Noisy, Non-Smooth, Non-Convex Estimation of Moment Condition Models

Jean-Jacques Forneron*

October 20, 2022

Abstract

A practical challenge in structural estimation is the requirement to minimize a sample objective function which is often non-smooth, non-convex, or both. This paper proposes a simple algorithm designed to find accurate solutions without performing an exhaustive search. It augments each iteration from a new Gauss-Newton algorithm with a grid search step. A finite sample analysis derives its optimization and statistical properties simultaneously under standard econometric assumptions. After a finite number of iterations, the algorithm transitions from global to fast local convergence, producing accurate estimates with high-probability. Simulated examples and an empirical application illustrate the properties and performance of the algorithm. Comparisons with commonly used optimizers and quasi-Bayesian estimation using MCMC are also given.

JEL Classification: C11, C12, C13, C32, C36.
Keywords: Generalized and Simulated Method of Moments, Non-Asymptotic bounds.

*Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA.
Email: jjmf@bu.edu, Website: http://jjforneron.com.
Part of this research was conducted while I was visiting Yale University, I am grateful for their hospitality. I would like to thank David Lagakos and Marc Rysman for their inquiries about non-linear estimation which motivated this research, as well as Xiaohong Chen, Yuichi Kitamura, Jessie Li, and Liang Zhong for helpful comments and suggestions, as well as participants at the CEME Conference at Duke, BU/BC, and Yale econometric workshops.

1 Introduction

Generalized method of moments (GMM) estimations are conceptually attractive as they enable parameter estimation without a fully specified model. It is, however, well known that GMM estimations can be challenging to implement because they entail solving non-trivial minimization problems. This is a serious concern because failed estimations can lead to very inaccurate predictions about causal effects, and welfare implications of couterfactual policies. The theoretical challenges in computing the GMM estimator are summarized in Andrews (1997, Section 2.1), and the practical implications are illustrated in Knittel and Metaxoglou (2014, Sections VI-VII) in the context of demand estimation and merger analyses.

Depending on the model and the choice of moments, the objective function may have multiple local minima. These attract local optimizers which return inconsistent estimates. It can also be non-smooth when the moments are simulated, or when the policy function is approximated numerically; sometimes after discretizing the model. Finally, even though the population problem is smooth and locally convex – in finite samples, sampling and simulation noise can make the problem non-convex, even locally, making some optimizers unstable. Stochastic gradient-descent methods have gained popularity in recent years; however, global convergence - a necessary circumscription for valid inferences - is only guaranteed for convex problems. This explains why they have yet to permeate to the estimations considered here.

This paper tackles these estimation challenges in two steps. First, it proposes a new Gauss-Newton algorithm that combines non-smooth moments with smoothed Jacobian estimates. The algorithm is shown to converge quickly, in an area that is nearly as large as in the smooth population problem. The resulting estimates are asymptotically unbiased under mild conditions on the bandwidth. Unlike existing methods that rely on smoothing, undersmoothing is not required here, making the estimates more robust to the choice of bandwidth. The local convergence analysis relies on standard local identification conditions.

Second, the paper shows that augmenting each Gauss-Newton iteration with a grid-search steps results in a globally convergent algorithm. After a finite number of iterations, the combined steps preserve the fast local convergence rate, albeit with a slightly slower rate. This contrasts with commonly used global optimizers that converge slowly. Note that, here, the Algorithm transitions from global to local optimization without user input. The global convergence results leverage standard global identification conditions.[1]

[1]While the main focus of this paper is GMM estimations, the procedure used to make local iterations globally convergent can be applied to other settings, including non-linear least-squares and maximum likelihood estimations.

This paper adopts a non-asymptotic (finite sample) approach to study the optimization and statistical properties of the algorithm simultaneously. This allows to better understand the effects that tuning parameters can have on both the estimation and the estimates. For instance, smoothing should facilitate optimization but could also degrade the properties of the estimator. Here, the derivations show how both are affected by the choice of smoothing parameters. The finite-sample approach is also conceptually attractive because it gives results that are valid across repeated samples. Textbook optimization results often consider fixed objective functions (Nocedal and Wright, 2006; Bertsekas, 2016). However, in empirical work, each dataset is associated with its own, unique, objective function. In that sense, guarantees over repeated samples provide a sense of robustness against sampling noise.

The first extension of the main results considers momentum, also known as the Polyak heavy-ball, which is often used to the accelerate convergence of stochastic gradient-descent algorithms. Here, it can be used to reduce the effect of sampling noise on optimization while maintaining the rate of convergence. The second extension proposes a computationally attractive Monte-Carlo quasi-Newton implementation of the Algorithm which makes it applicable to models where smoothing is not analytically tractable.

The properties of the algorithm are illustrated using simulated and empirical examples. A dynamic discrete choice model illustrates the algorithm's advantages over a benchmark which minimizes a smoothed objective function. From a distant starting value, the proposed algorithm performs well whilst the benchmark systematically fails to find consistent estimates. When initialized at the true value, the benchmark is consistent but sensitive to the choice of smoothing parameter. In contrast, the algorithm is fairly robust to the smoothing parameter. This is in line with the theoretical results. A simple Aiyagari heterogenous agent model illustrates the optimization properties on a small but fairly challenging estimation problem where commonly used optimizers fail to converge. The algorithm performs well for a range of tuning parameters. Finally, the empirical application considers a model of joint retirement decision from Honoré and de Paula (2018). Their estimation of 30 coefficients takes more than 5 hours whereas the Algorithm introduced in this paper finds accurate estimates within 11 minutes, using the same code and a more distant starting value.

Outline of the paper. After a brief literate review, Section 2 briefly describes the setting and introduces the Algorithm with a simple illustration. Section 3 provides the main assumptions, then derives local and global convergence results with infinite and finite samples. These population results provide a benchmark for the finite-sample ones. Section 4 provides

two extensions of the Algorithm and the results. Section 5 provides simulated and empirical applications to illustrate the properties and performance of the method. The Appendices provide proofs as well as additional simulation and empirical results.

Overview of the Problem and Litterature

Optimization is a defining feature of non-linear models evaluated by M-estimation:

$$\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} Q_n(\theta),$$

where Q_n is the sample objective function, θ are parameters of interest belonging to a compact, finite-dimensional set Θ. While some estimators have closed-form solution, this is the case of OLS regressions, non-linear models generally require numerical methods to find $\hat{\theta}_n$. Modern software provide empirical researchers with a wide array of optimizers.

When choosing among these methods, two properties are important to consider: convergence guarantees, typically over a class of functions, and convergence rates within that class. The former matters since a failed optimization yields inconsistent estimates, the latter is critical when Q_n is costly to evaluate. The following expands on these two properties. The goal here is to find, after $b \geq 1$ iterations, a θ_b such that:

$$\|\theta_b - \hat{\theta}_n\| \leq \text{err},$$

at a desired tolerance level $\text{err} > 0$. Suppose Q_n has a certain set of features, one could ask: how large does b need to be to guarantee the error $\|\theta_b - \hat{\theta}_n\|$ is at most err? This is the complexity of the optimization problem (Nemirovsky and Yudin, 1983). If Q_n is smooth and strongly convex,[2] then $b \geq O(|\log(\text{err})|)$ so that to divide err by 10 an additional $O(\log[10])$ iterations are required. Gradient-descent, and (quasi)-Newton methods can achieve this rate of convergence. However when Q_n is non-convex, they may only converge to a local optimum. If Q_n is r-times continuously differentiable but otherwise arbitrary, then $b \geq O(\text{err}^{-d_\theta/r})$ iterations are required, where $d_\theta = \dim(\theta)$. For continuous problems, $r = 1$, this rate is achieved by a grid search. Here, the curse of dimensionality is apparent: to divide err by 10, the number of iterations needs to be multiplied by $10^{d_\theta} = 10,000$ for $d_\theta = 4$. Nesterov (2018, p14) and Andrews (1997, p915) illustrate that as d_θ increases, optimization quickly is practically infeasible under this rate of convergence. Computer software provides an armada

[2] Q_n is strongly convex if its Hessian H_n is continuous and $0 < \underline{\lambda} \leq \lambda_{\min}[H_n(\theta)] \leq \lambda_{\max}[H_n(\theta)] \leq \overline{\lambda} < \infty$.

of global optimizers, however, as Griewank (1981, Sec1) points out, those that do not cover the set Θ are heuristic, in the sense that convergence is not guaranteed.

Since many GMM estimations are non-convex, a number of authors have proposed strategies to speed-up convergence. Building on Robinson (1988), Andrews (1997) proposed a stopping-rule to produce a consistent first-step estimate, with err $= o_p(1)$, which is followed by a single Newton-Raphson iteration. The main challenge is in computing the first estimate which still requires a global search. Chernozhukov and Hong (2003) introduce a quasi-Bayesian framework where MCMC sampling replaces optimization. The random-walk Metropolis-Hastings algorithm converges under weak conditions. However, convergence rates, which measure performance, are mostly derived for log-concave distributions.[3]

A companion paper, Zhong and Forneron (2022) gives a sufficient rank condition for the Gauss-Newton algorithm to be globally convergent in smooth GMM estimations, at rates similar to the convex case. However, the assumption does exclude local optima, requires adequate tuning, and smooth sample moments. This paper adds a global search step to make the algorithm more robust: it is also convergent without that rank condition.

A common strategy to handle non-smooth sample moments is to use smoothing. The complexity bounds above then apply to the smoothed optimization problem. McFadden (1989, pp1000-1001) and Bruins et al. (2018) suggest to replace an indicator function by a smooth approximation in simulation-based estimation of discrete choice models. Several papers consider smoothing in quantile regressions (Kaplan and Sun, 2017; Fernandes et al., 2021; He et al., 2021). A poor choice of smoothing parameter can have a significant impact on the estimates and their statistical properties (see Kaplan and Sun, 2017, Table 1, p127). Here, the local search step involves unsmoothed moments making the estimates more robust to the choice of bandwidth. This is illustrated in the next Section. Also, the new quasi-Newton Monte-Carlo algorithm allows to consider models where smoothing is not tractable.

2 Noisy, Non-Smooth, Non-Convex Estimation

Let $\overline{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^n g(\theta; x_i)$ be the sample moments, where x_i are iid random variables and θ are the parameters of interest. The goal is to find an estimate of $\theta^\dagger \in \Theta$ solving the moment condition $g(\theta^\dagger) := \mathbb{E}[\overline{g}_n(\theta^\dagger)] = 0$. In practice, this entails finding an approximate minimizer

[3]See e.g. Mengersen and Tweedie (1996), Brooks (1998), Belloni and Chernozhukov (2009).

$\hat{\theta}_n \in \Theta$ of the sample objective function $Q_n(\theta) = \|\overline{g}_n(\theta)\|^2_{W_n} = \overline{g}_n(\theta)' W_n \overline{g}_n(\theta)$ satisfying:

$$\|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n} \leq \inf_{\theta \in \Theta} \|\overline{g}_n(\theta)\|^2_{W_n} + o_p(n^{-1}),$$

where $(W_n)_{n \geq 1}$ is a sequence of symmetric and strictly positive definite weighting matrices with limit $W > 0$. To simplify the analysis, W_n is assumed not to depend on θ. Algorithm 1 is designed to find such an approximate minimizer.

Algorithm 1 Smoothed Gauss-Newton (sGN)

1) **Inputs** (a) a learning rate $\gamma \in (0, 1)$, (b) a smoothing parameter $\varepsilon > 0$, (c) a weighting matrix W_n, and (d) a sequence $(\theta^b)_{b \geq 0}$ covering the parameter space Θ.
2) **Iterations**
set $b = 0$, $\theta_0 = \theta^0$
repeat
compute $\theta_{b+1} = \theta_b - \gamma \Big[G_{n,\varepsilon}(\theta_b)' W_n G_{n,\varepsilon}(\theta_b) \Big]^{-1} G_{n,\varepsilon}(\theta_b)' W_n \overline{g}_n(\theta_b)$ ▷ Local Step
if $\|\overline{g}_n(\theta^{b+1})\|_{W_n} < \|\overline{g}_n(\theta_{b+1})\|_{W_n}$, set $\theta_{b+1} = \theta^{b+1}$ ▷ Global Step
increment $b := b + 1$
until a stopping critera is met
3) **Output** $\hat{\theta}_n = \text{argmin}_{0 \leq j \leq b} \|\overline{g}_n(\theta_j)\|_{W_n}$

Algorithm 1 implements Gauss-Newton iterations using a smoothed Jacobian $G_{n,\varepsilon}$, and augments them with a global search step. As discussed earlier, Gauss-Newton iterations can be globally convergent without convexity (Zhong and Forneron, 2022), the global step ensures that convergence will hold over a broader class of problems. A common approach to find a global solution is the multi-start strategy which runs a local optimizer many times from different starting values. This paper simply adds a grid search step for global convergence in a single run of the Algorithm. As discussed earlier, covering the set Θ is required to guaranteed convergence. The differences with multi-start are discussed in Appendix H.

The algorithm has three tuning parameters: the learning rate γ, the bandwidth ε, and the covering sequence $(\theta^b)_{b \geq 0}$. A benefit of the Gauss-Newton normalization $[G_{n,\varepsilon}(\theta_b)' W_n G_{n,\varepsilon}(\theta_b)]^{-1}$ is that the theory allows for any choice of $\gamma \in (0, 1)$. In contrast, gradient-descent methods require $\gamma > 0$ to be sufficiently small for convergence. While larger values are associated with faster convergence; smaller values, e.g. $\gamma \in [0.1, 0.3]$, make the local step less sensitive to sampling noise. The smoothing parameter ε should satisfy $\varepsilon = o(1)$ and $\sqrt{n}\varepsilon \to \infty$. For optimization $\varepsilon = O(n^{-1/4})$ is optimal, this implies that a "large bandwidth" is preferred. This contrasts with other methods that require undersmoothing, $\varepsilon = o(n^{-1/4})$, to avoid asymptotic bias. Choices of covering sequences $(\theta^b)_{b \geq 0}$ will be discussed in Section 3.2.

The Jacobian $G_{n,\varepsilon}$ is computed by convolution smoothing. Let ϕ be the standard Gaussian density. The smoothed moments are:

$$\overline{g}_{n,\varepsilon}(\theta) := \mathbb{E}_Z[\overline{g}_n(\theta + \varepsilon Z)] = \int \overline{g}_n(\theta + \varepsilon Z)\phi(Z)dZ,$$

and the corresponding smoothed Jacobian matrix is:[4]

$$G_{n,\varepsilon}(\theta) := \partial_\theta \overline{g}_{n,\varepsilon}(\theta) = \frac{1}{\varepsilon}\int \overline{g}_{n,\varepsilon}(\theta + \varepsilon Z)Z'\phi(Z)dZ.$$

While in theory the smoothing can be applied to any estimation problem, $\overline{g}_{n,\varepsilon}$ and $G_{n,\varepsilon}$ typically do not have closed-form. This is the case for the applications in Section 5. Nevertheless, an unbiased Monte-Carlo estimator of $G_{n,\varepsilon}$ can be computed as follows:

$$\hat{G}_{n,\varepsilon}(\theta) = \frac{1}{\varepsilon L}\sum_{\ell=0}^{L-1}\left[\overline{g}_{n,\varepsilon}(\theta + \varepsilon Z_\ell) - \overline{g}_{n,\varepsilon}(\theta)\right]Z_\ell', \quad Z_\ell \overset{iid}{\sim} \mathcal{N}(0, I_{d_\theta}),$$

where $L \geq 1$. The mean-zero adjustment $\overline{g}_{n,\varepsilon}(\theta)$ yields better finite-L properties. A similar estimator was proposed in Polyak (1987), Polyak and Tsybakov (1990) for minimizing smooth and globally convex objectives by stochastic gradient-descent (sGD), and was later studied in Nesterov and Spokoiny (2017). Convergence for gradient-descent is typically much slower than Gauss-Newton. Chen and Liao (2015) proposed a similar Jacobian for sieves. In an extension of the main results, a computationally attractive quasi-Newton implementation of $\hat{G}_{n,\varepsilon}$ is introduced. Its non-asymptotic local convergence properties are derived. All simulated and empirical examples in the paper are based on this implementation.

Stopping Criteria. The following describes a simple approach to determine when to terminate the optimization. Following Andrews (1997), let $k \geq 0$ be such that $\|\overline{g}_n(\theta_k)\|^2_{W_n} \leq c_{p,n}/n$, where $p = \dim(\overline{g}_n)$ and $c_{p,n} > 0$ is a threshold. If W_n is the optimal weighting matrix, $c_{p,n}$ can be a quantile of the χ^2_p distribution as in Andrews (1997, Table I, p921). Then, k is the first iteration when an Anderson-Rubin test does not reject $H_0 : \theta^\dagger = \theta_k$ at some level.

