\theta) - \bar{G}_n'\bar{G}_n\| \leq C\|\bar{G}_n(\theta) - \bar{G}_n\| \leq C\varepsilon$$

w.p.a.1 for all $\theta \in \mathcal{R}$. For each $\theta \in \mathcal{R}$, we can write

$$\begin{aligned}\bar{g}_n(\theta) - \bar{g}_n(\hat{\theta}_{n,W}) &= \int_0^1 \bar{G}_n(\theta_s)ds \cdot (\theta - \hat{\theta}_{n,W}) \\ &= \bar{G}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) + \int_0^1 (\bar{G}_n(\theta_s) - \bar{G}_n(\hat{\theta}_n))ds \cdot (\theta - \hat{\theta}_{n,W})\end{aligned}$$

where the line segment $\theta_s = (1-s)\hat{\theta}_{n,W} + s\theta$, $s \in [0,1]$, is contained in $\mathcal{R}$. Thus, we

can express $\bar{g}_n(\theta)$ as

$$\bar{g}_n(\theta) = \bar{G}_n(\theta - \hat{\theta}_{n,W}) + R_n(\theta)(\theta - \hat{\theta}_{n,W}) + \bar{g}_n(\hat{\theta}_{n,W}) \tag{38}$$

where $\|R_n(\theta)\| \leq C\epsilon$ for all $\theta \in \mathcal{R}$ w.p.a.1. Equation (37) implies that

$$\left|\bar{g}_n(\theta)'W_n(\bar{G}_n(\theta)'\bar{G}_n(\theta) - \bar{G}_n'\bar{G}_n)W_n\bar{g}_n(\theta)\right| \leq C\varepsilon\|W_n\bar{g}_n(\theta)\|^2 \leq C\varepsilon\bar{g}_n(\theta)'W_n\bar{g}_n(\theta)$$

w.p.a.1 by (D.2). Putting this together with (38) yields, for all $\theta \in \mathcal{R}$,

$$\begin{aligned}
&\|\bar{G}_n(\theta)'W_n\bar{g}_n(\theta)\|^2 \\
&= \bar{g}_n(\theta)'W_n\bar{G}_n(\theta)\bar{G}_n(\theta)'W_n\bar{g}_n(\theta) \\
&\geq \bar{g}_n(\theta)'W_n\bar{G}_n\bar{G}_n'W_n\bar{g}_n(\theta) - C\varepsilon\bar{g}_n(\theta)W_n'\bar{g}_n(\theta) \\
&= (\theta - \hat{\theta}_{n,W})' \begin{pmatrix} (\bar{G}_n'W_n\bar{G}_n)^2 + R_n(\theta)W_n'\bar{G}_n\bar{G}_n'W_n\bar{G}_n \\ + \bar{G}_n'W_n\bar{G}_n\bar{G}_n'W_nR_n(\theta) + R_n(\theta)'W_n\bar{G}_n\bar{G}_n'W_nR_n(\theta) \end{pmatrix} (\theta - \hat{\theta}_{n,W}) - C\varepsilon\bar{Q}_{n,W}(\theta) \\
&\geq \underline{c}\|\theta - \hat{\theta}_{n,W}\|^2 - C\varepsilon\bar{Q}_{n,W}(\theta) \geq \underline{c}\bar{Q}_{n,W}(\theta),
\end{aligned}$$

w.p.a.1, where $\underline{c} > 0$ is an absolute constant that may differ from line to line and is positive provided that $\varepsilon > 0$ is chosen sufficiently small. In the last line, we invoke (36) and Assumption (A8.4.8) to deduce that

$$\bar{Q}_{n,W}(\theta) \leq C\|\theta - \hat{\theta}_{n,W}\|^2, \quad \forall\theta \in \Theta.$$

This concludes that

$$\left\|\frac{\partial\bar{Q}_{n,W}(\theta)}{\partial\theta}\right\|^2 = \|\bar{G}_n(\theta)'W_n\bar{g}_n(\theta)\|^2 \geq \underline{c}(\bar{Q}_{n,W}(\theta) \wedge 1), \quad \theta \in \mathcal{R},$$

which addresses the first case.

To address the remaining case $\theta \in \mathcal{R}^c$, note that by construction $\mathcal{R}^c \subseteq \{\theta : \|\theta - \hat{\theta}_{n,W}\| \geq \delta\}$. Now, by Assumption (A8.4.7), we obtain, for all $\theta \in \mathcal{R}^c$,

$$\left\|\frac{\partial\bar{Q}_{n,W}(\theta)}{\partial\theta}\right\|^2 \geq c^2 \geq \underline{c}(\bar{Q}_{n,W}(\theta) \wedge 1)$$

by choosing $\underline{c} > 0$ to be no greater than c^2. This completes the proof of (D.8).

Proof of (D.9)*:* Observe that

$$\xi_t(\theta) - \xi_t(\hat{\theta}_{n,W}) = \tilde{G}_t(\theta)'(\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})) + (\tilde{G}_t(\theta) - \tilde{G}_t(\hat{\theta}_{n,W}))'\tilde{g}_t(\hat{\theta}_{n,W}).$$

Using the independence between $\tilde{G}_t(\theta)$ and $\tilde{g}_t(\theta)$ and Cauchy-Schwarz inequality, we find that

$$\begin{aligned}&\mathbb{E}_n^\star[\|\xi_t(\theta) - \xi_t(\hat{\theta}_{n,W})\|^2] \\ &\leq 2\left(\mathbb{E}_n^\star[\|\tilde{G}_t(\theta)\|^2]\mathbb{E}_n^\star[\|\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})\|^2] + \mathbb{E}_n^\star[\|\tilde{G}_t(\theta) - \tilde{G}_t(\hat{\theta}_{n,W})\|^2]\mathbb{E}_n^\star[\|\tilde{g}_t(\hat{\theta}_{n,W})\|^2]\right).\end{aligned}$$

By the triangle inequality and Assumption (A8.4.8), we have, for all $t \geq 1$ and $n \in \mathbb{N}$,

$$\sup_{\theta\in\Theta} \mathbb{E}_n^\star[\|\tilde{G}_t(\theta)\|^2] \leq \sup_{\theta\in\Theta} \mathbb{E}_n^\star[\|G(\tilde{z}_1, \theta)\|^2] = \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \|G(z_i, \theta)\|^2 \leq \frac{1}{n}\sum_{i=1}^n H(z_i)^2$$

and

$$\begin{aligned}\mathbb{E}_n^\star[\|\tilde{g}_t(\hat{\theta}_{n,W})\|^2] \leq \mathbb{E}_n^\star[\|g(\tilde{z}_1, \hat{\theta}_{n,W})\|^2] &= \frac{1}{n}\sum_{i=1}^n \|g(z_i, \hat{\theta}_{n,W})\|^2 \\ &\leq C(\bar{Q}_{n,W}(\hat{\theta}_{n,W}) + 1) \leq C(\|\bar{g}_n(\theta_o)\|^2 + 1),\end{aligned}$$

both of which are bounded by C w.p.a.1 by the SLLN. Thus, it suffices to establish that $\mathbb{E}_n^\star[\|\tilde{G}_t(\theta) - \tilde{G}_t(\hat{\theta}_{n,W})\|^2]$ and $\mathbb{E}_n^\star[\|\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})\|^2]$ are each bounded by a multiple of $\|\theta - \hat{\theta}_{n,W}\|^2$. Note that

$$\|\tilde{G}_t(\theta) - \tilde{G}_t(\hat{\theta}_{n,W})\| \leq \left(\frac{1}{B_{G,t}}\sum_{j=1}^{B_{G,t}} L(\tilde{z}_{j+B_{t-1}})\right)\|\theta - \hat{\theta}_{n,W}\|$$

by Assumption (A8.4.5), thereby implying that, for all $\theta \in \Theta$ and $t \geq 1$,

$$\begin{aligned}&\mathbb{E}_n^\star[\|\tilde{G}_t(\theta) - \tilde{G}_t(\hat{\theta}_{n,W})\|^2] \\ &\leq \|\theta - \hat{\theta}_{n,W}\|^2\left[\left(\frac{1}{n}\sum_{i=1}^n L(z_i)\right)^2 + \frac{1}{nB_{G,t}}\sum_{i=1}^n\left(L(z_i) - \frac{1}{n}\sum_{i=1}^n L(z_i)\right)^2\right].\end{aligned}$$

Since the moments on the right-hand side converge to their limits by the SLLN, we

have

$$\mathbb{E}_n^\star[\|\tilde{G}_t(\theta) - \tilde{G}_t(\hat{\theta}_{n,W})\|^2] \leq C\|\theta - \hat{\theta}_{n,W}\|^2$$

for all $\theta \in \Theta$ and $t \geq 1$ w.p.a.1.

Finally, $\mathbb{E}_n^\star[\|\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})\|^2]$ can be bounded as follows. Let $v = \tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})$. Applying the mean-value theorem to a function $\theta \mapsto v'(\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W}))$, we can find some $s \in (0,1)$ such that

$$\|\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})\|^2 = v'(\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})) = v'\tilde{G}_t(\theta_s)(\theta - \hat{\theta}_{n,W}) \leq \|v\|\|\tilde{G}_t(\theta_s)\|\|\theta - \hat{\theta}_{n,W}\|,$$

where $\theta_s = (1-s)\theta + s\hat{\theta}_{n,W}$ lies in the line segment joining θ and $\hat{\theta}_{n,W}$. Since $\|v\| = \|\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})\|$, dividing both sides by $\|v\|$ leads to the upper bound

$$\|\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})\| \leq \|\tilde{G}_t(\theta_s)\|\|\theta - \hat{\theta}_{n,W}\| \leq \frac{1}{B_{G,t}} \sum_{j=1}^{B_{G,t}} H(\tilde{z}_{j+B_{t-1}})\|\theta - \hat{\theta}_{n,W}\|,$$

whence it follows

$$\begin{aligned}
&\mathbb{E}_n^\star[\|\tilde{g}_t(\theta) - \tilde{g}_t(\hat{\theta}_{n,W})\|^2] \\
&\leq \left[\left(\frac{1}{n}\sum_{i=1}^n H(z_i)\right)^2 + \frac{1}{nB_{G,t}}\sum_{i=1}^n \left(H(z_i) - \frac{1}{n}\sum_{i=1}^n H(z_i)\right)^2\right]\|\theta - \hat{\theta}_{n,W}\|^2 \\
&\leq C\|\theta - \hat{\theta}_{n,W}\|^2
\end{aligned}$$

uniformly in $t \geq 1$ w.p.a.1.

Proof of (D.10)*:* There exists a constant $C_p > 0$ such that

$$\mathbb{E}_n^\star[\|\xi_t(\theta)\|^{2p}] \leq C_p\left(\mathbb{E}_n^\star[\|\tilde{G}_t(\theta)\|^{2p}]\mathbb{E}_n^\star[\|\tilde{g}_t(\theta)\|^{2p}] + \|\bar{G}_n(\theta)\|^{2p}\|\bar{g}_n(\theta)\|^{2p}\right).$$

By the triangle inequality and Assumption (A4.8), we have, for all $t \geq 1$, $n \in \mathbb{N}$, and $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{E}_n^\star[\|\tilde{G}_t(\theta)\|^{2p}] &\leq \mathbb{E}_n^\star[\|G(\tilde{z}_1,\theta)\|^{2p}] = \frac{1}{n}\sum_{i=1}^n \|G(z_i,\theta)\|^{2p} \leq \frac{1}{n}\sum_{i=1}^n H(z_i)^{2p}, \\
\mathbb{E}_n^\star[\|\tilde{g}_t(\theta)\|^{2p}] &\leq \mathbb{E}_n^\star[\|g(\tilde{z}_1,\theta)\|^{2p}] = \frac{1}{n}\sum_{i=1}^n \|g(z_i,\theta)\|^{2p} \leq M(\bar{Q}_{n,W}(\theta)^p + 1).
\end{aligned}$$

The rightmost quantities converge to finite values, and hence are bounded above by C w.p.a.1. Since $\sup_{\theta\in\Theta}\|\bar{G}_n(\theta)\| \leq C$ by (D.5) and $\|\bar{g}_n(\theta)\|^{2p} \leq [C\bar{g}_n(\theta)'W_n\bar{g}_n(\theta)]^p \leq C\bar{Q}_{n,W}(\theta)^p$, for all $\theta \in \Theta$ w.p.a.1, we conclude that, for all $\theta \in \Theta$,

$$\sup_{t\geq 1}\mathbb{E}_n^\star[\|\xi_t(\theta)\|^{2p}] \leq C(\bar{Q}_{n,W}(\theta)^p + 1)$$