Since θ_k is local to $\hat{\theta}_n$, the rate of convergence for optimization is of approximately $(1-\overline{\gamma})$ with $\overline{\gamma} \simeq \gamma$. After another $j \geq 0$ iterations, we have $b = k + j$ and the estimation error has been reduced by a factor of $(1-\overline{\gamma})^j$ compared to θ_k. This holds regardless of the dimension d_θ;

[4]Other choices densities ϕ could be used as long as they are: 1) continuously differentiable, and 2) satisfy $\int Z\phi(Z)dZ = 0$, $\int \|Z\|^2\phi(Z)dZ < \infty$, $\int \|\partial_Z\phi(Z)\|dZ < \infty$. The last equality can be derived using a change of variable argument similar to Powell et al. (1989) or Hazan et al. (2016). See Lemma A2 for several useful identities for implementation and derivations.

a feature of Gauss-Newton. Pick $\gamma = 0.1$ and $j = 45$, then the estimation error for θ_b is only $\simeq 1\%$ of θ_k's estimation error. Now pick the same $\gamma = 0.1$ and use momentum (described in an extension of the main results), with the parameters in Table 2, the estimation error for θ_b declines to $\simeq 10^{-6}\%$ of θ_k's estimation error, for the same $j = 45$. In the applications, using $200 \leq b_{\max} \leq 250$ with $\gamma = 0.1$ and momentum yielded good results.

Intuition for the Local Step. Figure 1 illustrates the main ideas behind the Gauss-Newton (GN) iterations in a stylized scalar example. Panel a) shows the population objective $Q(\theta)$ and the moments $g(\theta)$, minimizing Q is equivalent to finding $g(\theta^\dagger) = 0$; here $\theta^\dagger = 1$. Gradient-descent (GD) relies on Q (left panel) being decreasing when $\theta < 1$ and increasing when $\theta > 1$. GD iterations produce $\theta_{b+1} > \theta_b$ for $\theta_b < 1$, and $\theta_{b+1} < \theta_b$ for $\theta_b > 1$. In comparison, Gauss-Newton (GN) iterations rely on moments (middle panel) which are positive and decreasing for $\theta < 1$; positive and decreasing for $\theta > 1$. Much like GD, GN iterations produce $\theta_{b+1} > \theta_b$ for $\theta_b < 1$, and $\theta_{b+1} < \theta_b$ for $\theta_b > 1$.

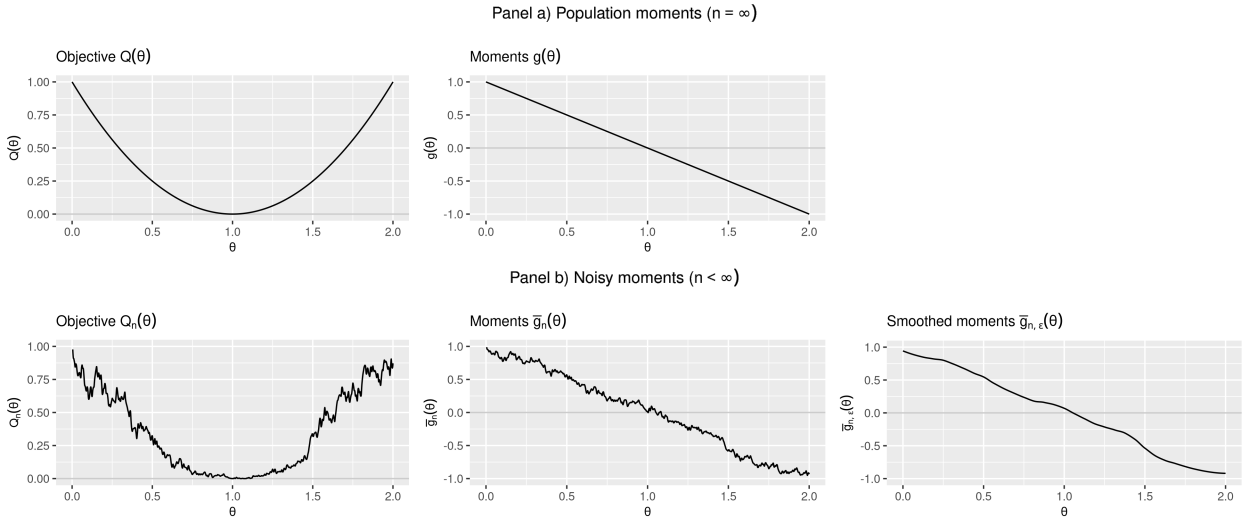
Panel b) shows a sample objective $Q_n(\theta)$ and moments $\overline{g}_n(\theta)$. Notice that Q_n has many local minima on both sides of $\theta = 1$ (left panel), making optimization difficult with GD. GNuses the information differently: $\overline{g}_n$ is strictly positive when $\theta < \hat{\theta}_n$, strictly negative for $\theta > \hat{\theta}_n$. Also, the smoothed $\overline{g}_{n,\varepsilon}$ is decreasing. Qualitatively, this is similar to the well behaved Panel a). Hence, sGN (Algorithm 1) should set $\theta_{b+1} > \theta_b$ when $\theta_b < \hat{\theta}_n$, and $\theta_{b+1} < \theta_b$ when $\theta_b > \hat{\theta}_n$.

The main idea here is that $Q_n(\theta) = \|\overline{g}_n(\theta)\|^2_{W_n}$ does not contain information about the sign and monotonicity of $\overline{g}_n(\theta)$. In contrast, sGN iterations explicitly use this information as illustrated here. In practice, the local step alone may not be sufficient to ensure global convergence. The global step incorporates additional information from Q_n (left panel).

A pen and pencil example. The following illustrates the differences between the local step in Algorithm 1 and smoothing the objective function. Consider estimating the t-th quantile of $x_i \overset{iid}{\sim} F$, $t \in (0, 1)$. The population and sample moments are $F(\theta) - t = 0$, and:

$$\overline{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} [\mathbb{1}_{x_i \leq \theta} - t] = 0,$$

Figure 1: Illustration: Population and Sample Objective, Moments



Legend: Standard optimizers only use information from Q_n in the left panel. Local sGN iterations use information from $\overline{g}_n$ and $\overline{g}_{n,\varepsilon}$ in the middle and right panels, the global step adds information about Q_n.

or more compactly $F_n(\theta) - t = 0$; where F_n is the empirical CDF of x_i. For any choice of bandwidth ε, the smoothed moment condition and its Jacobian are:[5]

$$\overline{g}_{n,\varepsilon}(\theta) = \frac{1}{n}\left[\sum_{i=1}^{n} \Phi\left(\frac{\theta - x_i}{\varepsilon}\right) - t\right], \quad G_{n,\varepsilon}(\theta) = \frac{1}{n\varepsilon}\sum_{i=1}^{n} \phi\left(\frac{\theta - x_i}{\varepsilon}\right),$$

where Φ and ϕ are the standard Gaussian CDF and density. Notice that $G_{n,\varepsilon}(\theta) = f_{n,\varepsilon}$ is the kernel density estimator of the density f. In this example, the local step in Algorithm 1 involves the non-smooth CDF F_n and the smoothed density $f_{n,\varepsilon}$:

$$\theta_{b+1} = \theta_b - \gamma \frac{F_n(\theta_b) - t}{f_{n,\varepsilon}(\theta_b)}.$$

The main appeal is that an exact solution $\hat{\theta}_n$ with $F_n(\hat{\theta}_n) - t = 0$ is a fixed-point of these iterations, regardless of the bandwidth ε. This implies that *if the Algorithm is convergent*, then the solution is robust to the choice of bandwidth.

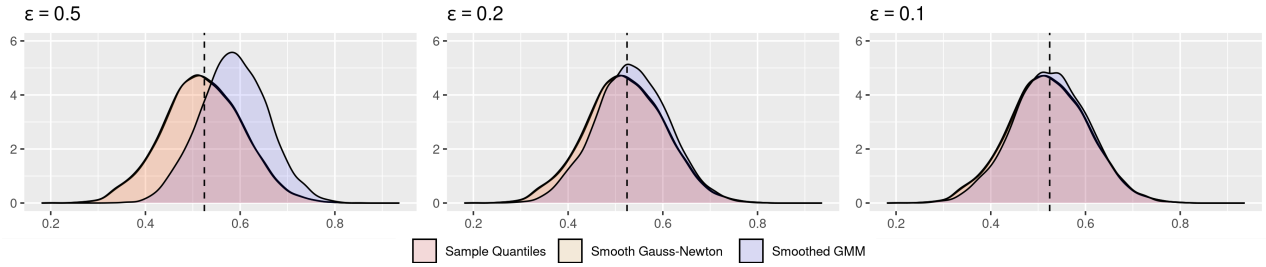
Applying Gauss-Newton to the smoothed moments further substitutes $\overline{g}_n$ for $\overline{g}_{n,\varepsilon}(\theta) = F_{n,\varepsilon}(\theta) - t$, $F_{n,\varepsilon}(\theta) = \int_{-\infty}^{\theta} f_{n,\varepsilon}(u)du$ is the smoothed CDF. This results in the iterations:

$$\theta_{b+1} = \theta_b - \gamma \frac{F_{n,\varepsilon}(\theta_b) - t}{f_{n,\varepsilon}(\theta_b)}.$$

[5]This is derived using $\mathbb{E}_Z(\mathbb{1}_{x_i \leq \theta + \varepsilon Z}) = \Phi([\theta - x_i]/\varepsilon)$.

Here, a fixed-point $\hat{\theta}_{n,\varepsilon}$ must satisfy $F_{n,\varepsilon}(\hat{\theta}_{n,\varepsilon}) - t = 0$. This is the t-th quantile of $F_{n,\varepsilon}$ which satisfies $\hat{\theta}_{n,\varepsilon} = \hat{\theta}_n + O_p(\varepsilon^2)$. The smoothing bias, captured by ε^2, is only negligible when $\varepsilon = o(n^{-1/4})$ which is more restrictive. Figure 2 illustrates this key difference. It compares

Figure 2: Illustration: estimating a sample quantile



Legend: sample quantile and SGN are nearly indistinguishable for all $\varepsilon \in \{0.5, 0.2, 0.1\}$.
$x_i \sim \mathcal{N}(0,1)$, $n = 250$, $t = 0.7$.

estimates based on R's *quantile* function[6] with both methods. Estimates from Algorithm 1 and the sample quantiles are nearly identical, for all ε. The blue curve corresponds to $\hat{\theta}_{n,\varepsilon}$, its bias is visible when $\varepsilon = 0.5, 0.2$. Since $n^{-1/4} = 0.25$, the choice $\varepsilon = 0.1$ represents the undersmoothing regime. Table 1 further illustrates the bias of the estimates and its implications for inference. Inferences using $\hat{\theta}_{n,\varepsilon}$ are size distorted compared to sample quantiles and Algorithm 1. Another example in Section 5 further compares the optimization and statistical properties of the two approaches.

Table 1: Illustration: estimating a sample quantile

		$\varepsilon = 0.5$			$\varepsilon = 0.2$			$\varepsilon = 0.1$		
	$\theta^\dagger$	$\hat{\theta}_n$	SGN	SGMM	$\hat{\theta}_n$	SGN	SGMM	$\hat{\theta}_n$	SGN	SGMM
avg	0.524	0.520	0.518	0.584	0.520	0.518	0.533	0.520	0.518	0.525
std	-	0.085	0.085	0.072	0.085	0.085	0.078	0.085	0.085	0.081
size	-	0.037	0.038	0.043	0.055	0.058	0.036	0.065	0.068	0.054

Legend: standard errors computed using $\sqrt{t(1-t)/n}/f_{n,\varepsilon}(\hat{\theta}_n)$ using the corresponding $\varepsilon \in \{0.5, 0.2, 0.1\}$, size reported for 5% significance level. SGN only uses the local step.

In this example, smoothing replaces the indicator function with the same $\Phi(\cdot/\varepsilon)$ for all i. This is generally not the case. Consider a quantile regression with moments $\overline{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^n [\mathbb{1}(y_i - x_i'\theta) - t]w_i$, w_i are the instruments. The smoothed moments are $\overline{g}_{n,\varepsilon}(\theta) = \frac{1}{n}\sum_{i=1}^n [\Phi(\frac{y_i - x_i'\theta}{\varepsilon\|x_i\|_2}) - t]w_i$. Here, the bandwidth is effectively $\varepsilon\|x_i\|_2$ which is individual specific.

[6]The sample quantile function in R returns an estimate obtained by a linear interpolation between order statistics, see e.g. Hyndman and Fan (1996) for details. As such, it does not exactly solve the moment condition $\overline{g}_n(\theta) = 0$, and its finite sample properties can differ slightly from the GMM estimator.

Similarly, the smoothed Jacobian is $G_{n,\varepsilon}(\theta) = \frac{-1}{n\varepsilon}\sum_{i=1}^{n} \frac{x_i'}{\|x_i\|_2}\phi(\frac{y_i - x_i'\theta}{\varepsilon\|x_i\|_2})w_i$ which, compared to smoothing the indicator function directly, has the additional $\|x_i\|_2$ term in the denominator. Notice that all moments of $x_i/\|x_i\|_2$ exist even when x_i has none.

3 Main Results

This section gives the main assumptions and derives the main results. First, the smooth population setting is considered to provide a benchmark for the finite sample problem. Because the global step in Algorithm 1 is new, the derivations for this benchmark will provide intuition in a simpler setting. Choices of covering sequences $(\theta^b)_{b\geq 0}$ are then discussed. Then, the finite sample convergence results are derived and compared with the benchmark.

3.1 Assumptions

Assumption 1. *Suppose the parameter space, moments, and weighting matrix satisfy: i.* $\theta^\dagger \in int(\Theta)$, $\Theta \subset \mathbb{R}^{d_\theta}$ *is compact and convex, ii.* $g(\cdot) = \mathbb{E}[\overline{g}_n(\cdot)]$ *is continuously differentiable on* Θ*, iii. the Jacobian* $G(\theta^\dagger) := \partial_\theta g(\theta^\dagger)$ *has full rank, iv. there exists* $L_G \geq 0$ *such that for any* $(\theta_1, \theta_2) \in \Theta^2$, $\|G(\theta_1) - G(\theta_2)\| \leq L_G\|\theta_1 - \theta_2\|$*, v. for all* $\eta > 0$*, there exists* $\delta(\eta) > 0$ *such that* $\inf_{\|\theta-\theta^\dagger\|\geq\eta} \|g(\theta)\|_W \geq \delta(\eta)$*, where* $\delta(\cdot)$ *is a continuous and weakly decreasing function, vi. there exists* $0 < \underline{\lambda}_W \leq \overline{\lambda}_W < \infty$ *such that* $\underline{\lambda}_W \leq \lambda_{\min}(W) \leq \lambda_{\max}(W) \leq \overline{\lambda}_W$*.*

Assumption 1 are the conditions required for the population quantities. Conditions ii, iii, vi imply local strong convexity of $Q(\theta) = \|g(\theta)\|_W^2$ around the solution $\theta^\dagger$.

Assumption 2. *Suppose the sample moments and weighting matrix are such that: i. for all* $\theta \in \Theta$, $\mathbb{E}(\|g(\theta; x_i)\|^2) < \infty$*, ii. for some* $L_g \geq 0$, $\psi \in (0, 1]$*, and any* $\delta > 0$, $[\mathbb{E}(\sup_{\|\theta_1-\theta_2\|\leq\delta} \|g(\theta_1; x_i) - g(\theta_2; x_i)\|^2)]^{1/2} \leq L_g\delta^\psi$*, iii.* $W_n \xrightarrow{p} W$*, there exists* $0 < \underline{\lambda}_W \leq \overline{\lambda}_W < \infty$ *such that* $\underline{\lambda}_W \leq \lambda_{\min}(W_n) \leq \lambda_{\max}(W_n) \leq \overline{\lambda}_W$*, with probability 1.*

Assumption 2 gives additional conditions for the sample moments. The assumptions allow for a variety of estimation problems, including simulated method of moments estimations with non-smooth moments. The data are assumed iid to apply concentration inequalities, mainly van der Vaart and Wellner (1996, Th2.14.2). Similar results exist under dependence (e.g. Dedecker and Louhichi, 2002, Sec4.3), so that the results could be extended to time-series data under mixing conditions. Under Assumptions 1-2, it can be shown that the approximate minimizer $\hat{\theta}_n$ is $\sqrt{n}$-consistent and asymptotically Gaussian:

$$\sqrt{n}(\hat{\theta}_n - \theta^\dagger) \xrightarrow{d} \mathcal{N}(0, V),$$

where V is the asymptotic variance of $\hat{\theta}_n$, see e.g. Newey and McFadden (1994, Th3.4).

3.2 Benchmark: the Population Problem

First, Algorithm 1 is applied to the population objective $\|g(\theta)\|_W^2$ without smoothing.

Local Convergence. Take a starting value $\theta_0 \in \Theta$ and iterate:

$$\theta_{b+1} = \theta_b - \gamma(G_b' W G_b)^{-1} G_b' W g(\theta_b), \quad b = 0, 1, \ldots \tag{1}$$

where $G_b = G(\theta_b)$. The following Lemma give the local convergence of θ_b.

Lemma 1. *Suppose Assumption 1 i-iv, vi hold. Let* $0 < \underline{\sigma} < \sigma_{\min}[G(\theta^\dagger)]$, *and* $R_G > 0$ *be such that* $\|\theta - \theta^\dagger\| \leq R_G \Rightarrow \underline{\sigma} \leq \sigma_{\min}[G(\theta)]$. *Pick any* $\gamma \in (0,1)$ *and* $\overline{\gamma} \in (0, \gamma)$, *then for any* $\theta_0 \in \Theta$ *such that:* $\|\theta_0 - \theta^\dagger\| \leq R := \min\left(R_G, [\gamma - \overline{\gamma}]\frac{\underline{\sigma}}{\gamma L_G \sqrt{\kappa_W}}\right)$, *we have:*

$$\|\theta_b - \theta^\dagger\| \leq (1 - \overline{\gamma})^b \|\theta_0 - \theta^\dagger\|.$$

Lemma 1 states that, for any choice of learning rate $\gamma \in (0,1)$, the iterations from (1) converge at a $(1-\overline{\gamma})$ rate to $\theta^\dagger$, for a suitable choice of starting value $\|\theta_0 - \theta^\dagger\| \leq R$. A large value of $\overline{\gamma} < \gamma$ requires a smaller R, convergence can be faster closer to $\theta^\dagger$. Optimization is easier when R is large, in the sense that an arbitrary θ_0 is more likely to result in convergence. The size of R depends on several factors. When $G(\theta^\dagger)$ is close to singular, $\underline{\sigma}$ is small and so is R. When $L_G = 0$, the moments are linear and convergence holds for any $\theta_0 \in \Theta$, this is the case in OLS regressions. Non-linear moments imply $L_G > 0$ and a finite R; a larger L_G implies a smaller R. The matrix W also matters. The quantity $\kappa_W = \overline{\lambda}_W / \underline{\lambda}_W \geq 1$ bounds its condition number. Under equal weighting, $W = I_d$, $\kappa_W = 1$ does not affect R. For other matrices W, $\kappa_W > 1$ is possible, which results in a smaller R. If W is ill-conditionned, κ_W is large so R is small and optimization is more difficult. This supports choices of W that are statistically inefficient but well conditionned. A short proof of Lemma 1 is given below.

Proof of Lemma 1. Take any θ_b such that $\|\theta_b - \theta^\dagger\| \leq R$ as defined in Lemma 1. Let $G_b = G(\theta_b)$. Since $g(\theta^\dagger) = 0$, the next θ_{b+1} is such that:

$$\begin{aligned}
\theta_{b+1} - \theta^\dagger &= \theta_b - \theta^\dagger - \gamma \left(G_b' W G_b\right)^{-1} G_b' W [g(\theta_b) - g(\theta^\dagger)] \\
&= (1-\gamma)(\theta_b - \theta^\dagger) - \gamma \left(G_b' W G_b\right)^{-1} G_b' W [G(\tilde{\theta}_b) - G_b](\theta_b - \theta^\dagger),
\end{aligned}$$

for some intermediate value $\tilde{\theta}_b$. Now we have, by Lipschitz-continuity of G:

$$\| \left(G_b' W G_b\right)^{-1} G_b' W [G(\tilde{\theta}_b) - G_b](\theta_b - \theta^\dagger)\| \leq \frac{\sqrt{\overline{\lambda}_W} L_G}{\sqrt{\lambda_{min}[G_b' W G_b]}} \|\theta_b - \theta^\dagger\|^2 \leq \frac{\sqrt{\kappa_W} L_G}{\underline{\sigma}} \|\theta_b - \theta^\dagger\|^2.$$

Taking this back into the previous equality, we have:

$$\|\theta_{b+1} - \theta^\dagger\| \leq \left(1 - \gamma + \gamma \frac{\sqrt{\kappa_W} L_G}{\underline{\sigma}} \|\theta_b - \theta^\dagger\|\right) \|\theta_b - \theta^\dagger\| \leq (1 - \overline{\gamma}) \|\theta_b - \theta^\dagger\|,$$

for $\|\theta_b - \theta^\dagger\| \leq R$. Iterating over b, we find: $\|\theta_{b+1} - \theta^\dagger\| \leq (1 - \overline{\gamma})^{b+1} \|\theta_0 - \theta^\dagger\|$. □

Global Convergence. For an arbitrary starting value θ_0, Lemma 1 does not guarantee convergence to $\theta^\dagger$. With a covering sequence $(\theta^b)_{b \geq 0}$, discussed below, the following augments the local search with a global grid search step:

$$\theta_{b+1} = \theta_b - \gamma (G_b' W G_b)^{-1} G_b' W g(\theta_b) \tag{1}$$

$$\text{if } \|g(\theta^{b+1})\|_W < \|g(\theta_{b+1})\|_W, \text{ set } \theta_{b+1} = \theta^{b+1}. \tag{2}$$

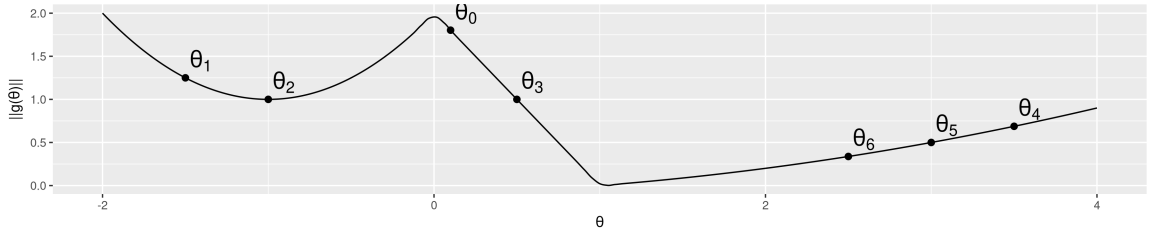
Step (2) resets the local search (1) when θ^{b+1} strictly reduces the objective function. The proof that the combined steps (1)-(2) yield a globally convergent sequence relies on three properties explained below and derived in the Appendix (Lemma A1).

First, locally to $\theta^\dagger$, the iteration (1) is a strict contraction for the objective function:

$$\|g(\theta_{b+1})\|_W \leq (1 - \overline{\gamma}) \|g(\theta_b)\|_W, \tag{3}$$

and $\|\theta_{b+1} - \theta^\dagger\| < \|\theta_b - \theta^\dagger\|$, cf. Lemma 1. However, if $\|g(\theta^{b+1})\|_W < (1 - \overline{\gamma}) \|g(\theta_b)\|_W$ at step (2), it is possible that $\|\theta^{b+1} - \theta^\dagger\| > R \geq \|\theta_b - \theta^\dagger\|$. When that is the case, the next iteration of (1) will be not locally convergent. Figure 3 illustrates this issue.

Figure 3: Illustration: a reduction of $\|g(\theta)\|$ does not imply that $\|\theta - \theta^\dagger\|$ also declines



Legend: stylized example, $\|g(\theta)\|$ decreases at each iteration but $\|\theta_b - \theta^\dagger\|$ increases at $b = 1, 4$.

This is the main challenge when proving global convergence. The key idea is to show that after a finite number of iterations, θ_b will be confined to a neighborhood of $\theta^\dagger$. Then,

within this neighborhood, step (2) preserves the local converge property of (1). These results are derived under the weighted norm:[7] $\|\theta - \theta^\dagger\|_{G'WG}$ which is closely related to $\|g(\theta)\|_W$ in step (2). More specifically, under Assumption 1, there exists a $\overline{r}_g > 0$ such that, for $\|\theta_b - \theta^\dagger\|_{G'WG} \leq \overline{r}_g$, the following norm equivalence holds:

$$(1 - \overline{\gamma}/2)\|\theta_b - \theta^\dagger\|_{G'WG} \leq \|g(\theta_b)\|_W \leq (1 + \overline{\gamma}/2)\|\theta_b - \theta^\dagger\|_{G'WG}, \tag{4}$$

and the contraction (3) above occurs. Under Assumption 1v, i.e. the global identification condition, there exists another $\underline{r}_g \in (0, \overline{r}_g]$ such that:[8]

$$\inf_{\|\theta - \theta^\dagger\|_{G'WG} \geq \overline{r}_g} \|g(\theta)\|_W \geq (1 + \overline{\gamma}/2)(1 - \overline{\gamma})\underline{r}_g. \tag{5}$$

Now, take $\|\theta_b - \theta^\dagger\|_{G'WG} \leq \underline{r}_g$, combine the contraction (3) and norm equivalence (4):

$$\|g(\theta_{b+1})\|_W \leq (1 + \overline{\gamma}/2)(1 - \overline{\gamma})\underline{r}_g,$$

after applying the local step (1). Suppose that step (2) finds $\|g(\theta^{b+1})\|_W < \|g(\theta_{b+1})\|_W$. Then $\|\theta^{b+1} - \theta^\dagger\|_{G'WG} \leq \overline{r}_g$ from (5). The local norm equivalence (4) applies and:

$$\|\theta^{b+1} - \theta^\dagger\|_{G'WG} < \frac{1 + \overline{\gamma}/2}{1 - \overline{\gamma}/2}(1 - \overline{\gamma})\underline{r}_g < \underline{r}_g, \tag{6}$$

which is a strict contraction. This implies local stability: $\|\theta_{b+j} - \theta^\dagger\| \leq \underline{r}_g$ for all $j \geq 1$. Now, the contraction (3) can be applied recursively with $j = 1, 2, \ldots$ This is sufficient to prove global convergence under the norm $\|\cdot\|_{G'WG}$.

Lemma 2. *Suppose Assumption 1 holds. Take $\gamma \in (0, 1)$ and $\overline{\gamma} \in (0, \gamma)$. Let $k \geq 0$ be such that:*

$$\sup_{\theta \in \Theta} \left(\inf_{0 \leq \ell \leq k} \|\theta^\ell - \theta\|_{G'WG} \right) \leq \underline{r}_g,$$

then for any $b = k + j$ with $j \geq 0$, we have:

$$\|\theta_b - \theta^\dagger\|_{G'WG} \leq (1 - \overline{\gamma})^j \frac{1 + \overline{\gamma}/2}{1 - \overline{\gamma}/2}\underline{r}_g.$$

Lemma 2 shows that after a finite number of steps k, (1)-(2) converges quickly to $\theta^\dagger$. What is particularly attractive in this result is the automatic transition from k global steps to the j local ones without user input. Because k only depends on $\underline{r}_g$ and $(\theta^\ell)_{\ell \geq 0}$, the Algorithm adapts to the complexity of the optimization problem in a single run.

[7] The weighted norm is defined as $\|\theta - \theta^\dagger\|^2_{G'WG} = (\theta - \theta^\dagger)'[G(\theta^\dagger)'WG(\theta^\dagger)](\theta - \theta^\dagger)$.

[8] Simply set $\underline{r}_g = [\inf_{\|\theta - \theta^\dagger\|_{G'WG} \geq \overline{r}_g} \|g(\theta)\|_W]/[(1 + \overline{\gamma}/2)(1 - \overline{\gamma})] > 0$, by assumption.

The global identification condition 1v. is key for the result. Without this assumption, the worst-case bound for minimization implies that finding $\|\theta_b - \theta^\dagger\| \leq \text{err}$ requires $b \geq O(\text{err}^{-d_\theta})$ iterations with approximation error $\text{err} > 0$. Lemma 2 implies that to attain the error err, $b \geq k + O(|\log[\text{err}]|)$ iterations are required. Here, only the finite $k \geq 0$ depends on d_θ. For well identified models, $k \ll \text{err}^{-d_\theta}$ implies significant improvements. However, poorly identified models could have $\delta(\cdot) \simeq \text{err}$ in condition 1v. and then $k = O(\text{err}^{-d_\theta})$ gives back the worst-case bound. A short proof of the Lemma is given below.

Proof of Lemma 2. Lemma A1 proves (3), (4), and (5) which are the building blocks of the result. After k iterations, with the choice of k defined in the Lemma, we must have:

$$\|\theta_k - \theta^\dagger\|_{G'WG} \leq \underline{r}_g,$$

because there is at least one θ^ℓ with $\ell \in \{0, \ldots, k\}$ such that this inequality holds and the strict contraction property (6) implies that after that ℓ, the distance cannot be greater than $\underline{r}_g$. The worst-case is $\ell = k$. Then, apply (6) from $b = k$ to $b = k + j$, to get that $\|\theta_{k+j} - \theta^\dagger\|_{G'WG} \leq \underline{r}_g$ for any $j \geq 0$. Now iterate over (3) to get $\|g(\theta_{b+j})\|_W \leq (1 - \overline{\gamma})^j \|g(\theta_b)\|_W \leq (1 - \overline{\gamma})^j (1 + \overline{\gamma}/2)\underline{r}_g$, for any $j \geq 0$. Then, (4) yields the result. □

Choices of covering sequences. A smaller value of k in Lemma 2 implies the transition to local convergence occurs sooner. This k depends on $\|g(\cdot)\|_W^2$, through $\underline{r}_g$, and the covering sequence $(\theta^\ell)_{\ell \geq 0}$. A lower bound on k can be derived using covering numbers arguments:

$$k \geq \underline{r}_g^{-d_\theta} \frac{\text{vol}(\Theta)}{\text{vol}(B)},$$

where vol is the volume in $\mathbb{R}^{d_\theta}$ and B is the unit ball. Suppose the parameter space Θ is a product space, $\Theta = [\underline{\theta}_1, \overline{\theta}_1] \times \cdots \times [\underline{\theta}_{d_\theta}, \overline{\theta}_{d_\theta}]$. One way to compare sequences is to consider their discrepancy D_{k-1} defined as:

$$D_{k-1} = \sup_{\theta \in \Theta} \left(\inf_{0 \leq \ell \leq k-1} \|\theta - \theta^\ell\| \right).$$

Ideally, D_{k-1} declines at, or close to, the k^{-1/d_θ} rate implied by covering. For product spaces, Theorem 3 in Niederreiter (1983) implies that low-discrepancy sequences satisfy:

$$D_{k-1} \leq O(\sqrt{d_\theta} \log(k) k^{-1/d_\theta}),$$

which is nearly the desired k^{-1/d_θ} rate. Examples include the Halton and Sobol sequences. The latter was used in the simulated and empirical applications. In comparison, uniform draws $\theta^j \overset{iid}{\sim} \mathcal{U}_\Theta$ have $D_{k-1} = O_p(k^{-1/[2d_\theta]})$ which converges more slowly.

3.3 Finite Sample Results

The results so far illustrate how convergence rates depend crucially on local and global identification conditions. The following extends the results to finite samples.

Local Convergence. The first set of results consider the local step only:

$$\theta_{b+1} = \theta_b - \gamma[G_b' W_n G_b]^{-1} G_b' W_n \overline{g}_n(\theta_n), \quad b = 0, 1, \ldots, \tag{1}$$

where now $G_b = G_{n,\varepsilon}(\theta_b)$. Convergence results are derived with respect to:

$$\hat{\theta}_n = \theta^\dagger - (G' W_n G)^{-1} G W_n \overline{g}_n(\theta^\dagger), \tag{7}$$

where $G = G(\theta^\dagger)$. In large samples, $\hat{\theta}_n$ is an approximate minimizer of $Q_n(\cdot) = \|\overline{g}_n(\cdot)\|^2_{W_n}$. Lemma 3 below measures the stability of (1) at $\theta = \hat{\theta}_n$.

Lemma 3. *Suppose Assumptions 1-2 hold, then for any $\varepsilon > 0$ and $c_n \geq 1$:*

$$\|G_{n,\varepsilon}(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n)\| \leq C_1 (c_n n^{-1/2})^{1+\psi} \left(1 + \frac{c_n n^{-1/2}}{\varepsilon} + \frac{\varepsilon}{(c_n n^{-1/2})^\psi}\right) := \Gamma_{n,\varepsilon},$$

with probability $1 - (1 + C)/c_n$ for some universal constant C, and C_1 which only depends on $p = dim(\overline{g}_n)$, the set Θ, $\Sigma = var[g(\theta^\dagger; x_i)]$, and the constants in Assumptions 1-2.

The term $\Gamma_{n,\varepsilon}$ measures how close $\hat{\theta}_n$ is to being a fixed-point of the smoothed Gauss-Newton iterations (1).[9] Here $c_n \geq 1$ controls the probability level for which the bound holds over repeated samples. It will be useful for understanding the relation between the probability of a successful optimization and the choice of tuning parameters.