F.3 Proof of Lemma 3

Part (i): Let us write

$$Q_t^\star = \bar{Q}_{n,W}(\theta_t^\star).$$

Conditional on the event E_n, by the second-order Taylor expansion, we have

$$Q_{t+1}^\star - Q_t^\star = \nabla\bar{Q}_{n,W}(\theta_t^\star)^\top(\theta_{t+1}^\star - \theta_t^\star) + \frac{1}{2}(\theta_{t+1}^\star - \theta_t^\star)^\top\frac{\partial^2\bar{Q}_{n,W}(\tilde{\theta}_t^\star)}{\partial\theta\partial\theta'}(\theta_{t+1}^\star - \theta_t^\star),$$

where $\tilde{\theta}_t^\star$ lies on the line segment $[\theta_t^\star, \theta_{t+1}^\star]$. Using the fact that the Hessian of $\bar{Q}_{n,W}$ is uniformly bounded above by MI_d and

$$\theta_{t+1}^\star - \theta_t^\star = -\gamma_{t+1}^\star P_n\nabla\bar{Q}_{n,W}(\theta_t^\star) - \gamma_{t+1}^\star P_n\xi_{t+1}(\theta_t^\star),$$

we find that

$$\begin{aligned}Q_{t+1}^\star - Q_t^\star \leq &-\underline{c}\gamma_{t+1}^\star\|\nabla\bar{Q}_{n,W}(\theta_t^\star)\|^2 - \gamma_{t+1}^\star\nabla\bar{Q}_{n,W}(\theta_t^\star)^\top P_n\xi_{t+1}(\theta_t^\star)\\ &+ C{\gamma_{t+1}^\star}^2(\|\nabla\bar{Q}_{n,W}(\theta_t^\star)\|^2 + \|\xi_{t+1}(\theta_t^\star)\|^2).\end{aligned}$$

where we use (D.1). By Lemma 2, conditional on E_n, $\|\nabla\bar{Q}_{n,W}(\theta_t^\star)\|^2 \geq \underline{c}(Q_t^\star \wedge 1)$ and

$$\mathbb{E}_n^\star\left[\|\nabla\bar{Q}_{n,W}(\theta_t^\star)\|^2 + \|\xi_{t+1}(\theta_t^\star)\|^2\right] \leq C(Q_t^\star + 1),$$

where we use $\mathbb{E}_n^\star\left[\|\xi_{t+1}(\theta_t^\star)\|^2\right] \leq \mathbb{E}_n^\star\left[\|\xi_{t+1}(\theta_t^\star)\|^{2p}\right]^{1/p} \leq C(Q_t^\star + 1)$. This leads to the basic inequality

$$\mathbb{E}_n^\star[Q_{t+1}^\star|\mathcal{F}_{n,t}^\star] \leq Q_t^\star - \underline{c}\gamma_{t+1}^\star(Q_t^\star \wedge 1) + C{\gamma_{t+1}^\star}^2(Q_t^\star + 1), \quad t \geq 0, \tag{39}$$

which holds for all $n \in \mathbb{N}$ and $N^\star \geq 0$ on the event E_n.

Let us define

$$\pi_t = \prod_{s=1}^{t}(1 + C\gamma_s^\star), \quad t \geq 0,$$

and let $\pi_\infty = \lim_{t\to\infty} \pi_t < \infty$ denote its limit, which is uniformly bounded in $N^\star$ since $\sum_{s\geq 1} \gamma_s^{\star 2} \leq \gamma_0^2 \sum_{s\geq 1} s^{-2a}$ is bounded by an absolute constant. Let $\tilde{Q}_t^\star := Q_t/\pi_t$. Dividing each side of (39) by π_{t+1}, we find that

$$\mathbb{E}_n^\star[\tilde{Q}_{t+1}^\star | \mathcal{F}_{n,t}^\star] \leq \tilde{Q}_t^\star - \underline{c}\gamma_{t+1}^\star(\tilde{Q}_t^\star \wedge 1) + C{\gamma_{t+1}^\star}^2, \quad t \geq 0. \tag{40}$$

(Recall that $\underline{c} > 0$ and $C > 0$ can be different in each equation.) By Lemma 8, (40) implies that

$$\mathbb{P}_n^\star\left(\lim_{t\to\infty} \tilde{Q}_t^\star \text{ exists}, \quad \sum_{t=1}^{\infty} \gamma_{t+1}^\star(\tilde{Q}_t^\star \wedge 1) < \infty\right) = 1.$$

These together imply that $\lim_{t\to\infty} \tilde{Q}_t^\star = 0$, yielding $\theta_T^\star - \hat{\theta}_{n,W} \to 0$ $\mathbb{P}_n^\star$-a.s. as $T \to \infty$ conditional on the event E_n by (D.7).

Part (ii): Define $W_t^\star$ by

$$W_t^\star = \tilde{Q}_t^\star + \underline{c}\sum_{s=0}^{t-1} \gamma_{s+1}^\star(\tilde{Q}_s^\star \wedge 1) - C\sum_{s=0}^{t-1} {\gamma_{s+1}^\star}^2, \quad t \geq 0,$$

where $\underline{c} > 0$ and $C > 0$ are the same as in (40). Note that

$$-C\sum_{s=0}^{t-1} {\gamma_{s+1}^\star}^2 \leq W_t^\star \leq \tilde{Q}_t^\star + \underline{c}\sum_{s=0}^{t-1} \gamma_{s+1}^\star,$$

and hence, for all $t \geq 0$, $\mathbb{E}_n^\star[|W_t^\star|] < \infty$ on the event E_n. By (40), it follows that $\mathbb{E}_n^\star[W_{t+1}^\star | \mathcal{F}_t^\star] \leq W_t^\star$ on the event E_n, that is, $(W_t^\star)_{t\geq 0}$ forms a supermartingale.

Conditional on the event E_n, consider a submartingale $(X_t^\star, \mathcal{F}_{n,t}^\star, \mathbb{P}_n^\star)_{t\geq 0}$ defined by

$$X_t^\star := -W_t^\star + W_0^\star, \quad t \geq 0.$$

By the submartingale maximal inequality (see Chung (2000, Theorem 9.4.1)), we have, for all $n \in \mathbb{N}$, $N^\star \geq 0$, and $T \in \mathbb{N}$,

$$\mathbb{P}_n^\star\left(\min_{0\leq t\leq T} X_t^\star \leq -\lambda\right) \leq \lambda^{-1}\mathbb{E}_n^\star[(X_T^\star \vee 0)], \quad \forall \lambda > 0$$

on the event E_n. Note that

$$X_T^\star \le \tilde{Q}_0^\star + C\sum_{s=0}^{\infty} {\gamma_{s+1}^\star}^2 \le C_1$$

by Assumption (A8.1) with $C_1 > 0$ being an absolute constant. Hence, for all $n \in \mathbb{N}$, $N^\star \ge 0$, and $T \in \mathbb{N}$, it holds $\mathbb{E}_n^\star[(X_T^\star \vee 0)] \le C_1$. This allows us to find an absolute constant $K < \infty$ such that

$$\mathbb{P}_n^\star\left(\min_{0\le t\le T} X_t^\star \le -(K - C_1)\right)\mathbb{1}_{E_n} < \varepsilon$$

for all $n \in \mathbb{N}$, $N^\star \ge 0$, and $T \in \mathbb{N}$. We then observe that

$$\max_{0\le t\le T}(-X_t^\star) \ge \max_{0\le t\le T}\tilde{Q}_t^\star - C\sum_{s=1}^{\infty}\gamma_s^{\star 2} - \tilde{Q}_0^\star \ge \max_{0\le t\le T}\tilde{Q}_t^\star - C_1.$$

This implies

$$\mathbb{P}\left(\max_{0\le t\le T}\tilde{Q}_t^\star \ge K\right) \le \mathbb{P}_n^\star\left(\max_{0\le t\le T}(-X_t^\star) \ge K - C_1\right) < \varepsilon,$$

which is equivalent to

$$\mathbb{P}_n^\star(T_K^\star \le T) < \varepsilon.$$

Since K is chosen to be independent of n, T, and $N^\star$, we let $T \to \infty$ to conclude that, for all $n \in \mathbb{N}$ and $N^\star \ge 0$,

$$\mathbb{P}_n^\star(T_K^\star < \infty)\mathbb{1}_{E_n} < \varepsilon.$$

This verifies part (ii).

Part (iii): Choose positive integers $\tilde{T} < T$ and consider a submartingale $(W_{\tilde{T}}^\star - W_t^\star, \mathcal{F}_{n,t}^\star, \mathbb{P}_n^\star)_{t=\tilde{T}}^T$. By part (ii), we can find K such that, for all $n \in \mathbb{N}$ and $N^\star \ge 0$, $\mathbb{P}_n^\star(T_K^\star < \infty)\mathbb{1}_{E_n} < \varepsilon$. By the submartingale maximal inequality (Chung (2000)), we have, for all $\varepsilon > 0$,

$$\mathbb{P}_n^\star\left(\min_{\tilde{T}\le t\le T}[W_{\tilde{T}}^\star - W_t^\star] < -\varepsilon \,\middle|\, \mathcal{F}_{n,\tilde{T}}\right)\mathbb{1}_{\mathcal{E}} \le \varepsilon^{-1}\mathbb{E}_n^\star[(W_{\tilde{T}}^\star - W_T^\star \vee 0)|\mathcal{F}_{n,\tilde{T}}]\mathbb{1}_{\mathcal{E}},$$

where the event $\mathcal{E} := \{T_K^\star \ge \tilde{T} + 1\}$ is $\mathcal{F}_{n,\tilde{T}}$-measurable since $T_K^\star$ is a stopping time

with respect to $(\mathcal{F}^\star_{n,t})_{t\geq 0}$. As T tends to infinity, we have, on the event E_n,

$$\max\{W^\star_{\tilde{T}} - W^\star_T, 0\} \leq \tilde{Q}^\star_{\tilde{T}} + C\sum_{s=\tilde{T}}^{T-1} {\gamma^\star_{s+1}}^2 \to \tilde{Q}^\star_{\tilde{T}} + C\sum_{s=\tilde{T}}^{\infty} {\gamma^\star_{s+1}}^2 =: U_0$$

By the monotone convergence theorem, we have, as $T \to \infty$,

$$\mathbb{P}^\star_n\left(\inf_{\tilde{T}\leq t<\infty}[W^\star_{\tilde{T}} - W^\star_t] < -\varepsilon \middle| \mathcal{F}_{n,\tilde{T}}\right)\mathbb{1}_{\mathcal{E}} \leq \varepsilon^{-1}\mathbb{E}^\star_n[U_0|\mathcal{F}_{n,\tilde{T}}]\mathbb{1}_{\mathcal{E}}.$$

By Lemma 4, which does not rely on part (iii), for all $n \in \mathbb{N}$ and $N^\star \geq 0$, it holds

$$\mathbb{E}^\star_n\left[U_0\mathbb{1}\left\{T^\star_K \geq \tilde{T}+1\right\}\right] \leq \mathbb{E}^\star_n\left[U_0\mathbb{1}\left\{T^\star_K \geq \tilde{T}\right\}\right] \leq C_K\gamma^\star_{\tilde{T}} + C\sum_{s\geq\tilde{T}} {\gamma^\star_{s+1}}^2.$$

As a result, we can find $\tilde{T} < \infty$ independent of $n \in \mathbb{N}$ and $N^\star \geq 0$ such that for all $n \in \mathbb{N}$ and $N^\star \geq 0$,

$$\mathbb{E}^\star_n\left[U_0\mathbb{1}\left\{T^\star_K \geq \tilde{T}+1\right\}\right] \leq C_K\gamma^\star_{\tilde{T}} + C\sum_{s\geq\tilde{T}} {\gamma^\star_{s+1}}^2 < (\varepsilon^2 \wedge \varepsilon)$$

on the event E_n. This implies that

$$\mathbb{P}^\star_n\left(\inf_{\tilde{T}\leq t<\infty}[W^\star_{\tilde{T}} - W^\star_t] < -\varepsilon, T^\star_K \geq \tilde{T}+1\right) < \varepsilon,$$

and hence, for all $n \in \mathbb{N}$ and $N^\star \geq 0$,

$$\mathbb{P}^\star_n\left(\inf_{\tilde{T}\leq t<\infty}[W^\star_{\tilde{T}} - W^\star_t] < -\varepsilon\right) < \varepsilon + \mathbb{P}^\star_n(T^\star_K < \infty) < 2\varepsilon$$

on the event E_n. Now, $\inf_{\tilde{T}\leq t<\infty}[W^\star_{\tilde{T}} - W^\star_t] \geq -\varepsilon$ implies that for all $t \geq \tilde{T}$,

$$\tilde{Q}^\star_t \leq \tilde{Q}^\star_{\tilde{T}} - \underline{c}\sum_{s=\tilde{T}}^{t-1}\gamma^\star_{s+1}(\tilde{Q}^\star_s \wedge 1) + C\sum_{s=\tilde{T}}^{t-1}{\gamma^\star_{s+1}}^2 + \varepsilon \leq U_0 + \varepsilon.$$