In Lemma 1, each iteration strictly reduces the distance between θ_b and $\theta^\dagger$. This need not be the case in finite samples because the moments are noisy. The following Proposition shows that within a neighborhood of $\hat{\theta}_n$, each iteration is a strict contraction towards $\hat{\theta}_n$, with high probability, up to a term that is asymptotically negligible.

Proposition 1. *Suppose Assumptions 1-2 hold. Take $\gamma \in (0, 1)$ and $\overline{\gamma} \in (0, \gamma)$, $c_n \geq 1$ as in Lemma 3, and R_G as in Lemma 1. Uniformly in $\theta_b \in \Theta$ such that $\|\theta_b - \theta^\dagger\| \leq R_{n,G}$:*

$$\begin{aligned}\|\theta_{b+1} - \hat{\theta}_n\| \leq &\left(1 - \gamma + \gamma \underline{\sigma}_{n,\varepsilon}^{-1} \sqrt{\kappa_W} [L_G \|\theta_b - \hat{\theta}_n\| + M_{1,Z} \varepsilon]\right) \|\theta_b - \hat{\theta}_n\| \\ &+ \gamma \Delta_{n,\varepsilon}(\|\theta_b - \hat{\theta}_n\|)\end{aligned}$$

[9] In just-identified models for which an exact solution $\overline{g}_n(\hat{\theta}_n) = 0$ exists, it is possible to set $\Gamma_{n,\varepsilon} = 0$ by using the exact minimizer.

with probability $1-(1+C)/c_n$*, where* $R_{n,G} := R_G - C_a c_n n^{-1/2}$*,* $\underline{\sigma}_{n,\varepsilon} = \underline{\sigma} - C_\sigma[\frac{c_n n^{-1/2}}{\varepsilon} + \varepsilon]$*,* $M_{1,Z} = \int \|\phi'(Z)\| dZ$*.* $C_a = \underline{\sigma}^{-1}\sqrt{\kappa_W \lambda_{\max}(\Sigma) p}$ *and* $C_\sigma \geq 0$ *is given in Lemma A5. The remainder term* $\Delta_{n,\varepsilon}$ *is*

$$\Delta_{n,\varepsilon}(\|\theta_b - \hat{\theta}_n\|) \leq \frac{C_2}{\underline{\sigma}^2_{n,\varepsilon}} \left(\Gamma_{n,\varepsilon} + \frac{(c_n n^{-1/2})^2}{\varepsilon} \|\theta_b - \hat{\theta}_n\|^\psi + \frac{c_n n^{-1/2}}{\varepsilon} \|\theta_b - \hat{\theta}_n\| \right).$$

Furthermore, if θ_b *is such that:*

$$\|\theta_b - \hat{\theta}_n\| \leq \left(\frac{\gamma - \overline{\gamma}}{\gamma} - M_{1,Z}\varepsilon \right) \frac{\underline{\sigma}_{n,\varepsilon}}{\sqrt{\kappa_W} L_G} := R_{n,\varepsilon},$$

then, with the same probability $1-(1+C)/c_n$*, uniformly in* $\|\theta_b - \hat{\theta}_n\| \leq \min(R_{n,G}, R_{n,\varepsilon})$*:*

$$\|\theta_{b+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})\|\theta_b - \hat{\theta}_n\| + \gamma \Delta_{n,\varepsilon}(\|\theta_b - \hat{\theta}_n\|). \tag{8}$$

The bound (8) is uniform in θ_b so that iterating on (8) does not change the probability level $1-(1+C)/c_n$. As a result, Proposition 1 will be used in Theorem 1 below to describe the full local optimization path, conditional on a single event of probability $1-(1+C)/c_n$. A larger value of c_n implies a greater probability for a successful contraction.

Relative to Lemma 1, the radius is $R_{n,\varepsilon} = R - O\left(\varepsilon + c_n n^{-1/2}\varepsilon^{-1}\right)$. This is nearly the same as the previous R when $\varepsilon = o(1)$ and $\sqrt{n}\varepsilon \to \infty$. The optimal choice of bandwidth, which maximizes $R_{n,\varepsilon}$, is $\varepsilon \asymp \sqrt{c_n} n^{-1/4}$. The same "large bandwidth" is also optimal for optimization with a smoothed objective function.

The dependence on c_n highlights the relationship between the tuning parameters and the probability of a successful contraction in (8). A higher probability (larger c_n) implies a larger optimal bandwidth $\varepsilon \asymp \sqrt{c_n} n^{-1/4}$. However, $R_{n,\varepsilon}$ is smaller: a larger bandwidth can increase the probability of a successful contraction but only within a smaller neighborhood.

Theorem 1. *Assume, without loss of generality, that* $R_{n,\varepsilon} \leq R_{n,G}$*. Take* $\gamma \in (0,1)$*,* $\overline{\gamma} \in (0,\gamma)$*, and* $\tau \in (0,1)$*. Suppose Assumptions 1-2 hold, and assume that* ε*,* $c_n n^{-1/2}$ *are small enough that the following two inequalities hold:*

$$\Delta_{n,\varepsilon}(R_{n,\varepsilon}) \leq \frac{\overline{\gamma}}{\gamma} R_{n,\varepsilon}, \tag{9}$$

$$\frac{(c_n n^{-1/2})^2}{\varepsilon} + \frac{c_n n^{-1/2}}{\varepsilon} < \frac{\overline{\gamma}\underline{\sigma}^2 \tau}{\gamma C_2}. \tag{10}$$

Then, with probability $1-(1+C)/c_n$*, uniformly in* $\|\theta_0 - \hat{\theta}_n\| \leq R_{n,\varepsilon}$*, for all* $b \geq 0$*:*

$$\begin{aligned} \|\theta_b - \hat{\theta}_n\| \leq & \left(1 - \overline{\gamma} + \tau\overline{\gamma}\right)^b \|\theta_0 - \hat{\theta}_n\| \\ & + \frac{\gamma}{\overline{\gamma}(1-\tau)} \frac{C_2}{\underline{\sigma}^2_{n,\varepsilon}} \left(\Gamma_{n,\varepsilon} + C_{n,\varepsilon} \left[\frac{(c_n n^{-1/2})^2}{\varepsilon} \right]^{\frac{1}{1-\psi}} \right), \end{aligned} \tag{11}$$

setting $[\frac{(c_n n^{-1/2})^2}{\varepsilon}]^{\frac{1}{1-\psi}} = 0$ *and* $C_{n,\varepsilon} = 0$ *if* $\psi = 1$*, while* $C_{n,\varepsilon} = \left(\frac{\overline{\gamma}\underline{\sigma}^2\tau}{\gamma C_2} - \frac{c_n n^{-1/2}}{\varepsilon}\right)^{\frac{\psi}{\psi-1}}$ *if* $\psi < 1$*. Suppose that* $\varepsilon = o(1)$ *and* $\sqrt{n}\varepsilon \to +\infty$*, then for* $c_n = O(1)$*:*

$$b \geq \frac{\log(\Gamma_{n,\varepsilon}) - \log(\|\theta_0 - \hat{\theta}_n\|)}{\log(1 - \overline{\gamma} + \tau\overline{\gamma})} \Rightarrow \sqrt{n}\|\theta_b - \hat{\theta}_n\| = o_p(1).$$

Theorem 1 extends Lemma 1 to finite samples. While Lemma 1 implies that $\theta_b \to \theta^\dagger$ as b increases, here (11) only implies that $\|\theta_b - \hat{\theta}_n\|$ is a $O(\Gamma_{n,\varepsilon})$ when b is large.[10] The final step of Algorithm 1 returns the arg-minimizer over (θ_b) which is the most accurate iterate.

Together, the contraction (8) and condition (9) imply local stability: if $\|\theta_0 - \hat{\theta}_n\| \leq R_{n,\varepsilon}$, then $\|\theta_b - \hat{\theta}_n\| \leq R_{n,\varepsilon}$ for all $b \geq 0$, with probability $1 - (1 + C)/c_n$. Because $\Delta_{n,\varepsilon} = o(1)$, (9) requires the sample size n to be sufficently large and ε sufficiently small.

The remainder $\Delta_{n,\varepsilon}$ depends on $\|\theta_b - \hat{\theta}_n\|$ which affects the rate of convergence. Condition (10) implies that $\Delta_{n,\varepsilon}$ inflates the rate in (8) by at most a factor of $\tau\overline{\gamma}$, where $\tau \in (0, 1)$ should satisfy (10). Smaller values of τ require large sample sizes n. Likewise, a smaller n requires a larger τ resulting in slower convergence. In both conditions, the ratio $\overline{\gamma}/\gamma < 1$ affects the sample size requirement. The first Extension modifies (1) in such a way that $\overline{\gamma}/\gamma > 1$ becomes feasible (Section 4.1).

The convergence rate in (11) is the same as in Lemma 1, up to $\tau\overline{\gamma}$. Even though the problem is non-smooth, only $b = O(|\log[\Gamma_{n,\varepsilon}]|) = O(\log[n])$ iterations are need to find $\|\theta_b - \hat{\theta}_n\| = O_p(\Gamma_{n,\varepsilon})$. The term $\Gamma_{n,\varepsilon}$ captures, among other things, the smoothing bias in the estimates. Using the optimal bandwidth, $\varepsilon \asymp n^{-1/4}$, yields $\Gamma_{n,\varepsilon} = O_p(n^{-3/4})$ for both $\psi = 1$ and $\psi = 1/2$. More generally, any $\varepsilon = o(1)$ such that $\sqrt{n}\varepsilon \to \infty$ yields $\sqrt{n}\Gamma_{n,\varepsilon} = o(1)$. Undersmoothing is not required, this makes the estimates more robust to the choice of bandwidth compared to minimizing a smoothed objective function. Notice that smaller bandwidths such that $n^{(1+\psi)/2}\varepsilon \to \infty$ also satisfy $\sqrt{n}\Gamma_{n,\varepsilon} = o(1)$, however they can make optimization unstable. This is because $\varepsilon = o(n^{-1/2})$ implies $R_{n,\varepsilon} < 0$ for n large enough which negates the optimization results. This highlights the importance of considering both the optimization and estimation properties of the Algorithm.

[10] Convergence of the sequence $(\theta_b)_{b\geq 0}$ would require (1) to have a fixed point which only holds if there exists a θ such that $G_{n,\varepsilon}(\theta)' W_n \overline{g}_n(\theta) = 0$. For discontinuous moments, there may be no such θ in a finite sample. In the simple quantile example, the sample moments are a step function: for some values of $t \in (0, 1)$, there is no exact solution to the sample moment equation.

Global Convergence. The second set of results consider the combined steps:

$$\theta_{b+1} = \theta_b - \gamma[G_b' W_n G_b]^{-1} G_b' W_n \overline{g}_n(\theta_n), \tag{1}$$

$$\text{if } \|\overline{g}_n(\theta^{b+1})\|_{W_n} < \|\overline{g}_n(\theta_{b+1})\|_{W_n}, \text{ set } \theta_{b+1} = \theta^{b+1}, \quad b = 0, 1, \ldots, \tag{2}$$

where again $G_b = G_{n,\varepsilon}(\theta_b)$. The main difference with Lemma 2 is that, in the non-asymptotic setting, the norm-equivalence (4) becomes:

$$(1 - \overline{\gamma}/2)\|\theta - \hat{\theta}_n\|_{G'W_nG} \leq \|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n} \leq (1 + \overline{\gamma}/2)\|\theta - \hat{\theta}_n\|_{G'W_nG},$$

up to asymptotically vanishing terms, locally to $\hat{\theta}_n$, where $G = G(\theta^\dagger)$. However, the global step in Algorithm 1 compares $\|\overline{g}_n(\theta^{b+1})\|_{W_n}$ with $\|\overline{g}_n(\theta_{b+1})\|_{W_n}$ which is different from $\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}$. An additional inequality is needed:

$$(1-\tau)^2\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 \leq \|\overline{g}_n(\theta)\|_{W_n}^2 - \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 \leq (1+\tau)^2\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}^2,$$

locally to $\hat{\theta}_n$, and up to asymptotically negligible terms. This holds for any choice of $\tau \in (0,1)$ as long as $c_n n^{-1/2}$ is small enough. A tighter inequality, i.e. a smaller τ, requires a larger sample size. Then, the global convergence proof follows the same outline as Lemma 2: using global identification to ensure local stability. The term $\tilde{\Gamma}_{n,\varepsilon}$ which will determine the estimation accuracy is more complicated than $\Gamma_{n,\varepsilon}$ found in Theorem 1, equation (11):

$$\begin{aligned}\tilde{\Gamma}_{n,\varepsilon} = \Gamma_{n,\varepsilon} &+ (c_n n^{-1/2})^2 \varepsilon^{\psi-1} + (c_n n^{-1/2})^{1+\psi/2+\psi^2/2} \\ &+ \sqrt{\varepsilon}(c_n n^{-1/2})^{1+\psi/2} + (c_n n^{-1/2})^{3/2+\psi}\varepsilon^{\psi-1} + a_n^{1+\psi/(1-\psi)} + c_n n^{-1/2} a_n^{1/(1-\psi)},\end{aligned}$$

where $a_n = c_n n^{-1/2} + (c_n n^{-1/2})^2 \varepsilon^{-1}$ for $\psi < 1$ and $a_n = 0$ otherwise. Even though the expression is more involved, it still holds that: $\sqrt{n}\tilde{\Gamma}_{n,\varepsilon} = o(1)$ when $\varepsilon = o(1)$ and $\sqrt{n}\varepsilon \to \infty$.

Theorem 2. *Suppose Assumptions 1-2 hold. Take $\gamma \in (0,1)$, $\overline{\gamma} \in (0, \gamma)$, and pick $\tau \in (0,1)$ small enough to satisfy:*

$$\frac{1+\tau}{1-\tau}\frac{1+\overline{\gamma}/2}{1-\overline{\gamma}/2}(1-\overline{\gamma}) < 1. \tag{12}$$

Let $\tilde{\gamma} \in (0,1)$ be such that:

$$1 - \tilde{\gamma} := \frac{1+\tau}{1-\tau}(1-\overline{\gamma}).$$

There exists $\underline{r}_{n,g}$ with $\liminf_{n\to\infty} \underline{r}_{n,g} > 0$ and $k_n \geq 0$ satisfying:

$$\sup_{\theta\in\Theta}\left(\inf_{0\leq\ell\leq k_n}\|\theta^\ell - \theta\|_{G'W_nG}\right) \leq \underline{r}_{n,g},$$

such that for $c_n n^{-1/2}, \varepsilon$ small enough, and $b = k_n + j$, with $j \geq 0$:

$$\|\theta_b - \hat{\theta}_n\|_{G'W_nG} \leq (1 - \tilde{\gamma} + \tau\tilde{\gamma})^j \underline{r}_{n,g} + \tilde{\Gamma}_{n,\varepsilon}, \tag{13}$$

with probability $1 - (1 + C)/c_n$. Suppose that $\varepsilon = o(1)$, and $\sqrt{n}\varepsilon \to \infty$, then for $c_n = O(1)$:

$$j \geq \frac{\log(\Gamma_{n,\varepsilon}) - \log[\underline{r}_{n,g}]}{\log(1 - \tilde{\gamma} + \tau\tilde{\gamma})} \Rightarrow \sqrt{n}\|\theta_b - \hat{\theta}_n\|_{G'W_nG} = o_p(1). \tag{14}$$

Theorem 2 extends Lemma 2 to finite samples. Condition (12) ensures the local contraction is preserved when looking at different norms. Similar to condition (10) in Theorem 1, a larger value of τ implies weaker requirements on the sample size n, but implies a slower rate of convergence. The coefficient $\tilde{\gamma}$ measures the effective rate of convergence (up to $\tau\tilde{\gamma}$) for the combined local and global steps. This rate is slower than in the local Theorem 1. Similar to $\underline{r}_g$ in Lemma 2, $\underline{r}_{n,g}$ is derived using the global identification condition. Note that $\liminf_{n\to\infty} \underline{r}_{n,g} > 0$ implies that k_n is bounded above. This ensures that the local convergence phase begins in finite time. The interpretation is similar to Lemma 2, the Algorithm transitions from global to fast local convergence after k_n iterations. Lastly, (14) gives the estimation result: θ_b is first-order equivalent to $\hat{\theta}_n$, when $b = k_n + j$ where $j = O(|\log[\Gamma_{n,\varepsilon}]|) = O(\log[n])$.

4 Extensions of the Main Results

This section provides two extensions for the local step in Algorithm 1 and Proposition 1. The first extension adds momentum to (1), a.k.a. the Polyak heavy-ball (Polyak, 1964). It is commonly used in stochastic gradient-descent to accelerate convergence. Here it can be used to either accelerate convergence or maintain the rate of convergence while reducing the effect of sampling noise on optimization. The second extension builds a quasi-Newton approximation of $G_{n,\varepsilon}$ which is useful when the smoothed Jacobian does not have closed-form.

4.1 Momentum: Acceleration or Noise Reduction

The first extension considers a modification of the local search step:

$$\theta_{b+1} = \theta_b - \gamma(G_b'W_nG_b)^{-1}G_b'W_n\overline{g}_n(\theta_b) + \alpha(\theta_b - \theta_{b-1}), \tag{1'}$$

where $\theta_{-1} = \theta_0$, $G_b = G_{n,\varepsilon}(\theta_b)$ and $\alpha \in [0,1)$ is the momentum parameter. Two new quantities affect convergence, the companion matrix $A(\gamma,\alpha)$ and the effective rate $\gamma(\alpha)$:

$$A(\gamma,\alpha) = \begin{pmatrix} 1-\gamma+\alpha & -\alpha \\ 1 & 0 \end{pmatrix}, \quad 1-\gamma(\alpha) = \sigma_{\max}[A(\gamma,\alpha)],$$

For any $\gamma \in (0,1)$, there exists $\alpha \in (0,1]$ such that $\gamma(\alpha) > \gamma$, and a $\alpha^\star$ which maximizes $\gamma(\alpha)$. Table 2 provides a selection of combinations $(\gamma, \alpha^\star)$. The last row measures $\gamma/\overline{\gamma(\alpha)} < 1$ which implies that setting $\overline{\gamma}/\gamma > 1$ is now possible: momentum can affect the convergence rate / sensitivity to sampling noise tradeoff in Proposition 1.

Table 2: Values of γ and optimal choice of α

γ	0.01	0.05	0.1	0.2	0.3	0.4	0.6	0.8
$\alpha^\star$	0.81	0.60	0.47	0.31	0.21	0.14	0.05	0.01
$\gamma(\alpha^\star)$	0.10	0.22	0.32	0.45	0.54	0.63	0.77	0.89
$\gamma/\gamma(\alpha^\star)$	0.10	0.22	0.32	0.45	0.55	0.63	0.78	0.90

Notice that $\theta_{b+1} - \alpha(\theta_b - \theta_{b-1}) = \theta_b - \gamma(G_b'W_nG_b)^{-1}G_b'W_n\overline{g}_n(\theta_b)$ is the same as (1) in Proposition 1, so convergence of (1') follows from the same derivations. The result are given for $\boldsymbol{\theta_b} = (\theta_b', \theta_{b-1}')'$ to denote θ_b and its lagged value θ_{b-1}; similarly $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} = (\hat{\theta}_n', \hat{\theta}_n')'$.

Proposition 2. *Suppose Assumptions 1-2 hold. Take $\gamma \in (0,1), \alpha \in [0,1)$ and $\overline{\gamma} \in (0, \gamma(\alpha))$. Take $c_n \geq 1$ and R_G as in Lemma 1. Uniformly in $\theta_b \in \Theta$ such that $\|\theta_b - \theta^\dagger\| \leq R_{n,G} := R_G - C_a c_n n^{-1/2}$:*

$$\begin{aligned}\|\boldsymbol{\theta_{b+1}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}\| \leq &\left(1 - \gamma(\alpha) + \gamma\underline{\sigma}_{n,\varepsilon}^{-1}\sqrt{\kappa_W}L_G\|\theta_b - \hat{\theta}_n\| + M_{1,Z}\varepsilon\right)\|\boldsymbol{\theta_b} - \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}\| \\ &+ \gamma\Delta_{n,\varepsilon}(\|\theta_b - \hat{\theta}_n\|) \end{aligned} \tag{15}$$

with probability $1-(1+C)/c_n$. The remainder $\Delta_{n,\varepsilon}$ is the same as in Proposition 1. Furthermore, if θ_b is such that:

$$\|\theta_b - \hat{\theta}_n\| \leq \left(\frac{\gamma(\alpha) - \overline{\gamma}}{\gamma} - M_{1,Z}\varepsilon\right)\frac{\underline{\sigma}_{n,\varepsilon}}{\sqrt{\kappa_W}L_G} := R_{n,\varepsilon}(\alpha), \tag{16}$$

then, with probability $1-(1+C)/c_n$:

$$\|\boldsymbol{\theta_{b+1}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}\| \leq (1-\overline{\gamma})\|\boldsymbol{\theta_b} - \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}\| + \gamma\Delta_{n,\varepsilon}(\|\theta_b - \hat{\theta}_n\|). \tag{17}$$

The convergence is now stated in terms of $\boldsymbol{\theta_{b+1}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$, similar to a first-order vector autoregression. Notice that $R_{n,\varepsilon}(\alpha) \geq R_{n,\varepsilon}$ when $\gamma(\alpha) \geq \gamma$: it may be possible to (slightly) increase the area of convergence for a well chosen value of α.

4.2 A Monte-Carlo quasi-Newton Implementation

In many empirical applications, the smoothed Jacobian $G_{n,\varepsilon}$ is not available in closed form. The following proposes a computationally attractive approximation.

Algorithm 2 quasi-Newton approximation $\hat{G}_b$ of $G_{n,\varepsilon}(\theta_b)$

1) **Input:** $L \geq d_\theta$,
2) **Moment Update:**
if $b = 0$ **then** ▷ Initialization
  draw $Z_{-\ell} \sim \mathcal{N}(0, I_{d_\theta})$, $\ell = 0, \ldots, L-1$
  compute $Y_{-\ell} = \frac{1}{\varepsilon}[\overline{g}_n(\theta_0 + \varepsilon Z_{-\ell}) - \overline{g}_n(\theta_0)]$
else ▷ Update
  draw $Z_b \sim \mathcal{N}(0, I_{d_\theta})$
  compute $Y_b = \frac{1}{\varepsilon}[\overline{g}_n(\theta_b + \varepsilon Z_b) - \overline{g}_n(\theta_b)]$
end if
3) **Least-Squares Approximation:**
de-mean $\tilde{Z}_{b-\ell} = Z_{b-\ell} - \sum_{\ell=0}^{L-1} Z_{b-\ell}/L$
compute $\hat{G}_L(\theta_b) = \sum_{\ell=0}^{L-1} Y_{b-\ell}\tilde{Z}'_{b-\ell} \left(\sum_{\ell=0}^{L-1} \tilde{Z}_{b-\ell}\tilde{Z}'_{b-\ell}\right)^{-1}$

Algorithm 2 involves an additional tuning parameter: $L \geq d_\theta$. A larger value of L reduces the Monte-Carlo error but involves more lagged values $\theta_{b-\ell}$, $\ell = 0, \ldots, L-1$, which slows convergence. The theoretical results require $L \to \infty$ but in practice, setting $L \geq \max(25, 1.5 \times d_\theta)$ yields good results. The Algorithm as presented above only updates one direction at a time but updating several directions can speed-up convergence because fewer lags are involved.[11] For simplicity, the results are only derived for the sample mean estimator:

$$\hat{G}_L(\theta_b) = \frac{1}{L}\sum_{\ell=0}^{L-1} \frac{1}{\varepsilon}[\overline{g}_n(\theta_{b-\ell} + \varepsilon Z_{b-\ell}) - \overline{g}_n(\theta_{b-\ell})]Z'_{b-\ell}.$$

The following extends Proposition 1 when $\hat{G}_L(\theta_b)$ is used in (1). It is assumed that the Algorithms runs for at most $1 \leq b_{\max} < \infty$ iterations (a stopping criterion).

Proposition 3. *Suppose Assumptions 1-2 hold. There exists* $0 < \hat{R}_G \leq R_G$ *such that*

[11] Updating several directions in parallel, e.g. 2 or 4, can improve the finite-sample optimization properties because, for the same choice of L, $\hat{G}_b$ depends on fewer lagged values, e.g. $L/2$ or $L/4$.

uniformly in $\mathcal{E}_b := (\max_{-L+1\leq\ell\leq 0}\|\theta_{b-\ell}-\hat{\theta}_n\|)\leq\hat{R}_G$*, we have with probability* $1-(5+C)/c_n$*:*

$$\begin{aligned}\|\theta_{b+1}-\hat{\theta}_n\| \leq (1-\gamma+\gamma\hat{\underline{\sigma}}_{n,\varepsilon}^{-1}\sqrt{\kappa_W}L_G[\|\theta_b-\hat{\theta}_n\|+\mu_{d_\theta}\mathcal{E}_b])\|\theta_b-\hat{\theta}_n\| \\ +\frac{\gamma}{\hat{\underline{\sigma}}_{n,\varepsilon}^2}\hat{\Delta}_{n,\varepsilon}(\|\theta_b-\hat{\theta}_n\|,\mathcal{E}_b),\end{aligned} \tag{18}$$

where $\mu_{d_\theta}=\mathbb{E}(\|ZZ'-I_{d_\theta}\|)$*,* $\hat{\underline{\sigma}}_{n,\varepsilon}=\underline{\sigma}/2-C_{\sigma,2}(c_nn^{-1/2}\varepsilon^{\psi-1}+\varepsilon+L^{-1/2})\delta_n^{3/2}$*, and* $\delta_n=\log(c_n)+\log(b_{\max}+L+1)$*. The remainder has the form:*

$$\hat{\Delta}_{n,\varepsilon}(\|\theta_b-\hat{\theta}_n\|,\mathcal{E}_b)=C_3\left(\hat{\Gamma}_{n,\varepsilon}+\delta_n^{3/2}[c_nn^{-1/2}\varepsilon^{\psi-1}+\varepsilon+L^{-1/2}]\mathcal{E}_b+c_nn^{-1/2}\|\theta_b-\hat{\theta}_n\|^\psi\right),$$

with $\hat{\Gamma}_{n,\varepsilon}=c_nn^{-1/2}\left[(c_nn^{-1/2})^\psi+\delta_n^{3/2}(c_nn^{-1/2}\varepsilon^{\psi-1}+\varepsilon+L^{-1/2})\right]$*. Suppose that:*

$$\mathcal{E}_b\leq\frac{\overline{\gamma}-\gamma}{\gamma}\frac{\hat{\underline{\sigma}}_{n,\varepsilon}}{L_G\sqrt{\kappa_W}(1+\mu_{d_\theta})}:=\hat{R}_{n,\varepsilon},$$

then with probability $1-(5+C)/c_n$*, we have:*

$$\|\theta_{b+1}-\hat{\theta}_n\|\leq(1-\overline{\gamma})\|\theta_b-\hat{\theta}_n\|+\frac{\gamma}{\hat{\underline{\sigma}}_{n,\varepsilon}^2}\hat{\Delta}_{n,\varepsilon}(\|\theta_b-\hat{\theta}_n\|,\mathcal{E}_b).$$

Compared to Proposition 1, $b_{\max}$ now plays a role in the results. This is because $\hat{G}_L(\theta_b)$ is computed using Monte-Carlo simulations and the worst-case approximation error can be arbitrarily large over infinitely many iterations. The dependence is only logarithmic because $\hat{G}_L(\theta_b)\simeq\frac{1}{L}\sum_{\ell=0}^{L-1}\partial_\theta g(\theta_{b-\ell})Z_{b-\ell}Z_{b-\ell}'$, where $Z_{b-\ell}Z_{b-\ell}'$ are iid Wishart distributed. Relying on properties of this distribution, the proof derives exponential tail bounds for the singular values of $\hat{G}_b$. This allows $b_{\max}$ to increase linearly or polynomially in n so that the Monte-Carlo error term δ_n only diverges logarithmically.

The probability bound is now $1-(5+C)/c_n$ to account for the Monte-Carlo error. Comparing $\Delta_{n,\varepsilon}$ with $\hat{\Delta}_{n,\varepsilon}$ in Propositions 1, 3 highlights the importance of $\mathcal{E}_b$ in the convergence. The dependence on lagged values underlines the lagged convergence of quasi-Newton methods compared to Proposition 1 or using the $\hat{G}_{n,\varepsilon}$ described in Section 2. The term $\hat{\Gamma}_{n,\varepsilon}$ depends on $(nL)^{-1/2}$ so that $L\to\infty$ is required. This divergence can, in theory, be arbitrarily slow.

5 Simulated and Empirical Examples

5.1 Dynamic Discrete Choice

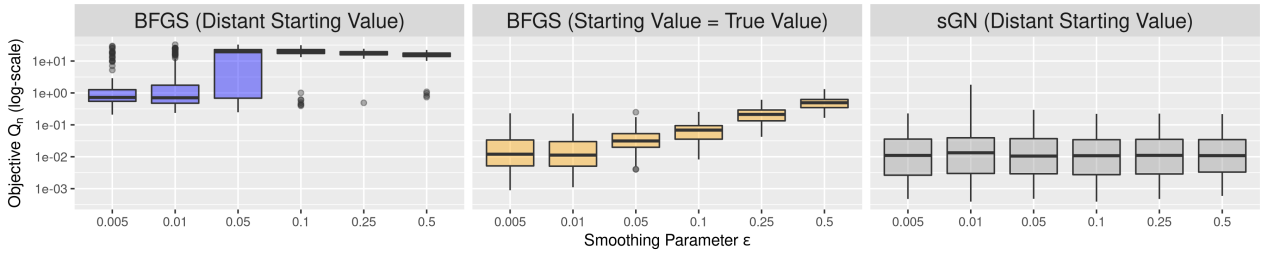
The first simulated example adapts an example from Bruins et al. (2018). The model is a simple panel dynamic discrete choice model. The data generating process (dgp) is given by:

$$y_{it} = \mathbb{1}\{x_{it}'\beta + u_{it} > 0\}, \quad u_{it} = e_{it} + \rho e_{it-1}, \quad e_{it} \overset{iid}{\sim} \mathcal{N}(0,1),$$

where x_{it} are iid, strictly exogenous regressors, $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The parameter of interest is $\theta = (\beta, \rho)$. The model is estimated by matching OLS estimates from a linear regression of y_{it} on x_{it} and y_{it-1} between sample and simulated datasets. Because of the indicator function $\mathbb{1}$, the objective function is discontinuous in θ.

Figure 4 and Table 3 compare the local step of sGN with smoothed GMM estimates based on replacing $\mathbb{1}\{\cdot > 0\}$ with $\Phi(\cdot/\varepsilon)$ in the dgp, where Φ is the CDF of the standard Gaussian. The simulations use $n = 250$, $T = 10$, $\dim(\beta) = 14$, $\beta_1^\dagger = \cdots = \beta_5^\dagger = 1/\sqrt{5}$, $\beta_6^\dagger = \cdots = \beta_{14}^\dagger = 0$, $\rho^\dagger = 0.7$. The initial value is $\theta_0 = (0, \ldots, 0)$ for sGN, and smoothed GMM estimates use BFGS for optimization. Because the latter systematically fails to converge, it is also evaluated using $\theta_0 = \theta^\dagger$.

Figure 4: Dynamic Discrete Choice: Distribution of Objection Function



Legend: 100 Monte-Carlo replications, sGN = Algorithm 1 initialized at $\theta_0 = (0, \ldots, 0)$. Smoothed Moments: initialized at distant $\theta_0 = (0, \ldots, 0)$ and true $\theta_0 = \theta^\dagger$ - computed using BFGS (R optim).

Figure 4 shows that sGN performs well for all values of ε considered, except for one replication when $\varepsilon = 0.01$. BFGS using smoothed moments systematically fails to converge from $\theta_0 = (0, \ldots, 0)$.[12] Starting from $\theta_0 = \theta^\dagger$, BFGS only fits the moments well for $\varepsilon \in [0.005, 0.01]$.[13] For $\varepsilon > 0.01$, the smoothing bias is large. Table 3 further illustrates how BFGS fails to converge when $\theta_0 = \theta^\dagger$. In comparison, sGN converges with properties that are stable over the range of ε. Estimates of ρ are downward biased, regardless of smoothing.

[12] Note that the logarithmic scale used for the y-axis implies that the discrepancy with sGN is large.

[13] For smaller values of ε, BFGS return $\theta = \theta_0$ because $\partial_\theta \overline{g}_n(\theta_0) = 0$ numerically.