Define an event $\mathcal{U} := \{U_0 \leq \varepsilon\}$. Since by Markov's inequality,

$$\mathbb{P}^\star_n(\mathcal{U}^c) \leq \varepsilon^{-1}\mathbb{E}^\star_n\left[U_0\mathbb{1}\left\{T^\star_K \geq \tilde{T}+1\right\}\right] + \mathbb{P}(T^\star_K < \infty) < 2\varepsilon,$$

we have

$$\begin{aligned}\mathbb{P}_n^\star\left(\inf_{\tilde{T}\le t<\infty}[W_{\tilde{T}}^\star - W_t^\star] < -\varepsilon\right) &\ge \mathbb{P}_n^\star\left(\left\{\inf_{\tilde{T}\le t<\infty}[W_{\tilde{T}}^\star - W_t^\star] < -\varepsilon\right\}\cap\mathcal{U}\right)\\ &\ge \mathbb{P}_n^\star\left(\left\{\sup_{\tilde{T}\le t<\infty}\tilde{Q}_t^\star \le 2\varepsilon\right\}^c\cap\mathcal{U}\right)\\ &\ge \mathbb{P}_n^\star\left(\sup_{\tilde{T}\le t<\infty}\tilde{Q}_t^\star > 2\varepsilon\right) - 2\varepsilon.\end{aligned}$$

This shows that, for all $n\in\mathbb{N}$ and $N^\star\ge 0$,

$$\mathbb{P}_n^\star\left(\sup_{\tilde{T}\le t<\infty}\tilde{Q}_t^\star > 2\varepsilon\right) < 4\varepsilon$$

on the event E_n. Since $Q_t^\star \le \pi_\infty\tilde{Q}_t^\star \le \exp\left(C\gamma_0^2\zeta(2a)\right)\tilde{Q}_t^\star$, this verifies part (iii), where $\zeta(2a)=\sum_{s\ge1}s^{-2a}$ denotes the Riemann zeta function.

F.4 Proof of Lemma 4

Define an auxiliary stopping time

$$\tilde{T}_K^\star := \inf\{t\ge 0 : \tilde{Q}_t^\star \ge K\} \tag{41}$$

for $K>0$. Since $Q_t^\star\le\pi_\infty\tilde{Q}_t^\star$, we observe that

$$T_K^\star \le \tilde{T}_{\pi_\infty K}^\star, \quad \forall K>0.$$

Let $\bar{x}$ be a positive number. Since $(x\wedge 1)\ge(1\wedge\bar{x}^{-1})x$ for all $x\in[0,\bar{x}]$, we find that

$$(\tilde{Q}_t^\star\wedge 1)\ge(K^{-1}\wedge 1)\tilde{Q}_t^\star \ge \underline{c}_K\tilde{Q}_t^\star$$

conditional on $\mathcal{E}_{t+1}:=\{\tilde{T}_K^\star\ge t+1\}$, which is an $\mathcal{F}_{n,t}^\star$-measurable event. Here, $\underline{c}_K>0$ denotes a generic absolute constant that may depend on K. Taking the expectations

of both sides of (40) on $\mathcal{E}_{t+1}$, this implies

$$\begin{aligned}\mathbb{E}_n^\star\left[\tilde{Q}_{t+1}^\star \mathbb{1}_{\mathcal{E}_{t+1}}\right] &\le \mathbb{E}_n^\star\left[\tilde{Q}_t^\star \mathbb{1}_{\mathcal{E}_{t+1}}\right](1-\underline{c}_K\gamma_{t+1}^\star) + C{\gamma_{t+1}^\star}^2 \\ &\le \mathbb{E}_n^\star\left[\tilde{Q}_t^\star \mathbb{1}_{\mathcal{E}_t}\right](1-\underline{c}_K\gamma_{t+1}^\star) + C{\gamma_{t+1}^\star}^2\end{aligned}$$

for all $t \ge 0$ and some absolute constants $0 < \underline{c}_K < 1/(2\gamma_0)$ and $C > 0$, where we use the fact that $\mathcal{E}_{t+1} \subseteq \mathcal{E}_t$. From this, we can also see that

$$\mathbb{E}_n^\star[\tilde{Q}_t^\star \mathbb{1}_{\mathcal{E}_t}] \le \mathbb{E}_n^\star[\tilde{Q}_{t-1}^\star \mathbb{1}_{\mathcal{E}_t}] + C{\gamma_{t+1}^\star}^2 \le K + C\gamma_0^2$$

is uniformly bounded in $t \ge 1$. Define a (random) sequence $v_t := \mathbb{E}_n^\star[\tilde{Q}_t^\star \mathbb{1}_{\mathcal{E}_t}]/\gamma_t^\star$, which must satisfy the following recursive bound:

$$v_{t+1} \le v_t(1-\underline{c}_K\gamma_{t+1}^\star)\frac{\gamma_t^\star}{\gamma_{t+1}^\star} + C\gamma_{t+1}^\star, \quad t \ge 0.$$

Using the fact that

$$\gamma_t^\star/\gamma_{t+1}^\star \le (1+1/(t+N^\star))^a \le 1 + \frac{a}{t+N^\star}$$

we have

$$(1-\underline{c}_K\gamma_{t+1}^\star)\frac{\gamma_t^\star}{\gamma_{t+1}^\star} \le 1 - \frac{\underline{c}_K\gamma_{t+1}^\star}{2}$$

provided that $a/(t+N^\star) \le \underline{c}_K\gamma_{t+1}^\star/2$. Since $\gamma_t^\star/\gamma_{t+1}^\star \le 2$, we deduce that $a/(t+N^\star) \le \underline{c}_K\gamma_{t+1}^\star/2$ is satisfied under

$$\frac{a}{t+N^\star} \le \underline{c}_K\gamma_t^\star/4$$

i.e.,

$$t \ge \mathcal{T}_K(N^\star) := \left(\left\lfloor\left(\frac{4a}{\underline{c}_K\gamma_0}\right)^{1/(1-a)}\right\rfloor - N^\star\right) \vee 0,$$

where we write $\mathcal{T} = \mathcal{T}_K(N^\star)$ for ease of notation.

We divide the analysis into two cases: (i) $\lfloor(\frac{4a}{\underline{c}_K\gamma_0})^{1/(1-a)}\rfloor \ge N^\star + 1$, and (ii) $\lfloor(\frac{4a}{\underline{c}_K\gamma_0})^{1/(1-a)}\rfloor \le N^\star$. In case (i), applying Lemma 9 to the sequence $(v_t)_{t\ge\mathcal{T}}$ satisfying

$$v_{t+1} \le v_t\left(1 - \frac{\underline{c}_K\,\gamma_{t+1}^\star}{2}\right) + C\,\gamma_{t+1}^\star,$$

we obtain

$$v_t \le C\left(v_{\mathcal{T}} + \frac{2}{\underline{c}_K}\right) \le C\left(K + C\gamma_0^2\right)\frac{\lfloor (\frac{4a}{\underline{c}_K\gamma_0})^{1/(1-a)}\rfloor^a}{\gamma_0} + \frac{2C}{\underline{c}_K}$$

for all $t \ge \mathcal{T}$ on the event E_n. Note that the upper bound on the right-hand side does not depend on $(n, N^\star, T)$ or the time index. For the remaining time indices $t \le \mathcal{T}$, we simply observe that

$$v_t = \frac{\mathbb{E}_n^\star\left[\tilde{Q}_t^\star \mathbb{1}_{\mathcal{E}_t}\right]}{\gamma_t^\star} \le \frac{K + C\gamma_0^2}{\gamma_{\mathcal{T}}^\star} \le \left(K + C\gamma_0^2\right)\frac{\lfloor (\frac{4a}{\underline{c}_K\gamma_0})^{1/(1-a)}\rfloor^a}{\gamma_0}$$

for $t \ge 1$. For $t = 0$, by assumption, $v_0 = \tilde{Q}_t^\star/\gamma_0^\star = Q_t^\star/\gamma_0^\star \le M$. Hence, there exists an upper bound that does not depend on $(n, N^\star, T)$ and applies to the entire sequence $(v_t)_{t\ge 1}$. This completes the analysis of case (i).

In case (ii), we have $\mathcal{T} = 0$, so that

$$v_{t+1} \le v_t\left(1 - \frac{\underline{c}_K\,\gamma_{t+1}^\star}{2}\right) + C\,\gamma_{t+1}^\star$$

holds for all $t \ge 0$. Lemma 9 then implies

$$v_t \le C\left(v_0 + \frac{2}{\underline{c}_K}\right) \le C\frac{\tilde{Q}_0^\star}{\gamma_0^\star} + \frac{2C}{\underline{c}_K}.$$

By the assumption that $Q_0^\star \le M\gamma_0^\star$ and the fact that $\tilde{Q}_0^\star = Q_0^\star$, we find that $v_t \le CM + \frac{2C}{\underline{c}_K}$, providing an absolute upper bound for v_t for all $t \ge 1$. This completes the analysis of case (ii).

Putting these together, we have shown that

$$\mathbb{E}_n^\star\left[\tilde{Q}_t^\star \mathbb{1}\left\{\tilde{T}_K^\star \ge t\right\}\right] \le C_K\gamma_t^\star, \quad t \ge 0,$$

where C_K is an absolute constant that depends on K. To translate this to the corresponding bound on $\mathbb{E}_n^\star\left[Q_t^\star \mathbb{1}\left\{T_K^\star \ge t\right\}\right]$, we recall that

$$Q_t^\star \le \pi_\infty \tilde{Q}_t^\star \le \exp\left(\gamma_0^2\zeta(2a)\right)\tilde{Q}_t^\star \quad \text{and} \quad T_K^\star \le \tilde{T}^\star_{\exp\left(\gamma_0^2\zeta(2a)\right)K}, \quad \forall K > 0.$$

Thus,

$$\mathbb{E}_n^\star\left[Q_t^\star \mathbb{1}\left\{T_K^\star \geq t\right\}\right] \leq \frac{C_{\exp\left(\gamma_0^2\zeta(2a)\right)K}}{\exp\left(\gamma_0^2\zeta(2a)\right)}\gamma_t^\star, \quad t \geq 0,$$

in which $C_{\exp\left(\gamma_0^2\zeta(2a)\right)K}/\exp\left(\gamma_0^2\zeta(2a)\right)$ is an absolute constant.

F.5 Proof of Lemma 5

For convenience of future reference, we divide the proof into two parts.

Part (i): Let $d_t := \theta_t^\star - \theta_t^\circ$ denote the approximation error. Then, d_t follows $d_t = d_{t-1} - \gamma_t^\star P_n \zeta_t$, where

$$\zeta_t := \bar{G}_n(\theta_{t-1}^\star)' W_n \bar{g}_n(\theta_{t-1}^\star) - \bar{\mathcal{H}}_n(\theta_{t-1}^\circ - \hat{\theta}_{n,W}).$$

Let

$$\kappa_t := \zeta_t - \bar{\mathcal{H}}_n d_{t-1} = \bar{G}_n(\theta_{t-1}^\star)' W_n \bar{g}_n(\theta_{t-1}^\star) - \bar{\mathcal{H}}_n(\theta_{t-1}^\star - \hat{\theta}_{n,W}).$$

Conditional on E_n, the first-order Taylor expansion of $\theta \mapsto \bar{G}_n(\theta)' W_n \bar{g}_n(\theta)$ around its critical point $\hat{\theta}_{n,W}$ gives

$$\bar{G}_n(\theta_{t-1}^\star)' W_n \bar{g}_n(\theta_{t-1}^\star) = \bar{\mathcal{H}}_n(\theta_{t-1}^\star - \hat{\theta}_{n,W}) + \kappa_t,$$

where, for some $\theta_s := s\theta_{t-1}^\star + (1-s)\hat{\theta}_{n,W}$ with $s \in (0,1)$,

$$\kappa_t = \left[\bar{G}_n(\theta_s)' W_n \bar{G}_n(\theta_s) + \sum_j \bar{g}_{nj}(\theta_s) W_n \frac{\partial^2}{\partial\theta\partial\theta'} \bar{g}_{nj}(\theta_s) - \bar{\mathcal{H}}_n\right](\theta_{t-1}^\star - \hat{\theta}_{n,W}).$$

Since

$$
\begin{aligned}
&\left\|\bar{G}_n(\theta_s)'W_n\bar{G}_n(\theta_s)+\sum_j \bar{g}_{nj}(\theta_s)W_n\frac{\partial^2}{\partial\theta\partial\theta'}\bar{g}_{nj}(\theta_s)-\bar{\mathcal{H}}_n\right\| \\
=&\left\|\bar{G}_n(\theta_s)'W_n\bar{G}_n(\theta_s)-\bar{G}_n(\hat{\theta}_{n,W})'W_n\bar{G}_n(\hat{\theta}_{n,W})+\sum_j(\bar{g}_{nj}(\theta_s)-\bar{g}_{nj}(\hat{\theta}_{n,W}))W_n\frac{\partial^2}{\partial\theta\partial\theta'}\bar{g}_{nj}(\theta_s)\right. \\
&\left.-\sum_j \bar{g}_{nj}(\hat{\theta}_{n,W})W_n\left(\frac{\partial^2}{\partial\theta\partial\theta'}\bar{g}_{nj}(\theta_s)-\frac{\partial^2}{\partial\theta\partial\theta'}\bar{g}_{nj}(\hat{\theta}_{n,W})\right)\right\| \\
\le&\left\|\bar{G}_n(\theta_s)'W_n\bar{G}_n(\theta_s)-\bar{G}_n(\hat{\theta}_{n,W})'W_n\bar{G}_n(\hat{\theta}_{n,W})\right\|+\left\|\sum_j(\bar{g}_{nj}(\theta_s)-\bar{g}_{nj}(\hat{\theta}_{n,W}))W_n\frac{\partial^2}{\partial\theta\partial\theta'}\bar{g}_{nj}(\theta_s)\right\| \\
&+\left\|\sum_j \bar{g}_{nj}(\hat{\theta}_{n,W})W_n\left(\frac{\partial^2}{\partial\theta\partial\theta'}\bar{g}_{nj}(\theta_s)-\frac{\partial^2}{\partial\theta\partial\theta'}\bar{g}_{nj}(\hat{\theta}_{n,W})\right)\right\| \\
\le& C(\|\theta^\star_{t-1}-\hat{\theta}_{n,W}\|+\|\bar{g}_n(\hat{\theta}_{n,W})\|)
\end{aligned}
$$

on the event E_n, we have

$$
\begin{aligned}
\|\kappa_t\| &\le C\|\theta^\star_{t-1}-\hat{\theta}_{n,W}\|(\|\theta^\star_{t-1}-\hat{\theta}_{n,W}\|+\|\bar{g}_n(\hat{\theta}_{n,W})\|) \\
&\le C(\|\theta^\star_{t-1}-\hat{\theta}_{n,W}\|^2+\|\bar{g}_n(\theta_o)\|\|\theta^\star_{t-1}-\hat{\theta}_{n,W}\|),
\end{aligned}
$$

where we use $\|\bar{g}_n(\hat{\theta}_{n,W})\| \le C\|\bar{g}_n(\theta_o)\|$. This shows

$$
\|\kappa_t\| \le C(\|\theta^\star_{t-1}-\hat{\theta}_n\|^2+\|\bar{g}_n(\theta_o)\|\|\theta^\star_{t-1}-\hat{\theta}_{n,W}\|)
$$

on the event E_n. Recall the definition of $\tau^\star_\varepsilon$ in (34). Conditional on $\{\tau^\star_{\underline{c}/2} \le t\} \cap E_n$, we have

$$
\underline{c}(\|\theta^\star_t-\hat{\theta}_{n,W}\|^2 \wedge 1) \le \bar{Q}_{n,W}(\theta^\star_t) \le \underline{c}/2,
$$

and hence $\|\theta^\star_t-\hat{\theta}_{n,W}\|<1$ and

$$
\begin{aligned}
\|\kappa_t\| &\le C\left((\|\theta^\star_{t-1}-\hat{\theta}_n\|^2 \wedge 1)+\|\bar{g}_n(\theta_o)\|(\|\theta^\star_{t-1}-\hat{\theta}_n\| \wedge 1)\right) \\
&\le C(Q^\star_{t-1}+\|\bar{g}_n(\theta_o)\|\sqrt{Q^\star_{t-1}}).
\end{aligned}
$$

Part (ii): Using the recurrence relation $d_t = d_{t-1}-\gamma^\star_t P_n\bar{\mathcal{H}}_n d_{t-1}-\gamma^\star_t P_n\kappa_t$, for each

$t \geq 1$ and $n \in \mathbb{N}$, it holds

$$\sum_{s=1}^{t}(\theta_s^\star - \theta_s^\circ) = \sum_{s=1}^{t} d_s = -\sum_{s=1}^{t} \alpha_s^t(P_n\bar{\mathcal{H}}_n)P_n\kappa_s,$$

where $\alpha_s^t(P_n\bar{\mathcal{H}}_n) = \gamma_s^\star \sum_{i=s}^{t}\prod_{k=s+1}^{i}(I_d - \gamma_k^\star P_n\bar{\mathcal{H}}_n)$. Note that the eigenvalues of $P_n\bar{\mathcal{H}}_n$ coincide with those of $P_n^{1/2}\bar{\mathcal{H}}_n P_n^{1/2}$, which are bounded above and below by positive constants on the event E_n by (D.1) and (D.4). Thus, the sequence $(P_n\bar{\mathcal{H}}_n)_{n\geq 1}$ lies in a compact subset of the space of negative Hurwitz matrices. Consequently, for all $n \in \mathbb{N}$, $\sup_{1\leq s\leq t<\infty}\|\alpha_s^t(P_n\bar{\mathcal{H}}_n)\| \leq C$ on the event E_n by Lemma 10.

Let us use the shorthand $\tau_{\underline{c}/2}^\star = \tau^\star$. We then have, on the event E_n,

$$\begin{aligned}
\sup_{1\leq t\leq T}\left\|\sum_{s=1}^{t}(\theta_s^\star - \theta_s^\circ)\right\| &\leq C\left(\sum_{s=1}^{\tau^\star}\|\kappa_s\| + \sum_{s=\tau^\star+1}^{T}\|\kappa_s\|\right) \\
&\leq C\sum_{s=1}^{\tau^\star}\|\kappa_s\| + C\sum_{s=\tau^\star+1}^{T}\left[Q_{t-1}^\star + \|\bar{g}_n(\theta_o)\|\sqrt{Q_{t-1}^\star}\right] \\
&=: S_1 + S_2.
\end{aligned}$$

Let $\varepsilon > 0$ be given. By Lemma 3, we can find $\tilde{T} < \infty$ that does not depend on n and $N^\star$ such that $\sup_{n\in\mathbb{N}}\mathbb{P}_n^\star(\tau^\star > \tilde{T})\mathbb{1}_{E_n} < \varepsilon$.

We address the first term S_1. Using $\theta_t^\star - \theta_{t-1}^\star = \gamma_t^\star P_n\nabla\bar{Q}_{n,W}(\theta_{t-1}^\star) - \gamma_t^\star P_n\xi_t(\theta_{t-1}^\star)$, we have

$$\begin{aligned}
\|\kappa_t\| &\leq C\left(\|\theta_{t-1}^\star - \hat{\theta}_{n,W}\|^2 + \|\bar{g}_n(\theta_o)\|^2\right) \\
&\leq C_t\left(\|\theta_0^\star - \hat{\theta}_{n,W}\|^2 + \sum_{s=1}^{t-1}\|\nabla\bar{Q}_{n,W}(\theta_{s-1}^\star)\|^2 + \sum_{s=1}^{t-1}\|\xi_s(\theta_{s-1}^\star)\|^2 + \|\bar{g}_n(\theta_o)\|^2\right)
\end{aligned}$$

on the event E_n. Thus, there exists an absolute constant $C_{\tilde{T}}$ depending on $\tilde{T}$ such that

$$\sum_{s=1}^{\tilde{T}}\|\kappa_s\| \leq C_{\tilde{T}}\left(\|\theta_0^\star - \hat{\theta}_{n,W}\|^2 + \sum_{s=1}^{\tilde{T}-1}\|\nabla\bar{Q}_{n,W}(\theta_{s-1}^\star)\|^2 + \sum_{s=1}^{\tilde{T}-1}\|\xi_s(\theta_{s-1}^\star)\|^2 + \|\bar{g}_n(\theta_o)\|^2\right).$$

Since $\|\nabla \bar{Q}_{n,W}(\theta^\star_{s-1})\|^2 \leq C Q^\star_{s-1}$, we have, on the event E_n,

$$
\begin{aligned}
&\mathbb{P}^\star_n\left(\frac{1}{\sqrt{T}}\sum_{s=1}^{\tau^\star}\|\kappa_s\| > \varepsilon\right) \\
&\leq \mathbb{P}^\star_n\left(\sum_{s=1}^{\tilde{T}}\|\kappa_s\| > \sqrt{T}\varepsilon, \tau^\star \leq \tilde{T}\right) + \mathbb{P}^\star_n(\tau^\star > \tilde{T}) \\
&\leq \mathbb{P}^\star_n\left(\sum_{s=1}^{\tilde{T}-1}(\bar{Q}^\star_s + \|\xi_s(\theta^\star_{s-1})\|^2) > \sqrt{T}\varepsilon/(2C_{\tilde{T}})\right) + \mathbb{1}\{\|\bar{g}_n(\theta_o)\|^2 + \|\theta^\star_0 - \hat{\theta}_{n,W}\|^2 > \sqrt{T}\varepsilon/(2C_{\tilde{T}})\} + \varepsilon.
\end{aligned}
$$

It is clear that

$$
\mathbb{P}(\mathbb{1}\{\|\bar{g}_n(\theta_o)\|^2 + \|\theta^\star_0 - \hat{\theta}_{n,W}\|^2 > \sqrt{T}\varepsilon/(2C_{\tilde{T}})\} > 0) \to 0
$$

as $n \to \infty$ and $T \to \infty$ uniformly in $N^\star \geq 0$. The first probability satisfies, for all $x > 0$ and $K > 0$,

$$
\begin{aligned}
\mathbb{P}^\star_n\left(\sum_{s=1}^{\tilde{T}-1}(\bar{Q}^\star_s + \|\xi_s(\theta^\star_{s-1})\|^2) > x\right) &\leq \mathbb{P}^\star_n\left(\sum_{s=1}^{\tilde{T}-1}(\bar{Q}^\star_s + \|\xi_s(\theta^\star_{s-1})\|^2) > x,\ T^\star_K \geq \tilde{T}\right) + \mathbb{P}^\star_n(T^\star_K < \infty) \\
&\leq \frac{1}{x}\mathbb{E}^\star_n\left[\sum_{s=1}^{\tilde{T}-1}(\bar{Q}^\star_s + \|\xi_s(\theta^\star_{s-1})\|^2)\mathbb{1}\left\{T^\star_K \geq \tilde{T}\right\}\right] + \mathbb{P}^\star_n(T^\star_K < \infty).
\end{aligned}
$$

We can find K such that, $\sup_{n\in\mathbb{N}}\mathbb{P}^\star_n(T^\star_K < \infty)\mathbb{1}_{E_n} < \varepsilon$ regardless of $N^\star \geq 0$. Since

$$
\mathbb{E}^\star_n[\|\xi_s(\theta^\star_{s-1})\|^2 | \mathcal{F}^\star_{n,s-1}] \leq C(Q^\star_{s-1} + 1)
$$

by (D.10), and the event $\{T_K^\star \geq s\}$ is $\mathcal{F}_{n,s-1}^\star$-measurable, we have, for all $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}_n^\star\left[\sum_{s=1}^{\tilde{T}-1}(\bar{Q}_s^\star + \|\xi_s(\theta_{s-1}^\star)\|^2)\mathbb{1}\left\{T_K^\star \geq \tilde{T}\right\}\right] &\leq \frac{1}{x}\mathbb{E}_n^\star\left[\sum_{s=1}^{\tilde{T}-1}(\bar{Q}_s^\star + \|\xi_s(\theta_{s-1}^\star)\|^2)\mathbb{1}\left\{T_K^\star \geq s\right\}\right] \\ &\leq \frac{C}{x}\mathbb{E}_n^\star\left[\sum_{s=1}^{\tilde{T}-1}(\bar{Q}_s^\star + 1)\mathbb{1}\left\{T_K^\star \geq s\right\}\right] \\ &\leq \frac{C}{x}\sum_{s=1}^{\tilde{T}-1}(C_K\gamma_s^\star + 1) \leq \frac{C}{x}\sum_{s=1}^{\tilde{T}-1}(s^{-a}+1)\end{aligned}$$

on the event E_n by Lemma 4. Thus, we have, as $T \to \infty$,

$$\sup_{n\in\mathbb{N}, N^\star \geq 0}\mathbb{P}_n^\star\left(\sum_{s=1}^{\tilde{T}-1}(\bar{Q}_s^\star + \|\xi_s(\theta_{s-1}^\star)\|^2) > \sqrt{T}\varepsilon/(2C_{\tilde{T}})\right)\mathbb{1}_{E_n} < 2\varepsilon.$$