Table 3: Dynamic Discrete Choice: Bias and Mean Absolute Error

	0.005	0.01	0.05	0.1	0.25	0.5	0.005	0.01	0.05	0.1	0.25	0.5
	Coefficient β_1						Coefficient ρ					
	Bias											
sGN	0.004	0.020	0.004	0.004	0.002	0.002	-0.088	-0.077	-0.089	-0.089	-0.091	-0.090
BFGS_1	1.035	3.775	14.43	20.78	22.61	21.10	0.285	0.292	0.296	0.300	0.300	0.300
BFGS_2	0.001	0.003	-0.002	-0.004	-0.004	0.021	-0.093	-0.090	-0.103	-0.114	-0.140	-0.152
	Mean Absolute Error											
sGN	0.037	0.052	0.038	0.037	0.037	0.037	0.094	0.100	0.095	0.095	0.097	0.096
BFGS_1	1.035	3.775	14.47	20.78	22.61	21.10	0.294	0.292	0.296	0.300	0.300	0.300
BFGS_2	0.036	0.036	0.037	0.036	0.035	0.039	0.097	0.095	0.107	0.115	0.140	0.152

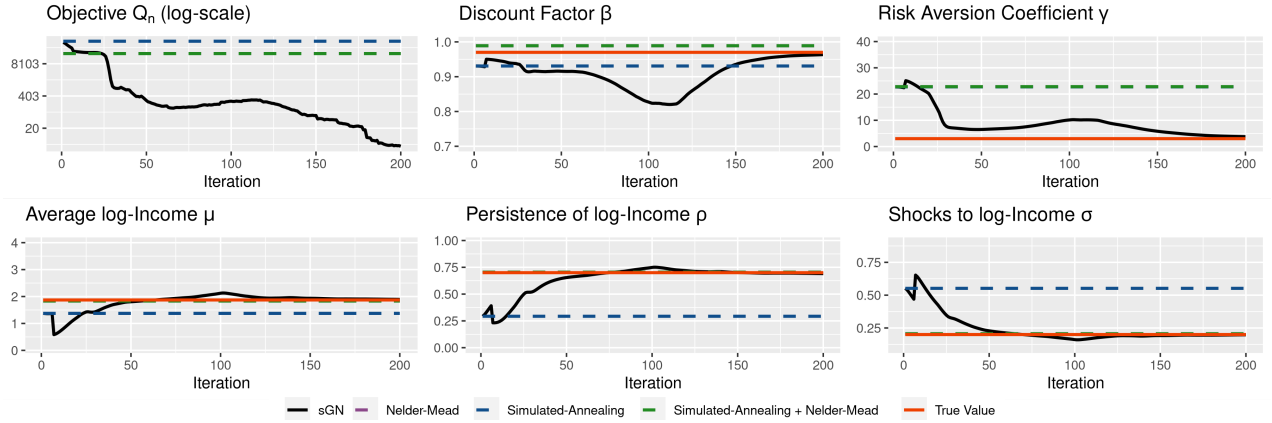
Legend: 100 Monte-Carlo replications, sGN = Algorithm 1 initialized at $\theta_0 = (0, \ldots, 0)$. $\text{BFGS}_{1,2}$: Smoothed Moments, initialized at $\theta_0 = (0, \ldots, 0)$ and true $\theta_0 = \theta^\dagger$, respectively - computed using BFGS (R optim).

5.2 Aiyagari Model

The second simulated example compares the estimation of a simple heterogenous agent model. These models are essential for understanding the distributional effects of macroeconomic policies. They are, however, very difficult to estimate. As a result, much of the literature report calibrated, rather than estimated, results. The following illustrates the properties of Algorithm 1 for an SMM estimation of a textbook Aiyagari model. The parameters of interest are $\theta = (\beta, \gamma, \mu, \rho, \sigma)$ which are the discount factor, risk-aversion, average log-income, persistence of log-income, and the standard deviation of log-income shocks. This type of estimation is challenging because the model is discretized, approximately solved by value function iterations, and the estimation is based on matching non-smooth moments (quantiles). The model and solution method are described in more details in Appendix G.1.

Figure 5 illustrates the optimization behaviour of Algorithm 1 for one simulated sample. A comparison with a local optimizer (Nelder-Mead), a global optimizer (Simulated Annealing), and a combination of both (Nelder-Mead after Simulated Annealing) are given as benchmarks. As shown in the top left panel, sGN converges quickly, whereas other methods fail to converge. Nelder-Mead finds accurate estimates for (μ, ρ, σ) but fails to accurately estimate (β, γ). Simulated Annealing fails to accurately estimate all coefficients. In comparison, sGN finds accurate estimates after 200 iterations for all coefficients. Figures G6, G7, G8 in the Appendix show similar results with different tuning parameters.

Figure 5: Aiyagari Model: local, global optimizers and sGN ($\varepsilon = 0.1$)



Legend: $n = 10000$, $T = 2$. $\gamma = 0.1$, $\alpha = 0.47$. sGN (black): Algorithm 1. Simulated-Annealing (dashed blue): 5000 iterations from θ_0. Simulated-Annealing + Nelder-Mead (dashed green): run Nelder-Mead after 5000 Simulated-Annealing iterations.

5.3 Interdependent Durations

The empirical example replicates the estimation of the joint duration model of employment by Honoré and de Paula (2018). They model the joint decision of optimal retirement age for married couples. Unlike single-agent duration models, their specification does not have a closed-form likelihood so they consider simulation-based estimation. Because the simulated outcomes are discrete, the resulting objective function is discontinuous and fairly difficult to minimize. Honoré and de Paula (2018) use a repeated succession of 5 optimizers.[14] The following compares their estimation from an accurate initial guess with Algorithm 1 initialized using a randomized Sobol sequence.

Table 4 replicates the first two columns of Tables 2 and 3 in Honoré and de Paula (2018, pp1319-1321). For brevity, only the first four coefficients are reported here. The estimates are similar for both specifications. However, the time required to compute the estimates is significantly reduced: the two specifications take 3.5 and 5.5 hours to estimate with their replication code. In comparison, Algorithm 1 finds estimates in 11 minutes using the same code. For comparison, for the 12 coefficient model, the random-walk Metropolis-Hastings (MH) algorithm has yet to converge after $100,000$ iterations, i.e. 10 hours, as illustrated in Figure G9, Appendix G.2, for the same distant starting value. Initialized from the same

[14]Their procedure: "The following loop of procedures was used until a loop produced a change in the parameter estimate of less than 10^{-5}. [...] (a) particle swarm [...], (b) Powell's conjugate direction method, (c) downhill simplex using Matlab's fminsearch routine (d) pattern search using Matlab's built-in routine, (e) particle swarm focusing on the jump-parameters [...]." See Honoré and de Paula (2018, p1330).

Table 4: Interdependent Duration Estimates: Honoré and de Paula (2018) and sGN

	Coefficients for Wives				Coefficients for Husbands			
	Honoré & de Paula		sGN		Honoré & de Paula		sGN	
δ	1.052	1.064	1.060	1.064	1.052	1.064	1.060	1.064
	(0.039)	(0.042)	(0.039)	(0.037)	(0.039)	(0.042)	(0.039)	(0.037)
θ_1	1.244	1.244	1.241	1.233	1.169	1.218	1.181	1.192
	(0.054)	(0.054)	(0.055)	(0.050)	(0.043)	(0.058)	(0.043)	(0.040)
$\geq$ 62 yrs-old	10.640	13.446	10.203	12.254	31.532	39.824	33.330	35.371
	(5.916)	(5.694)	(7.818)	(5.692)	(11.356)	(11.372)	(8.131)	(7.672)
$\geq$ 65 yrs-old	10.036	12.326	10.480	11.974	25.696	29.254	25.203	26.240
	(11.555)	(7.495)	(10.067)	(10.897)	(9.497)	(11.229)	(13.215)	(14.289)
...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Starting Obj. Value	93.70	89.77	2.10^4	5.10^4	-	-	-	-
Final Obj. Value	0.470	0.758	0.271	0.342	-	-	-	-
Number of Coef.	12	30	12	30	-	-	-	-
Computation Time	3h25m	5h34m	11min	11min	-	-	-	-

Legend: sGN: $\varepsilon = 10^{-2}$, $\gamma = 0.1$, $\alpha = 0.47$, $B = 250$ iterations in total. Husbands: - same as wives. Coefficients for wives and husbands are estimated jointly. Full estimation results are in Table G5.

starting value as Honoré and de Paula (2018), MH converges quickly (not reported).

6 Conclusion

This paper proposes an approach to tackle challenging GMM estimations. The procedure is simple to implement and has good finite sample theoretical properties. The empirical application illustrates the good performance of the Algorithm on a fairly large and difficult estimation problem. With estimation, inference is another essential part of empirical work. In the continuity of this paper, building on Algorithm 1 to design new Bayesian algorithms, or frequentist resampling methods in the spirit of Forneron (2022) could be useful.

References

Andrews, D. W. (1997): "A stopping rule for the computation of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, 913–931.

Belloni, A. and V. Chernozhukov (2009): "On the computational complexity of MCMC-based estimators in large samples," *The Annals of Statistics*, 37, 2011–2055.

Bertsekas, D. (2016): *Nonlinear Programming*, vol. 4, Athena Scientific.

Bhatia, R. (2013): *Matrix Analysis*, vol. 169, Springer Science & Business Media.

Birgé, L. and P. Massart (1998): "Minimum contrast estimators on sieves: exponential bounds and rates of convergence," *Bernoulli*, 329–375.

Brooks, S. P. (1998): "MCMC convergence diagnosis via multivariate bounds on log-concave densities," *The Annals of Statistics*, 26, 398–433.

Bruins, M., J. A. Duffy, M. P. Keane, and A. A. Smith Jr (2018): "Generalized indirect inference for discrete choice models," *Journal of econometrics*, 205, 177–203.

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

Chernozhukov, V. and H. Hong (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115, 293–346.

Dedecker, J. and S. Louhichi (2002): "Maximal inequalities and empirical central limit theorems," in *Empirical process techniques for dependent data*, Springer, 137–159.

Fernandes, M., E. Guerre, and E. Horta (2021): "Smoothing quantile regressions," *Journal of Business & Economic Statistics*, 39, 338–357.

Forneron, J. (2022): "Estimation and Inference by Stochastic Optimization," ArXiv:2004.09627.

Griewank, A. O. (1981): "Generalized descent for global optimization," *Journal of optimization theory and applications*, 34, 11–39.

Hazan, T., G. Papandreou, and D. Tarlow (2016): *Perturbations, Optimization, and Statistics*, MIT Press.

He, X., X. Pan, K. M. Tan, and W.-X. Zhou (2021): "Smoothed quantile regression with large-scale inference," *Journal of Econometrics*.

Honoré, B. E. and Á. de Paula (2018): "A new model for interdependent durations," *Quantitative Economics*, 9, 1299–1333.

Hyndman, R. J. and Y. Fan (1996): "Sample quantiles in statistical packages," *The American Statistician*, 50, 361–365.

Kaplan, D. M. and Y. Sun (2017): "Smoothed estimating equations for instrumental variables quantile regression," *Econometric Theory*, 33, 105–157.

Knittel, C. R. and K. Metaxoglou (2014): "Estimation of random-coefficient demand models: two empiricists' perspective," *Review of Economics and Statistics*, 96, 34–59.

Laurent, B. and P. Massart (2000): "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, 1302–1338.

McFadden, D. (1989): "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica: Journal of the Econometric Society*, 995–1026.

Mengersen, K. L. and R. L. Tweedie (1996): "Rates of convergence of the Hastings

and Metropolis algorithms," *The annals of Statistics*, 24, 101–121.

Muirhead, R. J. (1982): *Aspects of multivariate statistical theory*, Wiley series in probability and mathematical statistics. Probability and mathematical statistics.

Nemirovsky, A. S. and D. B. Yudin (1983): *Problem complexity and method efficiency in optimization*, Wiley-Interscience Series in Discrete Mathematics, Chichester, England: John Wiley & Sons.

Nesterov, Y. (2018): *Lectures on convex optimization*, Springer optimization and its applications, Cham, Switzerland: Springer International Publishing, 2 ed.

Nesterov, Y. and V. Spokoiny (2017): "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, 17, 527–566.

Newey, W. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, North Holland, vol. 36:4, 2111–2234.

Niederreiter, H. (1983): "A quasi-Monte Carlo method for the approximate computation of the extreme values of a function," in *Studies in pure mathematics*, Springer, 523–529.

Nocedal, J. and S. Wright (2006): *Numerical Optimzation*, Springer, second ed.

Polyak, B. T. (1964): "Some methods of speeding up the convergence of iteration methods," *Ussr computational mathematics and mathematical physics*, 4, 1–17.

——— (1987): "Introduction to optimization. optimization software," *Inc., Publications Division, New York*, 1, 32.

Polyak, B. T. and A. B. Tsybakov (1990): "Optimal order of accuracy of search algorithms in stochastic optimization," *Problemy Peredachi Informatsii*, 26, 45–53.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica: Journal of the Econometric Society*, 1403–1430.

Robinson, P. M. (1988): "The stochastic difference between econometric statistics," *Econometrica: Journal of the Econometric Society*, 531–548.

Tauchen, G. and R. Hussey (1991): "Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models," *Econometrica: Journal of the Econometric Society*, 371–396.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer Series in Statistics, New York, NY: Springer New York.

Vershynin, R. (2018): *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press.

Zhong, L. and J.-J. Forneron (2022): "Convexity Not Required: Estimation of Smooth Moment Condition Models," *Manuscript in preparation*.

Appendix A Preliminary Results for Section 3

A.1 Preliminary Results for the Population Problem

Lemma A1. *Suppose Assumptions 1-2 hold. There exists $\overline{r}_g > 0$ such that for any θ_b such that $\|\theta_b - \theta^\dagger\|_{G'WG} \leq \overline{r}_g$, the following hold:*

$$\|g(\theta_{b+1})\|_W \leq (1-\overline{\gamma})\|g(\theta_b)\|_W, \tag{3}$$

$$(1-\overline{\gamma}/2)\|\theta - \theta^\dagger\|_{G'WG} \leq \|g(\theta)\|_W \leq (1+\overline{\gamma}/2)\|\theta - \theta^\dagger\|_{G'WG}, \tag{4}$$

where $\theta_{b+1} = \theta_b - \gamma(G_b'WG_b)^{-1}G_b'Wg(\theta_b)$. There also exists a $\underline{r}_g \in (0, \overline{r}_g]$ such that:

$$\inf_{\|\theta-\theta^\dagger\|_{G'WG} \geq \overline{r}_g} \|g(\theta)\|_W \geq (1+\frac{\overline{\gamma}}{2})(1-\overline{\gamma})\underline{r}_g. \tag{5}$$

A.2 Preliminary Results for the Finite-Sample Problem

Lemma A2 (Useful Identities)**.** *For any $\theta \in \Theta$ and $\varepsilon > 0$, we have:*

i. $\overline{g}_{n,\varepsilon}(\theta) = \int_Z \overline{g}_n(\theta + \varepsilon Z)\phi(Z)dZ = \varepsilon \int_u \overline{g}_n(u)\phi(\frac{u-\theta}{\varepsilon})du$,

ii. $\overline{G}_{n,\varepsilon}(\theta) = -\frac{1}{\varepsilon}\int_Z \overline{g}_n(\theta + \varepsilon Z)\phi'(Z)dZ = -\int_u \overline{g}_n(u)\phi'(\frac{u-\theta}{\varepsilon})du$,

iii. $G_\varepsilon(\theta) = -\frac{1}{\varepsilon}\int_Z g(\theta + \varepsilon Z)\phi'(Z)dZ = \int_Z G(\theta + \varepsilon Z)\phi(Z)dZ$

Lemma A3 (Deterministic bounds)**.** *Suppose Assumption 1 holds then: i. For any $(\theta_1, \theta_2) \in \Theta$ and $\varepsilon > 0$: $\|G_\varepsilon(\theta_1) - G_\varepsilon(\theta_2)\| \leq L_G\|\theta_1 - \theta_2\|$, ii. For any $\theta \in \Theta$ and $\varepsilon > 0$: $\|G_\varepsilon(\theta) - G(\theta)\| \leq \varepsilon L_G M_{1,Z}$ where $M_{1,Z} = \int \|\phi'(Z)\|dZ$.*

Lemma A4 (Stochastic bounds)**.** *Suppose Assumptions 1-2 hold. Let $C_\Theta = \int_0^1 \sqrt{1 + \log[N(x, \Theta, \|\cdot\|)]}dx$, then there exists a universal constant $C > 0$ such that:*

$$\mathbb{E}\left(\sup_{\|\theta_1-\theta_2\|\leq\delta} \sqrt{n}\big\|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\big\|\right) \leq CC_\Theta L_g \delta^\psi.$$

For any $c_n \geq 1$, the above inequality implies:

a. Sample Moments:

$$\mathbb{P}\left(\sup_{\|\theta_1-\theta_2\|\leq\delta} \big\|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\big\| \leq c_n n^{-1/2} C_\Theta L_g \delta^\psi\right) \geq 1 - C/c_n,$$

b. Smoothed moments, for any $\varepsilon > 0$:

$$\mathbb{P}\left(\sup_{\|\theta_1-\theta_2\|\leq\delta}\left\|[\overline{g}_{n,\varepsilon}(\theta_1)-g_\varepsilon(\theta_1)]-[\overline{g}_{n,\varepsilon}(\theta_2)-g_\varepsilon(\theta_2)]\right\|\leq c_n n^{-1/2}C_\Theta L_g\delta^\psi\right)\geq 1-C/c_n,$$

c. Smoothed Jacobian, for any $\varepsilon > 0$:

$$\mathbb{P}\left(\sup_{\|\theta_1-\theta_2\|\leq\delta}\left\|[\overline{G}_{n,\varepsilon}(\theta_1)-G_\varepsilon(\theta_1)]-[\overline{G}_{n,\varepsilon}(\theta_2)-G_\varepsilon(\theta_2)]\right\|\leq c_n\varepsilon^{-1}n^{-1/2}C_\Theta L_g M_{1,Z}\delta^\psi\right)\geq 1-C/c_n,$$

where $M_{1,Z} = \int \|\phi'(Z)\|dZ$.

All three events a-c. hold jointly with the same probability bound $1 - C/c_n$.

Lemma A5 (Singular Values)**.** *Suppose Assumptions 1-2 hold. Let $\theta \in \Theta$ such that $\|\theta - \theta^\dagger\| \leq R_G$, where R_G is defined in Lemma 1, then:*

$$\sigma_{\min}[\overline{G}_{n,\varepsilon}(\theta)]\geq\underline{\sigma}-\left[\frac{c_n}{\varepsilon\sqrt{n}}C_\Theta L_g M_{1,Z}R_G^\psi+\varepsilon L_G M_{1,Z}\right]:=\underline{\sigma}_{n,\varepsilon},$$

with probability $1 - C/c_n$.

Lemma A6. *Suppose Assumptions 1-2 hold, then for any $\delta \geq 0$ and $\varepsilon > 0$, $c_n \geq 1$. We have for C and C_Θ as in Lemma A4:*

$$\begin{aligned}\mathbb{P}\Big(&\sup_{\|\theta_1-\theta_2\|\leq\delta}\|\overline{g}_n(\theta_1)-\overline{g}_n(\theta_2)-G_{n,\varepsilon}(\theta_1)(\theta_1-\theta_2)\|\\&\leq L_G\delta^2+L_gC_\Theta c_n n^{-1/2}[\delta^\psi+\varepsilon^{\psi-1}\delta]+L_G M_{1,Z}\varepsilon\delta\Big)\geq 1-C/c_n.\end{aligned}\tag{A.1}$$

Lemma A7. *Suppose Assumptions 1-2 hold. For $\eta \in (0, 1)$ and $c_n \geq 1$, let:*

$$R_n(\eta):=\frac{\eta\underline{\sigma}^2}{\kappa_W L_G}-C_a c_n n^{-1/2}-\frac{L_g}{L_G}C_\Theta(c_n n^{-1/2})^\psi.$$

Then uniformly in $\|\theta - \hat{\theta}_n\| \leq R_n(\eta)$, we have with probability $1 - (1 + C)/c_n$:

$$\begin{aligned}(1-\eta)\|\theta-\hat{\theta}_n\|_{G'W_nG}-\overline{\lambda}_W L_g C_\Theta(c_n n^{-1/2})^{1+\psi}&\leq\|\overline{g}_n(\theta)-\overline{g}_n(\hat{\theta}_n)\|_{W_n}\\&\leq(1+\eta)\|\theta-\hat{\theta}_n\|_{G'W_nG}+\overline{\lambda}_W L_g C_\Theta(c_n n^{-1/2})^{1+\psi},\end{aligned}\tag{A.2}$$

where $G = G(\theta^\dagger)$.

Lemma A8. *Suppose Assumptions 1-2 hold. For $\eta \in (0,1)$ and $R_n(\eta)$ defined in Lemma A7 let:*

$$\overline{r}_n(\eta) = \frac{R_n(\eta)}{\sqrt{\kappa_G \kappa_W}} - \|\theta^\dagger - \hat{\theta}_n\| \geq \frac{R_n(\eta)}{\sqrt{\kappa_G \kappa_W}} - C_a c_n n^{-1/2}$$
$$\underline{r}_{n,g}(\eta) = \frac{\delta(\overline{r}_{n,g}(\eta))}{\sqrt{\kappa_W}} - \overline{\lambda}_W^{1/2} c_n n^{-1/2} \left(L_g C_\Theta diam(\Theta)^\psi + \lambda_{\max}(\Sigma)^{1/2} p \right),$$

where δ is defined in Assumption 1, $p = dim(\overline{g}_n(\theta))$, $\Sigma = var(\overline{g}_n(\theta^\dagger))$, $\kappa_G = \overline{\sigma}/\underline{\sigma}$ and $\overline{\sigma} = \sigma_{\max}[G(\theta^\dagger)]$. Then with probability $1 - (1+C)/c_n$:

$$\inf_{\|\theta - \hat{\theta}_n\|_{G' W_n G}} \|\overline{g}_n(\theta)\|_{W_n} \geq \underline{r}_{n,g}(\eta).$$

Lemma A9. *Suppose Assumptions 1-2 hold. Take $\eta \in (0,1)$ and let $x_b = \|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}$. For any $\varepsilon > 0$, $c_n \geq 1$, we have:*

$$x_{b+1} \leq (1 - \overline{\gamma}) x_b + \frac{\underline{\sigma}_{n,\varepsilon}^{-3}}{(1-\eta)^2} \Delta_{2,n,\varepsilon}(x_b), \tag{A.3}$$

uniformly in $\|\theta_b - \hat{\theta}_n\| \leq R_n(\eta)$, defined in Lemma A7, with probability $1 - (1+C)/c_n$. The remainder has the form:

$$\Delta_{2,n,\varepsilon}(x_b) = C_4 \left(\Gamma_{n,\varepsilon} + \varepsilon^{\psi - 1} (c_n n^{-1/2})^2 + [c_n n^{-1/2} + (c_n n^{-1/2})^2 \varepsilon^{-1}] x_b^\psi + (c_n n^{-1/2}) x_b \right),$$

Lemma A10. *Suppose Assumptions 1-2 hold, take τ and $\eta \in (0,1)$. Then for $c_n n^{-1/2}$ small enough, we have uniformly in $\|\theta - \hat{\theta}_n\| \leq R_n(\eta)$:*

$$\begin{aligned}(1-\tau)^2 \|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 - \Gamma_{2,n,\varepsilon}^2(\eta,\tau) &\leq \|\overline{g}_n(\theta)\|_{W_n}^2 - \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 \\ &\leq (1+\tau)^2 \|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 + \Gamma_{2,n,\varepsilon}^2(\eta,\tau),\end{aligned}$$

with probability $1-(1+C)/c_n$. $n\Gamma_{2,n,\varepsilon}^2(\eta,\tau) = o(1)$ for any fixed $\eta, \tau \in (0,1)$ when $c_n = O(1)$, $\varepsilon = o(1)$, and $\sqrt{n}\varepsilon \to \infty$.

Appendix B Proofs for the Main Results

Proof of Lemma 3. We have:

$$\|G_{n,\varepsilon}(\hat{\theta}_n)'W_n\overline{g}_n(\hat{\theta}_n)\| \leq \|G(\theta^\dagger)'W_n\overline{g}_n(\hat{\theta}_n)\| \tag{B.4}$$
$$+ \|[G_{n,\varepsilon}(\hat{\theta}_n) - G_\varepsilon(\hat{\theta}_n)]'W_n\overline{g}_n(\hat{\theta}_n)\| \tag{B.5}$$
$$+ \|[G_\varepsilon(\hat{\theta}_n) - G_\varepsilon(\theta^\dagger)]'W_n\overline{g}_n(\hat{\theta}_n)\| \tag{B.6}$$
$$+ \|[G_\varepsilon(\theta^\dagger) - G(\theta^\dagger)]'W_n\overline{g}_n(\hat{\theta}_n)\|. \tag{B.7}$$

In the following, each term on the right-hand side of the inequality will be bounded.

Let $p = \dim(\overline{g}_n)$ and $\Sigma = \text{var}(g(\theta^\dagger; x_i))$, by construction of $\hat{\theta}_n$:

$$\|\hat{\theta}_n - \theta^\dagger\| \leq \underline{\sigma}^{-1}\sqrt{\kappa_W}\|\overline{g}_n(\theta^\dagger)\| \leq \underline{\sigma}^{-1}\sqrt{\kappa_W\lambda_{\max}(\Sigma)p}c_n n^{-1/2} = C_a c_n n^{-1/2},$$

with probability $1-1/c_n$. The last inequality follows from Markov's inequality: $\mathbb{P}(\|\Sigma^{-1/2}\overline{g}_n(\theta^\dagger)\| \geq \sqrt{p}c_n n^{-1/2}) \leq \mathbb{E}(\|\Sigma^{-1/2}\overline{g}_n(\theta^\dagger)\|^2)^{1/2}/[\sqrt{p}c_n] = 1/c_n$. Then, using the bounds in Lemma A4 and the Lipschitz continuity of G, we have:

$$\|\overline{g}_n(\hat{\theta}_n) - \overline{g}_n(\theta^\dagger) - G(\theta^\dagger)(\hat{\theta}_n - \theta^\dagger)\| \leq L_G C_a^2 c_n^2 n^{-1} + C_\Theta L_g C_a^\psi (c_n n^{-1/2})^{1+\psi} \leq C_b(c_n n^{-1/2})^{1+\psi},$$

with probability $1-(1+C)/c_n$ since $\psi \in (0,1]$, with C from Lemma A4. Now note that:

$$G(\theta^\dagger)'W_n[\overline{g}_n(\theta^\dagger) - G(\theta^\dagger)(\hat{\theta}_n - \theta^\dagger)] = 0,$$

by construction of $\hat{\theta}_n$ and using projection arguments. This implies that:

$$\|G(\theta^\dagger)'W_n\overline{g}_n(\hat{\theta}_n)\| \leq \overline{\sigma}\overline{\lambda}_W C_b(c_n n^{-1/2})^{1+\psi} = C_{(B.4)}(c_n n^{-1/2})^{1+\psi}, \tag{B.4}$$

with probability $1-(1+C)/c_n$ for some constant $C_{(B.4)}$, where $\overline{\sigma} = \sigma_{\max}[G(\theta^\dagger)]$. Now, in order to bound (B.5), use:

$$\|[G_{n,\varepsilon}(\hat{\theta}_n) - G_\varepsilon(\hat{\theta}_n)]'W_n\overline{g}_n(\hat{\theta}_n)\| \leq \overline{\lambda}_W \sup_{\theta\in\Theta}\|G_{n,\varepsilon}(\theta) - G_\varepsilon(\theta)\| \times \|\overline{g}_n(\hat{\theta}_n)\|.$$

Using the previous bounds, we have:

$$\|\overline{g}_n(\hat{\theta}_n)\| \leq \|\overline{g}_n(\hat{\theta}_n) - \overline{g}_n(\theta^\dagger) - G(\theta^\dagger)(\hat{\theta}_n - \theta^\dagger)\| + \|\overline{g}_n(\theta^\dagger) + G(\theta^\dagger)(\hat{\theta}_n - \theta^\dagger)\| \leq C_c c_n n^{-1/2},$$

with probability $1-(1+C)/c_n$ for some C_c which depends on C_a and C_b. We can also bound the difference between the sample and population smoothed Jacobian:

$$\begin{aligned}
&\sup_{\theta\in\Theta}\|G_{n,\varepsilon}(\theta)-G_\varepsilon(\theta)\| \\
&\leq \frac{1}{\varepsilon}\sup_{\theta\in\Theta}\|\overline{g}_n(\theta)-g(\theta)\|\int_Z\|\phi'(Z)\|dZ \\
&\leq \frac{1}{\varepsilon}\left(\sup_{\theta\in\Theta}\|[\overline{g}_n(\theta)-g(\theta)]-[\overline{g}_n(\theta^\dagger)-g(\theta^\dagger)]\|+\|\overline{g}_n(\theta^\dagger)-g(\theta^\dagger)\|\right)\int_Z\|\phi'(Z)\|dZ \\
&\leq \frac{1}{\varepsilon}\left(C_\Theta L_g\text{diam}(\Theta)c_n n^{-1/2}+\sqrt{\lambda_{\max}(\Sigma)p}c_n n^{-1/2}\right)M_{1,Z},
\end{aligned}$$

with probability $1-C/c_n$. Putting these bounds together, we get:

$$\|[G_{n,\varepsilon}(\hat{\theta}_n)-G_\varepsilon(\hat{\theta}_n)]'W_n\overline{g}_n(\hat{\theta}_n)\|\leq\frac{C_{(B.5)}}{\varepsilon}(c_n n^{-1/2})^{2+\psi}, \tag{B.5}$$

with probability $1-(1+C)/c_n$. For (B.6), the bound for $\overline{g}_n(\hat{\theta}_n)$ can be used in tandem with the Lipschitz-continuity of the smoothed Jacobian and the bound on $\|\hat{\theta}_n-\theta^\dagger\|$, this yields:

$$\|[G_\varepsilon(\hat{\theta}_n)-G_\varepsilon(\theta^\dagger)]'W_n\overline{g}_n(\hat{\theta}_n)\|\leq\overline{\lambda}_W L_G C_a C_c(c_n n^{-1/2})^{2+\psi}=C_{(B.6)}(c_n n^{-1/2})^{2+\psi}, \tag{B.6}$$

with probability $1-(1+C)/c_n$. For the last (B.7), we have

$$\|[G_\varepsilon(\theta^\dagger)-G(\theta^\dagger)]'W_n\overline{g}_n(\hat{\theta}_n)\|\leq\overline{\lambda}_W C_c L_G M_{1,Z}\varepsilon c_n n^{-1/2}=C_{(B.7)}\varepsilon c_n n^{-1/2}, \tag{B.7}$$

with probability $1-(1+C)/c_n$. Combining all the bounds into the smoothed first-order condition, we have:

$$\|G_{n,\varepsilon}(\hat{\theta}_n)'W_n\overline{g}_n(\hat{\theta}_n)\|\leq C_{(B.8)}(c_n n^{-1/2})^{1+\psi}\left(1+\frac{c_n n^{-1/2}}{\varepsilon}+\frac{\varepsilon}{(c_n n^{-1/2})^\psi}\right), \tag{B.8}$$

with probability $1-(1+C)/c_n$ using $C_{(B.8)}=\max(C_{(B.4)}+C_{(B.6)},C_{(B.5)},C_{(B.7)})$ and assuming (without loss of generality) that $(c_n n^{-1/2})^{2+\psi}\leq(c_n n^{-1/2})^{1+\psi}$, i.e. $c_n n^{-1/2}\leq 1$. □

Proof of Proposition 1 Let $G_b=G_{n,\varepsilon}(\theta_b)$ with $\|\theta_b-\hat{\theta}_n\|\leq R_G-C_a c_n n^{-1/2}$. Using the constants from Lemma 3, we have with a probability of at least $1-1/c_n$ that $\|\hat{\theta}_n-\theta^\dagger\|\leq R_G$. By definition of θ_{b+1}, we have:

$$\theta_{b+1}=\theta_b-\gamma(G_b'W_nG_b)^{-1}G_b'W_n\overline{g}_n(\theta_b),$$

which we can re-write as:

$$\theta_{b+1} - \hat{\theta}_n - (1-\gamma)(\theta_b - \hat{\theta}_n) = -\gamma(G_b' W_n G_b)^{-1} G_b' W_n \left(\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n) - G_b(\theta_b - \hat{\theta}_n) \right) \quad \text{(B.9)}$$

$$+ \gamma(G_b' W_n G_b)^{-1} G_b' W_n \overline{g}_n(\hat{\theta}_n). \quad \text{(B.10)}$$

The proof boils down to finding bounds for (B.9) and (B.10). First, using the singular value bound $\underline{\sigma}_{n,\varepsilon}$ from Lemma A5, we have:

$$\|(B.10)\| \leq \gamma \underline{\sigma}_{n,\varepsilon}^{-2} \, \underline{\lambda}_W^{-1} \|G_{n,\varepsilon}(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n)\| \quad \text{(B.11)}$$

$$+ \gamma \underline{\sigma}_{n,\varepsilon}^{-2} \, \kappa_W \|G_{n,\varepsilon}(\hat{\theta}_n) - G_b\| \times \|\overline{g}_n(\hat{\theta}_n)\| \quad \text{(B.12)}$$