Turning to the second term S_2, note that

$$\sum_{s=\tau^\star+1}^{T}\left[Q_{s-1}^\star + \|\bar{g}_n(\theta_o)\|\sqrt{Q_{s-1}^\star}\right] \leq \sum_{s=1}^{T}\left[Q_{s-1}^\star + \|\bar{g}_n(\theta_o)\|\sqrt{Q_{s-1}^\star}\right].$$

Conditional on E_n, we have, by Lemma 4,

$$\begin{aligned}&\mathbb{E}_n^\star\sum_{s=1}^{T}\left[Q_{s-1}^\star + \|\bar{g}_n(\theta_o)\|\sqrt{Q_{s-1}^\star}\right]\mathbb{1}\left\{T_K^\star \geq T\right\} \\ \leq\ &\mathbb{E}_n^\star\sum_{s=1}^{T}\left[Q_{s-1}^\star + \|\bar{g}_n(\theta_o)\|\sqrt{Q_{s-1}^\star}\right]\mathbb{1}\left\{T_K^\star \geq s-1\right\} \\ \leq\ &\sum_{s=1}^{T}(C_K\gamma_{s-1}^\star + \|\bar{g}_n(\theta_o)\|\sqrt{C_K\gamma_{s-1}^\star}).\end{aligned}$$

This yields, on the event E_n,

$$\begin{aligned}\mathbb{P}_n^\star\left(S_2 > \sqrt{T}\varepsilon\right) &\leq \mathbb{P}_n^\star\left(S_2 > \sqrt{T}\varepsilon,\ T_K^\star \geq T\right) + \mathbb{P}_n^\star(T_K^\star < \infty) \\ &\leq \frac{1}{\varepsilon\sqrt{T}}\mathbb{E}_n^\star[S_2\mathbb{1}\left\{T_K^\star \geq T\right\}] + \varepsilon\end{aligned}$$

$$\leq \frac{C}{\varepsilon\sqrt{T}}\left(\sum_{t=1}^{T}(\gamma_t^\star + \|\bar{g}_n(\theta_o)\|\sqrt{\gamma_t})\right) + \varepsilon$$
$$\leq C\varepsilon^{-1}\left(T^{1/2-a} + \|\bar{g}_n(\theta_o)\|T^{1/2-a/2}\right) + \varepsilon.$$

As $n \to \infty$, $T \to \infty$, and $T^{1-a}/n \to 0$,

$$T^{1/2-a} + \|\bar{g}_n(\theta_o)\|T^{1/2-a/2} = T^{1/2-a} + \sqrt{n}\|\bar{g}_n(\theta_o)\|(T^{1-a}/n)^{1/2} < \varepsilon^2/C$$

w.p.a.1 uniformly in $N^\star \geq 0$.

Putting these pieces together, we conclude, as $n \to \infty$, $T \to \infty$, and $T^{1-a}/n \to 0$,

$$\mathbb{P}\left[\mathbb{P}_n^\star\left(\frac{1}{\sqrt{T}}\sup_{1\leq t\leq T}\left\|\sum_{s=1}^{t}(\theta_s^\star - \theta_s^\circ)\right\| \geq 2\varepsilon\right) \geq 6\varepsilon\right]$$
$$\leq \mathbb{P}\left[\mathbb{P}_n^\star\left(\frac{1}{\sqrt{T}}(S_1 + S_2) \geq 2\varepsilon\right)\mathbb{1}_{E_n} \geq 6\varepsilon\right] + \mathbb{P}(E_n^c)$$
$$\leq \mathbb{P}\left[\mathbb{P}_n^\star\left(\frac{1}{\sqrt{T}}S_1 \geq \varepsilon\right)\mathbb{1}_{E_n} \geq 3\varepsilon\right] + \mathbb{P}\left[\mathbb{P}_n^\star\left(\frac{1}{\sqrt{T}}S_2 \geq \varepsilon\right)\mathbb{1}_{E_n} \geq 3\varepsilon\right] + \mathbb{P}(E_n^c) \to 0,$$

where the convergence is uniform in $N^\star \geq 0$.

F.6 Proof of Lemma 6

We use a decomposition analogous to Chen et al. (2025). However, in the presence of n and $N^\star$, the argument differs in that (i) $\theta_0^\star$ and $\gamma_t^\star$ can depend on $N^\star$, (ii) $\alpha_s^t(P_n\bar{\mathcal{H}}_n)$ and $w_s^t(P_n\bar{\mathcal{H}}_n)$ vary with n, (iii) the Gaussian approximation holds as n, T, and $N^\star$ shift simultaneously.

We begin by writing $\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr\rfloor}(\theta_t^\circ - \hat{\theta}_{n,W}) = I_0(r) + I_1(r) + I_2(r)$, where

$$I_0(r) = \frac{1}{\sqrt{T}\gamma_0^\star}\alpha_0^{\lfloor Tr\rfloor}(P_n\bar{\mathcal{H}}_n)P_n(\theta_0^\star - \hat{\theta}_{n,W}),$$
$$I_1(r) = -\bar{\mathcal{H}}_n^{-1}\frac{1}{\sqrt{T}}\sum_{s=1}^{\lfloor Tr\rfloor}\xi_s(\theta_{s-1}^\star),$$
$$I_2(r) = \frac{1}{\sqrt{T}}\sum_{s=1}^{\lfloor Tr\rfloor}w_s^{\lfloor Tr\rfloor}(P_n\bar{\mathcal{H}}_n)P_n\xi_s(\theta_{s-1}^\star).$$

and $\alpha_s^t(P_n\bar{\mathcal{H}}_n) = \gamma_s^\star\sum_{i=s}^{t}\prod_{k=s+1}^{i}(I_d - \gamma_k^\star P_n\bar{\mathcal{H}}_n)$ and $w_s^t(P_n\bar{\mathcal{H}}_n) := (P_n\bar{\mathcal{H}}_n)^{-1} - \alpha_s^t(P_n\bar{\mathcal{H}}_n)$

for $1 \le s \le t$. As noted in the proof of Lemma 5, $(P_n \bar{\mathcal{H}}_n)_{n\ge 1}$ lies in a compact subset of the space of negative Hurwitz matrices conditional on the event E_n. Hence, on the event E_n, $\sup_{1\le s\le t} \|\alpha_s^t(P_n\bar{\mathcal{H}}_n)\| \le C$ and $\sum_{s=1}^t \|w_s^t(P_n\bar{\mathcal{H}}_n)\| \le C(t+N^\star)^a$ for all $t \ge 1$ by Lemma 10.

Since $\frac{1}{\sqrt{T}\gamma_0^\star} \le C\frac{(N^\star+1)^a}{\sqrt{T}}$, $\sup_{0\le r\le 1}\|\alpha_0^{\lfloor Tr\rfloor}(P_n\bar{\mathcal{H}}_n)P_n\| \le C$, and

$$(\|\theta_0^\star - \hat{\theta}_{n,W}\| \wedge 1)^2 \le CQ_0^\star \le C\gamma_0^\star \le C(N^\star+1)^{-a},$$

we have, for every $\varepsilon > 0$, $\mathbb{P}_n^\star\left(\sup_{0\le r\le 1}\|I_0(r)\| \ge \varepsilon\right)\mathbb{1}_{E_n} \xrightarrow{\mathbb{P}} 0$ as $n, T \to \infty$, $N^\star/T = O(1)$, and $N^{\star a}/T \to 0$.

For $I_1(r)$, we verify that $I_1(r) \rightsquigarrow ((G_o'WG_o)^{-1}\Sigma_o(G_o'WG_o)^{-1})^{1/2}W(r)$ at the end of this proof.

To address the last term $\sup_{0\le r\le 1}\|I_2(r)\| \xrightarrow{\mathbb{P}}\to 0$, we first note that, for all $\varepsilon > 0$ and $K > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{r\in[0,1]}\|I_2(r)\| > \varepsilon\right)\\
&\le \mathbb{P}\left(\sup_{r\in[0,1]}\|I_2(r)\| > \varepsilon, E_n\right) + \mathbb{P}(E_n^c)\\
&\le \mathbb{P}\left(\sup_{r\in[0,1]}\|I_2(r)\| > \varepsilon, E_n, T_K^\star = \infty\right) + \mathbb{P}(E_n, T_K^\star < \infty) + \mathbb{P}(E_n^c)\\
&\le \mathbb{P}\left(\sup_{r\in[0,1]}\|I_2(r)\| > \varepsilon, E_n, T_K^\star = \infty\right) + \mathbb{E}\left[\sup_{n\in\mathbb{N}}\mathbb{P}_n^\star(T_K^\star < \infty)\mathbb{1}_{E_n}\right] + \mathbb{P}(E_n^c).
\end{aligned}$$

By Lemma 3, for every $\varepsilon > 0$, we can find $K > 0$ independent of $N^\star$ such that $\sup_{n\in\mathbb{N}}\mathbb{P}_n^\star(T_K^\star < \infty)\mathbb{1}_{E_n} < \varepsilon$. Moreover, $\mathbb{P}(E_n^c) \to 0$ as $n \to \infty$. Thus, it suffices to show that, for each given $K > 0$, $\sup_{r\in[0,1]}\|I_2(r)\| \xrightarrow{\mathbb{P}} 0$ on the event $E_n \cap \{T_K^\star = \infty\}$. Specifically, we establish that, for all $n \in \mathbb{N}$ and $T \in \mathbb{N}$,

$$\mathbb{E}_n^\star\left[\sum_{t=1}^T\left\|\sum_{s=1}^t w_s^t(P_n\bar{\mathcal{H}}_n)P_n\xi_s(\theta_{s-1}^\star)\mathbb{1}\left\{T_K^\star \ge s\right\}\right\|^{2p}\right]\mathbb{1}_{E_n} \le C_K(T+N^\star)^{ap+1}$$

for an absolute constant $C_K > 0$ that depends only on K. This bound is sufficient,

since conditional on the event $E_n \cap \{T_K^\star = \infty\}$,

$$\begin{aligned}\sup_{0\le r\le 1}\|I_2(r)\|^{2p} &\le T^{-p}\sup_{1\le t\le T}\left\|\sum_{s=1}^{t} w_s^t(P_n\bar{\mathcal{H}}_n)P_n\xi_s(\theta_{s-1}^\star)\mathbb{1}\left\{T_K^\star\ge s\right\}\right\|^{2p}\mathbb{1}_{E_n}\\ &\le T^{-p}\sum_{t=1}^{T}\left\|\sum_{s=1}^{t} w_s^t(P_n\bar{\mathcal{H}}_n)P_n\xi_s(\theta_{s-1}^\star)\mathbb{1}\left\{T_K^\star\ge s\right\}\right\|^{2p}\mathbb{1}_{E_n},\end{aligned}$$

which must be on the order of $O_{\mathbb{P}}(T^{1-(1-a)p}) = o_{\mathbb{P}}(1)$ under $N^\star/T = O(1)$ and $p > (1-a)^{-1}$.

By applying Burkholder's inequality to the martingale difference partial sum, $\sum_{s=1}^{t} w_s^t(P_n\bar{\mathcal{H}}_n)P_n\xi_s(\theta_{s-1}^\star)\mathbb{1}\left\{T_K^\star\ge s\right\}$, we observe that for all $t\in\mathbb{N}$,

$$\begin{aligned}&\mathbb{E}_n^\star\left\|\sum_{s=1}^{t} w_s^t(P_n\bar{\mathcal{H}}_n)P_n\xi_s(\theta_{s-1}^\star)\mathbb{1}\left\{T_K^\star\ge s\right\}\right\|^{2p}\\ &\le C_p\mathbb{E}_n^\star\left[\left(\sum_{s=1}^{t}\|w_s^t(P_n\bar{\mathcal{H}}_n)P_n\|^2\|\xi_s(\theta_{s-1}^\star)\|^2\mathbb{1}\left\{T_K^\star\ge s\right\}\right)^p\right].\end{aligned}$$

Here, C_p denotes an absolute constant depending only on p. Applying Hölder's inequality

$$\sum_s |a_s b_s| \le \left(\sum_s |a_s|^\alpha\right)^{1/\alpha}\left(\sum_s |b_s|^\beta\right)^{1/\beta}$$

with $(a_s, b_s) = (\|w_s^t(P_n\bar{\mathcal{H}}_n)P_n\|^{2-2/p}, \|w_s^t(P_n\bar{\mathcal{H}}_n)P_n\|^{2/p}\|\xi_s(\theta_{s-1}^\star)\|^2\mathbb{1}\left\{T_K^\star\ge s\right\})$ and $(\alpha,\beta) = (p/(p-1), p)$, we get

$$\begin{aligned}&\sum_{s=1}^{t}\|w_s^t(P_n\bar{\mathcal{H}}_n)P_n\|^2\|\xi_s(\theta_{s-1}^\star)\|^2\mathbb{1}\left\{T_K^\star\ge s\right\}\\ &\le \left(\sum_{s=1}^{t}\|w_s^t(P_n\bar{\mathcal{H}}_n)P_n\|^2\right)^{\frac{p-1}{p}}\left(\sum_{s=1}^{t}\|w_s^t(P_n\bar{\mathcal{H}}_n)P_n\|^2\|\xi_s(\theta_{s-1}^\star)\|^{2p}\mathbb{1}\left\{T_K^\star\ge s\right\}\right)^{\frac{1}{p}}.\end{aligned}$$

This implies that, for all $t \in \mathbb{N}$,

$$
\begin{aligned}
&\mathbb{E}_n^\star \left[\left(\sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \|\xi_s(\theta_{s-1}^\star)\|^2 \mathbb{1}\left\{T_K^\star \geq s\right\} \right)^p \right] \\
&\leq \left(\sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \right)^{p-1} \sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \mathbb{E}_n^\star \left[\|\xi_s(\theta_{s-1}^\star)\|^{2p} \mathbb{1}\left\{T_K^\star \geq s\right\} \right] \\
&\leq \left(\sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \right)^{p-1} \sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \mathbb{E}_n^\star \left[\mathbb{E}_n^\star [\|\xi_s(\theta_{s-1}^\star)\|^{2p} | \mathcal{F}_{n,s-1}^\star] \mathbb{1}\left\{T_K^\star \geq s\right\} \right] \\
&\leq C \left(\sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \right)^{p-1} \sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \mathbb{E}_n^\star \left[({Q_{s-1}^\star}^p + 1) \mathbb{1}\left\{T_K^\star \geq s\right\} \right] \\
&\leq C \left(\sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \right)^{p} (K^p + 1)
\end{aligned}
$$

on the event E_n, where we use the fact that

$$
\mathbb{E}_n^\star \left[\|\xi_s(\theta_{s-1}^\star)\|^{2p} | \mathcal{F}_{n,s-1}^\star \right] = \mathbb{E}_n^\star [\|\xi_s(\theta)\|^{2p}] \big|_{\theta = \theta_{s-1}^\star} \leq M({Q_{s-1}^\star}^p + 1)
$$

on the event E_n, and $Q_{s-1}^\star \leq K$ conditional on $\{T_K^\star \geq s\}$. Uniformly in $s \in \mathbb{N}$, $t \geq s$, and $n \in \mathbb{N}$, $\|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\| \leq \|\alpha_s^t(P_n \bar{\mathcal{H}}_n) P_n\| + \|\bar{\mathcal{H}}_n^{-1}\| \leq C$ on the event E_n, whence it follows, for all $t \geq 1$ and $n \in \mathbb{N}$,

$$
\sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \leq C \sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\| \leq C(t + N^\star)^a,
$$

and therefore,

$$
\begin{aligned}
\mathbb{E}_n^\star \left[\sum_{t=1}^T \left\| \sum_{s=1}^t w_s^t(\bar{\mathcal{H}}_n) \xi_s(\theta_{s-1}^\star) \mathbb{1}\left\{T_K^\star \geq s\right\} \right\|^{2p} \right] &\leq C_K \sum_{t=1}^T \left(\sum_{s=1}^t \|w_s^t(P_n \bar{\mathcal{H}}_n) P_n\|^2 \right)^p \\
&\leq C_K \sum_{t=1}^T (t + N^\star)^{ap} \leq C_K (T + N^\star)^{ap+1}.
\end{aligned}
$$

This concludes $\sup_{0 \leq r \leq 1} \|I_2(r)\| \xrightarrow{\mathbb{P}} 0$, as discussed before.

Finally, we show that the FCLT applies to $\{I_1(r) : r \in [0,1]\}$ by establishing the conditional Lindeberg conditions for $\frac{1}{\sqrt{T}} \sum_{s=1}^{\lfloor Tr \rfloor} \xi_s(\theta_{s-1}^\star)$. On the event E_n, we show

that as $n \to \infty$ and $T \to \infty$, and uniformly in $N^\star \geq 0$,

$$\frac{1}{T}\sum_{s=1}^{T} \mathbb{E}_n^\star \left[\|\xi_s(\theta_{s-1}^\star)\|^2 \mathbb{1}\left\{ \|\xi_s(\theta_{s-1}^\star)\| > T\delta \right\} \middle| \mathcal{F}_{n,s-1}^\star \right] \xrightarrow{\mathbb{P}} 0 \quad \text{for any } \delta > 0, \tag{42}$$

$$\frac{1}{T}\sum_{s=1}^{\lfloor Tr \rfloor} \mathbb{E}_n^\star \left[\xi_s(\theta_{s-1}^\star)\xi_s(\theta_{s-1}^\star)' \middle| \mathcal{F}_{n,s-1}^\star \right] \xrightarrow{\mathbb{P}} r\Sigma_o \quad \text{for any } r \in [0,1]. \tag{43}$$

These conditions together imply that $\frac{1}{\sqrt{T}}\sum_{s=1}^{\lfloor Tr \rfloor} \xi_s(\theta_{s-1}^\star) \rightsquigarrow \Sigma_o^{1/2} W(r)$. Independence between $W(r)$ and $\sigma((z_i)_{i\geq 1}) \subseteq \sigma(\cup_{n=1}^\infty \mathcal{F}_{n,0}^\star)$ follows from the fact that the asymptotic variance is a non-random constant and hence weak convergence becomes mixing (Hall and Heyde (1980)). Combined with $\bar{\mathcal{H}}_n \xrightarrow{\mathbb{P}} G_o' W G_o$, which is shown in the proof of (D.4), this implies that $I_1(r) \rightsquigarrow ((G_o' W G_o)^{-1}\Sigma_o (G_o' W G_o)^{-1})^{1/2} W(r)$. Now, we verify the conditions (42) and (43).

Proof of (42)*:* By Lemma 3, there exists K that does not depend on $N^\star$ such that $\limsup_{n\to\infty, T\to\infty} \mathbb{P}(T_K^\star < \infty) < \varepsilon$ uniformly in $N^\star$. Thus, it suffices to prove instead

$$\frac{1}{T}\sum_{s=1}^{T} \mathbb{E}_n^\star \left[\|\xi_s(\theta_{s-1}^\star)\|^2 \mathbb{1}\left\{ \|\xi_s(\theta_{s-1}^\star)\| > T\delta \right\} \middle| \mathcal{F}_{n,s-1}^\star \right] \mathbb{1}\left\{T_K^\star \geq s\right\} \mathbb{1}_{E_n} \xrightarrow{\mathbb{P}} 0,$$

for any given $K > 0$, which suffices for (42). On the event E_n, by Markov's inequality,

$$\begin{aligned}\mathbb{E}_n^\star \left[\|\xi_s(\theta_{s-1}^\star)\|^2 \mathbb{1}\left\{ \|\xi_s(\theta_{s-1}^\star)\| > T\delta \right\} \middle| \mathcal{F}_{n,s-1}^\star \right] &\leq (T\delta)^{2(1-p)} \mathbb{E}_n^\star \left[\|\xi_s(\theta_{s-1}^\star)\|^{2p} \middle| \mathcal{F}_{n,s-1}^\star \right] \\ &\leq C(T\delta)^{2(1-p)} ({Q_{s-1}^\star}^p + 1).\end{aligned}$$

This implies, for any $K > 0$,

$$\begin{aligned}&\frac{1}{T}\sum_{s=1}^{T} \mathbb{E}_n^\star \left[\|\xi_s(\theta_{s-1}^\star)\|^2 \mathbb{1}\left\{ \|\xi_s(\theta_{s-1}^\star)\| > T\delta \right\} \middle| \mathcal{F}_{n,s-1}^\star \right] \mathbb{1}\left\{T_K^\star \geq s\right\} \mathbb{1}_{E_n} \\ &\leq C(T\delta)^{2(1-p)} \frac{1}{T}\sum_{s=1}^{T} ({Q_{s-1}^\star}^p + 1)\mathbb{1}\left\{T_K^\star \geq s\right\} \leq C(T\delta)^{2(1-p)}(K^p + 1) \xrightarrow{\mathbb{P}} 0\end{aligned}$$

as $n \to \infty$ and $T \to \infty$ uniformly in $N^\star$. The desired claim follows.

Proof of (43)*:* Recall the definition of $\Sigma_{n,t}(\theta)$: $\Sigma_{n,t}(\theta) = \mathbb{E}_n^\star[\xi_t(\theta)\xi_t(\theta)']$, which depends

on both $n \in \mathbb{N}$ and $t \geq 1$. We have also defined the following matrices:

$$\Sigma_{n,t} := \Sigma_{n,t}(\hat{\theta}_{n,W}), \quad \Sigma_t := \operatorname*{plim}_{n\to\infty} \Sigma_{n,t}, \quad \Sigma_o := \lim_{t\to\infty} \Sigma_t.$$

We begin by writing

$$\begin{aligned}\frac{1}{T}\sum_{s=1}^{\lfloor Tr \rfloor} \mathbb{E}_n^\star \left[\xi_s(\theta_{s-1}^\star)\xi_s(\theta_{s-1}^\star)' \middle| \mathcal{F}_{n,s-1}^\star \right] &= \frac{1}{T}\sum_{s=1}^{\lfloor Tr \rfloor} \Sigma_{n,s}(\theta_{s-1}^\star) \\ &= \frac{1}{T}\sum_{s=1}^{\lfloor Tr \rfloor} \Sigma_s + \frac{1}{T}\sum_{s=1}^{\lfloor Tr \rfloor} [\Sigma_{n,s} - \Sigma_s] + \frac{1}{T}\sum_{s=1}^{\lfloor Tr \rfloor} [\Sigma_{n,s}(\theta_{s-1}^\star) - \Sigma_{n,s}].\end{aligned}$$

We show that the latter two terms are negligible. Observe that

$$\begin{aligned}\xi_t(\hat{\theta}_{n,W}) &= \tilde{G}_t(\hat{\theta}_{n,W})' W_n \tilde{g}_t(\hat{\theta}_{n,W}) = \tilde{G}_t(\hat{\theta}_{n,W})' W_n \tilde{g}_t(\hat{\theta}_{n,W}) - \bar{G}_n' W_n \bar{g}_n(\hat{\theta}_{n,W}) \\ &= (\tilde{G}_t(\hat{\theta}_{n,W}) - \bar{G}_n)' W_n \tilde{g}_t(\hat{\theta}_{n,W}) + \bar{G}_n' W_n (\tilde{g}_t(\hat{\theta}_{n,W}) - \bar{g}_n(\hat{\theta}_{n,W})).\end{aligned}$$

Since $\tilde{G}_t(\hat{\theta}_{n,W}) - \bar{G}_n$ is centered at zero and independent of $\tilde{g}_t(\hat{\theta}_{n,W})$ conditional on $z_{1:n}$, these two terms are uncorrelated to each other, yielding

$$\begin{aligned}\Sigma_{n,t} &= \mathbb{E}_n^\star[(\tilde{G}_t(\hat{\theta}_{n,W}) - \bar{G}_n)' W_n \tilde{g}_t(\hat{\theta}_{n,W}) \tilde{g}_t(\hat{\theta}_{n,W})' W_n (\tilde{G}_t(\hat{\theta}_{n,W}) - \bar{G}_n)] \\ &\quad + \mathbb{E}_n^\star[\bar{G}_n' W_n (\tilde{g}_t(\hat{\theta}_{n,W}) - \bar{g}_n(\hat{\theta}_{n,W}))(\tilde{g}_t(\hat{\theta}_{n,W}) - \bar{g}_n(\hat{\theta}_{n,W}))' W_n \bar{G}_n] \\ &= \frac{1}{B_g}\left(\frac{1}{B_{G,t}} \cdot \frac{1}{n}\sum_{i=1}^n G(z_i, \hat{\theta}_{n,W})' W_n \tilde{\Omega}_n W_n G(z_i, \hat{\theta}_{n,W}) + \left(1 - \frac{1}{B_{G,t}}\right) \bar{G}_n' W_n \tilde{\Omega}_n W_n \bar{G}_n \right).\end{aligned}$$

Since both averages in the last line converge to their respective limits as $n \to \infty$, we find that

$$\sup_{t\geq 1} \|\Sigma_{n,t} - \Sigma_t\| \xrightarrow{\mathbb{P}} 0 \quad \text{as} \quad n \to \infty.$$

In turn, this leads to $\frac{1}{T}\sum_{s=1}^{\lfloor Tr \rfloor} [\Sigma_{n,t} - \Sigma_t] \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$, addressing the second term. To show that the last term converges in probability to zero, it suffices to show

$$\frac{1}{T}\sum_{s=1}^{\lfloor Tr \rfloor} [\Sigma_{n,s}(\theta_{s-1}^\star) - \Sigma_{n,s}] \mathbb{1}\left\{T_K^\star \geq s\right\} \mathbb{1}_{E_n} \xrightarrow{\mathbb{P}} 0$$

for any given K. For all $n \in \mathbb{N}$ and $t \geq 1$, it holds that $\|\Sigma_{n,t}(\theta_{s-1}^\star) - \Sigma_{n,t}\| \leq$

$C\|\theta^\star_{s-1} - \hat{\theta}_{n,W}\|$ on the event E_n by (D.9). Moreover, uniformly in $t \geq 1$, on the event $E_n \cap \{T^\star_K \geq s\}$,

$$\begin{aligned}\|\Sigma_{n,t}(\theta^\star_{s-1}) - \Sigma_{n,t}\| &\leq \mathbb{E}^\star_n[\|\xi_t(\theta^\star_{t-1})\|^2 + \|\xi_t(\hat{\theta}_{n,W})\|^2 | \mathcal{F}^\star_{n,t-1}] \\ &\leq C(Q^\star_{s-1} + \bar{Q}_{n,W}(\hat{\theta}_{n,W}) + 1) \leq C(\|\bar{g}_n(\theta_o)\|^2 + K + 1),\end{aligned}$$

which is bounded by C w.p.a.1. Putting these together, we find that

$$\forall t \geq 1 : \|\Sigma_{n,t}(\theta^\star_{t-1}) - \Sigma_{n,t}\| \mathbb{1}\left\{T^\star_K \geq s\right\} \mathbb{1}_{E_n} \leq C(\|\theta^\star_{s-1} - \hat{\theta}_{n,W}\| \wedge 1) \mathbb{1}\left\{T^\star_K \geq s\right\} \mathbb{1}_{E_n}$$

w.p.a.1. Thus, as $n \to \infty$ and $T \to \infty$,

$$\begin{aligned}&\mathbb{E}\mathbb{E}^\star_n \frac{1}{T} \sum_{s=1}^{\lfloor Tr \rfloor} \left\|\Sigma_{n,t}(\theta^\star_{s-1}) - \Sigma_{n,t}\right\| \mathbb{1}\left\{T^\star_K \geq s\right\} \mathbb{1}_{E_n} \\ &\leq \mathbb{E}\frac{C}{T} \sum_{s=1}^{\lfloor Tr \rfloor} \mathbb{E}^\star_n[(\|\theta^\star_{s-1} - \hat{\theta}_{n,W}\| \wedge 1) \mathbb{1}\left\{T^\star_K \geq s-1\right\}] \mathbb{1}_{E_n} \\ &\leq \mathbb{E}\frac{C}{T} \sum_{s=1}^{T} \mathbb{E}^\star_n[\sqrt{Q^\star_{s-1}} \mathbb{1}\left\{T^\star_K \geq s-1\right\}] \mathbb{1}_{E_n} \leq \frac{1}{T} \sum_{s=1}^{T} \sqrt{C_K \gamma^\star_{s-1}} \to 0\end{aligned}$$

by Lemma 4, where the convergence is uniform in $N^\star \geq 0$. This verifies

$$\frac{1}{T} \sum_{s=1}^{\lfloor Tr \rfloor} [\Sigma_{n,t}(\theta^\star_{s-1}) - \Sigma_{n,t}] \mathbb{1}\left\{T^\star_K \geq s\right\} \mathbb{1}_{E_n} \xrightarrow{\mathbb{P}} 0$$

uniformly in $N^\star$. Now, (43) follows from the fact that $\frac{1}{T}\sum_{s=1}^{\lfloor Tr \rfloor} \Sigma_s \to r\Sigma_o$ since the deterministic sequence $(\Sigma_t)_{t \geq 1}$ converges to Σ_o as $t \to \infty$.

F.7 Proof of Lemma 7

We note that Lemma 7 arises as a special case of the approximation in Lemma 6 at $r = 1$. Relative to the proof of Lemma 6, the relaxed assumption of $p > 1$ only changes the way we address $I_1(1)$ and $I_2(1)$.

For $I_1(1)$, we first observe that the Lindeberg conditions (42) and (43) remain valid, as the arguments in the proof of Lemma 6 continue to work under $p > 1$. As a result,

the CLT applies to $I_1(r)$ with $r = 1$, yielding

$$I_1(1) \xrightarrow{d} \mathcal{N}(0, (G_o' W G_o)^{-1} \Sigma_o (G_o' W G_o)^{-1})$$

as $n, T \to \infty$.

For $I_2(1)$, by applying Burkholder's inequality with exponent 2 rather than $2p$, we have, for all $n \in \mathbb{N}$ and $T \in \mathbb{N}$, and for each $K > 0$,

$$\begin{aligned}
&\mathbb{E}_n^\star \left\| \sum_{s=1}^{T} w_s^T (P_n \bar{\mathcal{H}}_n) P_n \xi_s(\theta_{s-1}^\star) \mathbb{1}\left\{T_K^\star \geq s\right\} \right\|^2 \mathbb{1}_{E_n} \\
&\leq C \sum_{s=1}^{T} \|w_s^T (P_n \bar{\mathcal{H}}_n) P_n\|^2 \mathbb{E}_n^\star [\|\xi_s(\theta)\|^2 | \mathcal{F}_{n,s-1}] \mathbb{1}\left\{T_K^\star \geq s\right\} \mathbb{1}_{E_n} \\
&\leq C \sum_{s=1}^{T} \|w_s^T (P_n \bar{\mathcal{H}}_n) P_n\| \leq C(T + N^\star)^a.
\end{aligned}$$

As the learning rate exponent $a < 1$, we have, for all $K > 0$,

$$\sup_{n \in \mathbb{N}} \mathbb{E}_n^\star [\|I_2(1)\|^2 \mathbb{1}\left\{T_K^\star = \infty\right\}] \mathbb{1}_{E_n} \leq C(T + N^\star)^a / T \to 0$$

as $T \to \infty$ and $N^\star / T = O(1)$. This implies that $I_2(1) \xrightarrow{\mathbb{P}} 0$.

F.8 Proof of Lemma 8

See Lemma 5.2.2 in Benveniste et al. (2012).

F.9 Proof of Lemma 9

This follows from Lemma 14 in Gadat and Panloup (2023) with $u_n := v_n / n$ and $\beta_n := \gamma_n / n$.

F.10 Proof of Lemma 10

In this proof, let $V = V(A)$ denote the Lyapunov matrix of A, i.e., the unique positive definite solution to the equation $A'V + VA = I_d$. It is known that V depends continuously on A, and so do its maximum and minimum eigenvalues, denoted by

$L = L(A) = \lambda_{\max}(V(A))$ and $\ell = \ell(A) = \lambda_{\min}(V(A))$, respectively. Lemma 1 in Polyak and Juditsky (1992) reveals that

$$\left\| \prod_{j=s+1}^{i} (I_d - \gamma_j^\star A) \right\| \leq \sqrt{\frac{L}{\ell}} \exp\left(- \sum_{j=s+1}^{i} \left(\frac{\gamma_j^\star}{L} - \frac{\|A\|^2 L \gamma_j^{\star 2}}{\ell} \right) \right)$$
$$\leq \sqrt{\frac{L}{\ell}} \exp\left(\frac{\|A\|^2 L}{\ell} \gamma_0^2 \zeta(2a) \right) \exp\left(-\frac{1}{L} \sum_{j=s+1}^{i} \gamma_j^\star \right) \tag{44}$$

for all $i \geq s$, where $\zeta(2a) = \sum_{k=1}^{\infty} k^{-2a}$. For notational convenience, let us write $\alpha_s^t = \alpha_s^t(A)$, $\lambda = \lambda(A) = 1/L$, and $\kappa = \kappa(A, a, \gamma_0) := \sqrt{\frac{L}{\ell}} \exp(\frac{\|A\|^2 L}{\ell} \gamma_0^2 \zeta(2a))$. Additionally, define $w_s^t := \alpha_s^t - A^{-1}$ and $\beta_s^t := \sum_{i=s}^{t} (\gamma_s^\star - \gamma_{i+1}^\star) \prod_{k=s+1}^{i} (I_d - \gamma_k^\star A)$ so that

$$\begin{aligned}
\beta_s^t &= \alpha_s^t - \sum_{i=s}^{t} \gamma_{i+1}^\star \prod_{k=s+1}^{i} (I_d - \gamma_k^\star A) \\
&= \alpha_s^t + A^{-1} \sum_{i=s}^{t} \left(\prod_{k=s+1}^{i+1} (I_d - \gamma_k^\star A) - \prod_{k=s+1}^{i} (I_d - \gamma_k^\star A) \right) \\
&= \alpha_s^t + A^{-1} \prod_{k=s+1}^{t+1} (I_d - \gamma_k^\star A) - A^{-1} \\
&= w_s^t + A^{-1} \prod_{k=s+1}^{t+1} (I_d - \gamma_k^\star A).
\end{aligned}$$

This observation allows us to derive the asserted bound for $\sum_{s=1}^{t} \|w_s^t\|$ by tackling $\sum_{s=1}^{t} \|\beta_s^t\|$.

Let us define $f(u) := (u + N^\star)^{-a}$ for all $u > 0$. By (44), we obtain for all $s \geq 0$ and $t \geq s$,

$$\begin{aligned}
\|\alpha_s^t\| \leq \kappa \gamma_s^\star \sum_{i=s}^{t} \exp\left(-\lambda \gamma_0 \sum_{k=s+1}^{i} f(k) \right) &\leq \kappa \gamma_s^\star \sum_{i=s}^{t} \exp\left(-\lambda \gamma_0 \int_{s+1}^{i+1} f(u) du \right) \\
&= \kappa \gamma_s^\star \left(1 + \sum_{i=s+1}^{t} \exp\left(-\lambda \gamma_0 \int_{s+1}^{i+1} f(u) du \right) \right) \\
&\leq \kappa \gamma_s^\star \left(1 + \int_{s}^{t-1} \exp\left(-\lambda \gamma_0 \int_{s+1}^{x+1} f(u) du \right) dx \right)
\end{aligned}$$

where we employ the fact that $f(x)$ and $\exp(-\lambda\gamma_0 \int_{s+1}^{x+1} f(u)du)$ are both decreasing in x. Substituting x with z, defined as

$$z := (1-a)\int_{s+1}^{x+1} f(u)du = (x+N^\star+1)^{1-a} - (s+N^\star+1)^{1-a},$$

or equivalently $x = (z + (s+N^\star+1)^{1-a})^{\frac{1}{1-a}} - N^\star - 1$, we get

$$\begin{aligned}
\|\alpha_s^t\| &\le \kappa\left(\gamma_s^\star + \frac{\gamma_s^\star}{1-a}\int_0^\infty \exp\left(-\frac{\lambda\gamma_0}{1-a}z\right)(z+(s+N^\star+1)^{1-a})^{\frac{a}{1-a}}dz\right) \\
&= \kappa\left(\gamma_0 + \frac{\gamma_0}{1-a}\frac{(s+N^\star+1)^a}{((s+N^\star)\vee 1)^a}\int_0^\infty \exp\left(-\frac{\lambda\gamma_0}{1-a}z\right)\left(\frac{z}{(s+N^\star+1)^{1-a}}+1\right)^{\frac{a}{1-a}}dz\right) \\
&\le \kappa\left(\gamma_0 + \frac{\gamma_0}{1-a}2^a\int_0^\infty \exp\left(-\frac{\lambda\gamma_0}{1-a}z\right)(z+1)^{\frac{a}{1-a}}dz\right) \\
&\le \kappa\left(\gamma_0 + \frac{\gamma_0}{1-a}2^a\mathcal{L}_a\left(\frac{\lambda\gamma_0}{1-a}\right)\right),
\end{aligned}$$

where $\mathcal{L}_a(s) = \int_0^\infty e^{-sz}(z+1)^{\frac{a}{1-a}}dz$ denotes the Laplace transform of $(z+1)^{a/(1-a)}$. We conclude

$$\|\alpha_s^t\| \le \underbrace{\kappa\gamma_0\left(1 + \frac{1}{1-a}2^a\mathcal{L}_a\left(\frac{\lambda\gamma_0}{1-a}\right)\right)}_{=:M_1(A,a,\gamma_0)}$$

uniformly in $s \ge 0$ and $t \ge s$. One can see that $M_1(A,a,\gamma_0)$ as defined above is continuous in its arguments.

Next, we bound the term $\sum_{s\le t}\|\beta_s^t\|$. From (44), it follows that for all $s \ge 0$ and $t \ge s$,

$$\begin{aligned}
\|\beta_s^t\| &\le \kappa\sum_{i=s}^t(\gamma_s^\star - \gamma_{i+1}^\star)\exp\left(-\lambda\sum_{k=s+1}^i\gamma_k^\star\right) \\
&\le \frac{\kappa e^{\lambda\gamma_0}}{\lambda}\sum_{i=s}^t(e^{-\lambda\gamma_{i+1}^\star} - e^{-\lambda\gamma_s^\star})\exp\left(-\lambda\sum_{k=s+1}^i\gamma_k^\star\right)
\end{aligned}$$

where we use the fact that $\sup_{0<x<y\le\gamma_s^\star}\frac{x-y}{e^{-\lambda y}-e^{-\lambda x}} \le \frac{e^{\lambda\gamma_s^\star}}{\lambda} \le \frac{e^{\lambda\gamma_0}}{\lambda}$ in order to move from

the second line to the third. This implies that for all $t \geq 1$,

$$\begin{aligned}\sum_{s\leq t}\|\beta_s^t\| &\leq \frac{\kappa e^{\lambda\gamma_0}}{\lambda}\sum_{s=1}^{t}\sum_{i=s}^{t}(e^{-\lambda\gamma_{i+1}^\star}-e^{-\lambda\gamma_s^\star})\exp\left(-\lambda\sum_{k=s+1}^{i}\gamma_k^\star\right)\\ &= \frac{\kappa e^{\lambda\gamma_0}}{\lambda}\sum_{s=1}^{t}\left(\exp\left(-\lambda\sum_{k=s+1}^{t+1}\gamma_k^\star\right)-\exp\left(-\lambda\sum_{k=1}^{s}\gamma_k^\star\right)\right),\end{aligned}$$

where we employ telescoping sums to simplify. This in turn yields

$$\begin{aligned}\sum_{s\leq t}\|\beta_s^t\| &\leq \frac{\kappa e^{\lambda\gamma_0}}{\lambda}\sum_{s=1}^{t}\exp\left(-\lambda\sum_{k=s+1}^{t+1}\gamma_k^\star\right)\\ &\leq \frac{\kappa e^{\lambda\gamma_0}}{\lambda}\sum_{s=1}^{t}\exp\left(-\lambda(t-s)\gamma_t^\star\right)\\ &\leq \frac{\kappa e^{\lambda\gamma_0}}{\lambda}\frac{1}{1-e^{-\lambda\gamma_t^\star}} \leq \frac{\kappa e^{\lambda\gamma_0}}{\lambda}\frac{e^{\lambda\gamma_0}}{\lambda\gamma_t^\star} = \frac{\kappa e^{2\lambda\gamma_0}}{\lambda^2\gamma_0}(t+N^\star)^a.\end{aligned}$$

This establishes the second assertion by

$$\begin{aligned}\sum_{s=1}^{t}\|\alpha_s^t(A)-A^{-1}\| &\leq \sum_{s=1}^{t}\|\beta_s^t\| + \|A^{-1}\|\sum_{s=1}^{t}\left\|\prod_{k=s+1}^{t+1}(I_d-\gamma_k^\star A)\right\|\\ &\leq \frac{\kappa e^{2\lambda\gamma_0}}{\lambda^2\gamma_0}(t+N^\star)^a + \kappa\|A^{-1}\|\underbrace{\sum_{s=1}^{t}\exp\left(-\lambda\sum_{k=s+1}^{t+1}\gamma_k^\star\right)}_{\leq\frac{e^{\lambda\gamma_0}}{\lambda\gamma_0}(t+N^\star)^a}\\ &\leq \kappa\left(\frac{e^{2\lambda\gamma_0}}{\lambda^2\gamma_0}+\|A^{-1}\|\frac{e^{\lambda}\gamma_0}{\lambda\gamma_0}\right)(t+N^\star)^a =: M_2(A,a,\gamma_0)(t+N^\star)^a.\end{aligned}$$

F.11 Proof of Lemma 11

We follow the same line of arguments as in the proof of Lemma 5 with the following modifications. Let $n \in \mathbb{N}$ be given and fixed and let $N^\star = 0$. In this proof, we use the symbol $*$ rather than $\star$, and write $\gamma_t = \gamma_0 t^{-a}$ for the learning rate. The conditioning and weighting matrices are set to identity matrices and we write $\hat{\theta}_n$ for the target estimator.

The event E_n holds under Assumptions 6 and 7. Additionally, we assume that the Hessian $\theta \mapsto \frac{\partial^2 \bar{g}_{nj}}{\partial\theta\partial\theta'}(\theta)$ is Lipscthiz continuous on the δ-neighborhood of $\hat{\theta}_n$ for each

$j = 1, \ldots, d_g$ in Assumption (A6.3). Let $(\bar{L}_j)_{j=1}^{d_g}$ denote the Lipschitz constants for these Hessians. The same first-order Taylor expansion as in the proof of Lemma 5 yields, for all $t \geq 1$,

$$\|\kappa_t\| \leq C \left(\|\theta^*_{t-1} - \hat{\theta}_n\|^2 + \|\bar{g}_n(\hat{\theta}_n)\| \sqrt{\sum_{j=1}^{d_g} \bar{L}_j^2} \|\theta^*_{t-1} - \hat{\theta}_n\|^2 \right) \leq C \|\theta^*_{t-1} - \hat{\theta}_n\|^2$$

under fixed n with $\|\bar{g}_n(\theta_o)\|$ fixed. Conditional $\{\tau^* < t\}$, this implies $\|\kappa_t\| \leq C Q^*_{t-1}$.

We write, for $1 \leq t \leq N$,

$$\sum_{s=1}^{t} d_s = -\sum_{s=1}^{t} \alpha_s^t \kappa_s$$

where $\alpha_s^t := \gamma_s \sum_{i=s}^{m} \prod_{k=s+1}^{i} (I_d - \gamma_k \bar{\mathcal{H}}_n)$ is uniformly bounded in $s \geq 1$ and $t \geq s$. This yields

$$\sup_{1 \leq t \leq N} \left\| \sum_{s=1}^{t} (\theta^*_s - \theta^\circ_s) \right\| \leq C \sum_{t=1}^{N} \|\kappa_t\| \leq \sum_{t=1}^{\tau^*} \|\kappa_t\| + C \sum_{t=\tau^*+1}^{N} Q^*_{t-1}.$$

Following the argument in the proof of Lemma 5, we can show that, for every $\varepsilon > 0$,

$$\mathbb{P}^*_n \left(\frac{1}{\sqrt{N}} \sum_{t=1}^{\tau^*} \|\kappa_t\| > \varepsilon \right) \to 0 \quad \text{as} \quad N \to \infty,$$

which addresses the first term. For the second term, we have

$$\mathbb{P}^*_n \left(\frac{1}{\sqrt{N}} \sum_{t=\tau^*+1}^{N} Q^*_{t-1} > \varepsilon \right) \leq \mathbb{P}^*_n \left(\frac{1}{\sqrt{N}} \sum_{t=1}^{N} Q^*_{t-1} > \varepsilon \right) \leq \frac{C_K}{\varepsilon \sqrt{N}} \sum_{t=1}^{N} \gamma_{t-1} + \mathbb{P}^*_n (T^*_K < \infty).$$

For arbitrary $\eta > 0$, we choose K to satisfy $\mathbb{P}^*_n(T^*_K < \infty) < \eta$. Then, as $N \to \infty$,

$$\limsup_{N \to \infty} \mathbb{P}^*_n \left(\frac{1}{\sqrt{N}} \sum_{t=\tau^*+1}^{N} Q^*_{t-1} > \varepsilon \right) \leq \eta, \quad \forall \eta > 0.$$

Putting these together, we conclude that

$$\lim_{N \to \infty} \mathbb{P}^*_n \left(\frac{1}{\sqrt{N}} \sup_{1 \leq t \leq N} \left\| \sum_{s=1}^{t} (\theta^*_s - \theta^\circ_s) \right\| > \varepsilon \right) = 0.$$

F.12 Proof of Lemma 12

We adapt the proof of Lemma 6 to the case with fixed data. Consider the decomposition $\frac{1}{\sqrt{N}}\sum_{t=1}^{\lfloor Nr\rfloor}(\theta_t^* - \hat{\theta}_n) = I_0(r) + I_1(r) + I_2(r)$, where

$$I_0(r) = \frac{1}{\sqrt{N}\gamma_0}\alpha_0^{\lfloor Nr\rfloor}(\theta_0^* - \hat{\theta}_n),$$
$$I_1(r) = -\bar{\mathcal{H}}_n^{-1}\frac{1}{\sqrt{N}}\sum_{s=1}^{\lfloor Nr\rfloor}\xi_s(\theta_{s-1}^*),$$
$$I_2(r) = \frac{1}{\sqrt{N}}\sum_{s=1}^{\lfloor Nr\rfloor}w_s^{\lfloor Nr\rfloor}\xi_s(\theta_{s-1}^*),$$

where $\alpha_s^t = \gamma_s\sum_{i=s}^t\prod_{k=s+1}^i(I_d - \gamma_k\bar{\mathcal{H}}_n)$ and $w_s^t = \bar{\mathcal{H}}_n^{-1} - \alpha_s^t$ for $1 \le s \le t$. Under Assumption 6, we have $\sup_{s\le t}\|\alpha_s^t\| \le C$ and $\sum_{s=1}^t\|w_s^t\| \le Ct^a$ for all $t \ge 1$ as in the proof of Lemma 6.

Since the event E_n is assumed to hold, $\sup_{r\in[0,1]}\|I_0(r)\| \xrightarrow{\mathbb{P}^*} 0$ and $\sup_{r\in[0,1]}\|I_2(r)\| \xrightarrow{\mathbb{P}^*} 0$ continue to hold.

To establish the FCLT for $I_1(r)$, we need to verify the following conditions: as $N \to \infty$,

$$\frac{1}{N}\sum_{s=1}^N\mathbb{E}_n^*\left[\|\xi_s(\theta_{s-1}^*)\|^2\mathbb{1}\left\{\|\xi_s(\theta_{s-1}^*)\|^2 > N\delta\right\}\middle|\mathcal{F}_{n,s-1}^*\right] \to 0 \quad \text{for any } \delta > 0, \tag{45}$$

$$\frac{1}{N}\sum_{s=1}^{\lfloor Nr\rfloor}\mathbb{E}_n^*\left[\xi_s(\theta_{s-1}^*)\xi_s(\theta_{s-1}^*)'\middle|\mathcal{F}_{n,s-1}^*\right] \to r\Sigma_n \quad \text{for any } r \in [0,1], \tag{46}$$

where Σ_o is replaced by its sample analogue Σ_n relative to (43). Proofs of (45) and (46) derive from the same arguments as in the proof of Lemma 6, whose details are omitted to avoid repetition. As a result, $I_1(r) \rightsquigarrow (\bar{\mathcal{H}}_n^{-1}\Sigma_n\bar{\mathcal{H}}_n^{-1})^{1/2}W(r)$ as $N \to \infty$. Putting it all together, we obtain, as $N \to \infty$,

$$\left\{\frac{1}{\sqrt{N}}\sum_{t=1}^{\lfloor Nr\rfloor}(\theta_t^\circ - \hat{\theta}_n)\right\}_{r\in[0,1]} \rightsquigarrow \left(\bar{\mathcal{H}}_n^{-1}\Sigma_n\bar{\mathcal{H}}_n^{-1}\right)^{1/2}W(r).$$