Lemma 3 gives the bound $\|(B.11)\| \leq \gamma \underline{\sigma}_{n,\varepsilon}^{-2} \, \underline{\lambda}_W^{-1} \Gamma_{n,\varepsilon}$, with probability $1 - (1+C)/c_n$, for the C in Lemma A4. A bound $\|\overline{g}_n(\hat{\theta}_n)\| \leq C_c c_n n^{-1/2}$ with probability $1 - (1+C)/c_n$, is derived in the proof of Lemma 3. It remains to bound $\|G_{n,\varepsilon}(\hat{\theta}_n) - G_{n,\varepsilon}(\theta_b)\|$. Using Lemma A4:

$$\begin{aligned}
\|(B.12)\| &\leq \frac{\gamma \kappa_W}{\underline{\sigma}_{n,\varepsilon}^2} C_c c_n n^{-1/2} \left(\|G_\varepsilon(\theta_b) - G_\varepsilon(\hat{\theta}_n)\| + \|[G_\varepsilon(\theta_b) - G_\varepsilon(\hat{\theta}_n)] - [G_{n,\varepsilon}(\theta_b) - G_{n,\varepsilon}(\hat{\theta}_n)]\| \right) \\
&\leq \frac{\gamma \kappa_W}{\underline{\sigma}_{n,\varepsilon}^2} C_c c_n n^{-1/2} \left(L_G \|\theta_b - \hat{\theta}_n\| + C_\Theta L_g M_{1,Z} \frac{c_n n^{-1/2}}{\varepsilon} \|\theta_b - \hat{\theta}_n\|^\psi \right),
\end{aligned}$$

with probability $1 - C/c_n$. Putting the bounds together, we have:

$$\|(B.10)\| \leq \frac{\gamma C_{(B.10)}}{\underline{\sigma}_{n,\varepsilon}^2} \left(\Gamma_{n,\varepsilon} + c_n n^{-1/2} \|\theta_b - \hat{\theta}_n\| + \frac{(c_n n^{-1/2})^2}{\varepsilon} \|\theta_b - \hat{\theta}_n\|^\psi \right),$$

with probability $1 - (1+C)/c_n$.[15] For the next part, consider $\|(B.9)\| \leq \gamma \underline{\sigma}_{n,\varepsilon}^{-1} \sqrt{\kappa_W} \|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n) - G_b(\theta_b - \hat{\theta}_n)\|$ so we can focus on bounding the difference found in the norm:

$$\|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n) - G_b(\theta_b - \hat{\theta}_n)\| \leq \|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n) - g(\theta_b) - g(\hat{\theta}_n)\| \quad \text{(B.13)}$$

$$+ \|g(\theta_b) - g(\hat{\theta}_n) - G(\theta_b)(\theta_b - \hat{\theta}_n)\| \quad \text{(B.14)}$$

$$+ \|G(\theta_b) - G_\varepsilon(\theta_b)\| \times \|(\theta_b - \hat{\theta}_n)\| \quad \text{(B.15)}$$

$$+ \|G_{n,\varepsilon}(\theta_b) - G_\varepsilon(\theta_b)\| \times \|(\theta_b - \hat{\theta}_n)\|. \quad \text{(B.16)}$$

Lemma A4 gives $\|(B.13)\| \leq C_\Theta L_g c_n n^{-1/2} \|\theta_b - \hat{\theta}_n\|^\psi$, with probability $1 - C/c_n$. Lipschitz continuity of G implies $\|(B.14)\| \leq L_G \|\theta_b - \hat{\theta}_n\|^2$, which plays the same role for convergence

[15] The probability does not change when combining the bounds together because they all are derived from the same two events, one is the empirical process bound used for the results in Lemma A4, and the other is a bound based on Markov's inequality used in Lemma 3.

as in Lemma 1. Lemma A3 gives $\|(B.15)\| \leq M_{1,Z} L_G \varepsilon \|\theta_b - \hat{\theta}_n\|$. For the last term:

$$\begin{aligned}\|G_{n,\varepsilon}(\theta_b) - G_\varepsilon(\theta_b)\| &\leq \|[G_{n,\varepsilon}(\theta_b) - G_\varepsilon(\theta_b)] - [G_{n,\varepsilon}(\theta^\dagger) - G_\varepsilon(\theta^\dagger)]\| \\ &\quad + \|G_{n,\varepsilon}(\theta^\dagger) - G_\varepsilon(\theta^\dagger) - \frac{1}{\varepsilon}[\overline{g}_n(\theta^\dagger) - g(\theta^\dagger)]\| + \frac{\|\overline{g}_n(\theta^\dagger)\|}{\varepsilon} \\ &\leq C_\Theta L_g M_{1,Z} R_G^\psi \frac{c_n n^{-1/2}}{\varepsilon} + C_\Theta L_g \varepsilon^{\psi-1} c_n n^{-1/2} + \sqrt{p\lambda_{\max}(\Sigma)} c_n n^{-1/2},\end{aligned}$$

with probability $1 - (1+C)/c_n$. For $\varepsilon \leq 1$ this implies $\|(B.16)\| \leq C_{(B.16)} \frac{c_n n^{-1/2}}{\varepsilon} \|\theta_b - \hat{\theta}_n\|$, with the same probability. Putting everything together, we have:

$$\|\theta_{b+1} - \hat{\theta}_n - (1-\gamma)(\theta_b - \hat{\theta}_n)\| \leq \gamma \frac{\sqrt{\kappa_W} L_G}{\underline{\sigma}_{n,\varepsilon}} \left(\|\theta_b - \hat{\theta}_n\| + M_{1,Z}\varepsilon \right) \|\theta_b - \hat{\theta}_n\| + \gamma \Delta_{n,\varepsilon}, \quad \text{(B.17)}$$

with probability $1 - (1+C)/c_n$ where, after some simplifications:

$$\Delta_{n,\varepsilon} \leq \frac{C_2}{\underline{\sigma}_{n,\varepsilon}^2} \left(\Gamma_{n,\varepsilon} + \frac{(c_n n^{-1/2})^2}{\varepsilon} \|\theta_b - \hat{\theta}_n\|^\psi + \frac{c_n n^{-1/2}}{\varepsilon} \|\theta_b - \hat{\theta}_n\| \right).$$

□

Proof of Theorem 1 If (9) holds, then Proposition 1 implies that for $\|\theta_0 - \hat{\theta}_n\| \leq R_{n,\varepsilon}$, with probability $1 - (1+C)/c_n$, we have by recursion for all $b \geq 0$:

$$\|\theta_{b+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})\|\theta_b - \hat{\theta}_n\| + \gamma \Delta_{n,\varepsilon}(R_{n,\varepsilon}) \leq R_{n,\varepsilon},$$

since $\Delta_{n,\varepsilon}$ is increasing so that we have stability of the contraction property.

Let $a_n = (c_n n^{-1/2})^2/\varepsilon$ and $b_n = (c_n n^{-1/2})/\varepsilon$. Let $x_b = \|\theta_b - \hat{\theta}_n\|$, then equation (8) can be re-written as:

$$x_{b+1} \leq (1-\overline{\gamma}) x_b + \gamma \frac{C_2}{\underline{\sigma}_{n,\varepsilon}^2} \left(\Gamma_{n,\varepsilon} + a_n x_b^\psi + b_n x_b \right).$$

For $\psi = 1$, condition (10) implies that $a_n + b_n \leq \tau\overline{\gamma}$ and:

$$x_{b+1} \leq (1 - \overline{\gamma} + \tau\overline{\gamma}) x_b + \gamma \frac{C_2}{\underline{\sigma}_{n,\varepsilon}^2} \Gamma_{n,\varepsilon},$$

from which (11) immediately follows. For $\psi < 1$, using $a_n x_b^\psi = (a_n x_b^{\psi-1}) x_b$, either:

- x_b is such that $x_b \geq a_n^{\frac{1}{1-\psi}} C_{n,\varepsilon}^{1/\psi}$. Then we have $1 - \overline{\gamma} + \gamma \frac{C_2}{\underline{\sigma}_{n,\varepsilon}^2}(b_n + a_n x_b^{\psi-1}) \leq 1 - \overline{\gamma} + \tau\overline{\gamma}$ which implies:

$$x_{b+1} \leq (1 - \overline{\gamma} + \tau\overline{\gamma}) x_b + \gamma \frac{C_2}{\underline{\sigma}_{n,\varepsilon}^2} \Gamma_{n,\varepsilon},$$

- or x_b is such that $x_b \leq a_n^{\frac{1}{1-\psi}} C_{n,\varepsilon}^{1/\psi}$. Then we have $1-\overline{\gamma}+\gamma\frac{C_2}{\underline{\sigma}^2_{n,\varepsilon}} b_n \leq 1-\overline{\gamma}+\tau\overline{\gamma}$, which implies:

$$x_{b+1} \leq (1-\overline{\gamma}+\tau\overline{\gamma})x_b + \gamma\frac{C_2}{\underline{\sigma}^2_{n,\varepsilon}}\left(\Gamma_{n,\varepsilon} + C_{n,\varepsilon} a_n^{\frac{1}{1-\psi}}\right).$$

Majoring these two inequalities yields (11) for $\psi < 1$. Take $b \geq \frac{\log(\Gamma_{n,\varepsilon}) - \log(\|\theta_0 - \hat{\theta}_n\|)}{\log(1-\overline{\gamma}+\tau\overline{\gamma})}$, then

$$(1-\overline{\gamma})^b \|\theta_0 - \hat{\theta}_n\| \leq \Gamma_{n,\varepsilon} \Rightarrow \|\theta_b - \hat{\theta}_n\| \leq (1+\gamma\frac{C_2}{\underline{\sigma}^2_{n,\varepsilon}})\Gamma_{n,\varepsilon} + \gamma\frac{C_2}{\underline{\sigma}^2_{n,\varepsilon}} C_{n,\varepsilon} a_n^{\frac{1}{1-\psi}},$$

with probability $1-(1+C)/c_n$. If $\varepsilon = o(1)$ and $\sqrt{n}\varepsilon \to \infty$, then $\sqrt{n}\Gamma_{n,\varepsilon} = o(1)$ and $\sqrt{n} a_n^{\frac{1}{1-\psi}} = o(1)$ for $c_n = O(1)$ which implies the desired result. □

Proof of Theorem 2: The main steps broadly follow those in Lemma 2 with some additional terms that account for non-smoothness and sampling uncertainty. In the following Lemmas A7, A8, A9, and A10 will be applied with $\eta = \overline{\gamma}/2 \in (0,1)$.

Take $k_n \geq 1$ such that: $\inf_{0\leq j\leq k_n} \sup_{\theta-\theta^j} \|\theta - \theta^j\|_{G'W_nG} \leq \underline{r}_{n,g}(\overline{\gamma}/2)$, defined in Lemma A8. As in the proof of Lemma 2 assume, without loss of generality, that for $b = k_n$ we have $\|\hat{\theta}_n - \theta_b\| \leq \underline{r}_{n,g}(\overline{\gamma}/2)$. Using $\underline{r}_{n,g}(\overline{\gamma}/2) \leq \underline{\sigma}\underline{\lambda}_W^{1/2} R_n(\overline{\gamma}/2)$ we also have $\|\theta_b - \hat{\theta}_n\| \leq R_n(\overline{\gamma}/2)$ so that Lemmas A7 and A9 can be applied. Let $x_b = \|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}$, we have:

$$x_{b+1} \leq (1-\overline{\gamma})x_b + \Delta_{2,n,\varepsilon}(x_b), \tag{B.18}$$

with probability $1-(1+C)/c_n$. Using the reverse triangular inequality, this implies that:

$$\|\overline{g}_n(\theta_{b+1})\|_{W_n} \leq (1-\overline{\gamma})x_b + \Delta_{2,n,\varepsilon}(x_b) + \|\overline{g}_n(\hat{\theta}_n)\|_{W_n},$$

with probability $1-(1+C)/c_n$. And now Lemma A7 applied to x_b yields:

$$\begin{aligned}\|\overline{g}_n(\theta_{b+1})\|_{W_n} \leq\ & (1+\overline{\gamma}/2)(1-\overline{\gamma})\underline{r}_{n,g}(\overline{\gamma}/2) \\ & + \Delta_{2,n,\varepsilon}(x_b) + \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi},\end{aligned} \tag{B.19}$$

with probability $1-(1+C)/c_n$. Suppose θ^{b+1} is such that:

$$\|\overline{g}_n(\theta^{b+1})\|_{W_n} \leq \|\overline{g}_n(\theta_{b+1})\|_{W_n},$$

then θ^{b+1} satisfies the same inequality (B.19). Notice that $(1+\overline{\gamma}/2)(1-\overline{\gamma}) < 1$, $\liminf_{n\to\infty} \underline{r}_{n,g}(\overline{\gamma}/2) > 0$, since $\overline{\gamma}$ is fixed but $(\Delta_{2,n,\varepsilon}(x_b) + \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi}) \leq (\Delta_{2,n,\varepsilon}(x_b) + C_c c_n n + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi})$ for which the $\limsup_{n\to\infty}$ is zero when $c_n n^{-1/2} \to 0$, $\varepsilon \to 0$,

and $\sqrt{n}\varepsilon/c_n \to \infty$. This implies that for $c_n n^{-1/2}$, ε small enough and $\sqrt{n}\varepsilon/c_n$ large enough, we have:

$$\|\overline{g}_n(\theta^{b+1})\|_{W_n} \leq \underline{r}_{n,g}(\overline{\gamma}/2),$$

with probability $1-(1+C)/c_n$. Now apply Lemma A8 to find that it implies $\|\theta^{b+1} - \hat{\theta}_n\|_{G'W_nG} \leq \overline{r}_{n,g}(\overline{\gamma}/2)$ with the same probability, which in turn implies $\|\theta^{b+1} - \hat{\theta}_n\| \leq R_n(\overline{\gamma}/2)$. Lemma A7 now applies to θ^{b+1}:

$$(1-\overline{\gamma}/2)\|\theta^{b+1} - \hat{\theta}_n\|_{G'W_nG} \leq \|\overline{g}_n(\theta^{b+1}) - \overline{g}_n(\hat{\theta}_n)\|_{W_n} + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi}, \tag{B.20}$$

again with probability $1-(1+C)/c_n$. The main issue here is that $\|\overline{g}_n(\theta^{b+1})\|_{W_n} \leq \|\overline{g}_n(\theta_{b+1})\|_{W_n}$ does not directly imply an ordering between $\|\overline{g}_n(\theta^{b+1}) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}$ and x_{b+1}. Pick $\tau \in (0,1)$; for $c_n n^{-1/2}$ small enough, Lemma A10 applies and we have:

$$(1-\tau)\|\overline{g}_n(\theta^{b+1}) - \overline{g}_n(\hat{\theta}_n)\|_{W_n} \leq (1+\tau)\|\overline{g}_n(\theta_{b+1}) - \overline{g}_n(\hat{\theta}_n)\|_{W_n} + \sqrt{2}\Gamma_{2,n,\varepsilon}(\overline{\gamma}/2, \tau),$$

with probability $1-(1+C)/c_n$. Plug this back into (B.18) to find:

$$\begin{aligned}(1-\tau)(1-\overline{\gamma}/2)\|\theta^{b+1} - \hat{\theta}_n\|_{G'W_nG} \leq{}& (1+\tau)(1+\overline{\gamma}/2)(1-\overline{\gamma})\underline{r}_{n,g}(\overline{\gamma}/2) \\ &+ \sqrt{2}\Gamma_{2,n,\varepsilon}(\overline{\gamma}/2,\tau) + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi} + \Delta_{2,n,\varepsilon}(x_b),\end{aligned}$$

with probability $1-(1+C)/c_n$. When $\tau \in (0,1)$ is small enough that:

$$\frac{1+\tau}{1-\tau}\frac{1+\overline{\gamma}/2}{1-\overline{\gamma}/2}(1-\overline{\gamma}) < 1,$$

then using the same arguments as above, we have for $c_n c^{-1/2}$, ε small enough and $\sqrt{n}\varepsilon/c_n$ large enough that $\|\theta^{b+1} - \hat{\theta}_n\|_{G'W_nG} \leq \underline{r}_{n,g}(\overline{\gamma}/2)$ with probability $1-(1+C)/c_n$. This implies that with the same probability $\|\theta_{b+j} - \hat{\theta}_n\| \leq \underline{r}_{n,g}(\overline{\gamma}/2)$, for all $j \geq 0$. Combine (B.18) and (B.20) to find that even after setting $\theta_{b+1} = \theta^{b+1}$ we have:

$$x_{b+1} \leq \frac{1+\tau}{1-\tau}(1-\overline{\gamma})x_b + \Delta_{2,n,\varepsilon}(x_b) + \Gamma_{2,n,\varepsilon}(\overline{\gamma}/2,\tau) + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi}, \tag{B.21}$$

with probability $1-(1+C)/c_n$, where:

$$1 - \tilde{\gamma} := \frac{1+\tau}{1-\tau}(1-\overline{\gamma}) < 1,$$

by assumption about τ. Now we will use similar steps as in the proof of Theorem 1 to bound $\Delta_{2,n,\varepsilon}(x_b)$ and get a $(1-\tilde{\gamma}+\tau\tilde{\gamma})$ convergence rate. Set $a_n = [c_n n^{-1/2} + (c_n n^{-1/2})^2 \varepsilon^{-1}]$ and $b_n = c_n n^{-1/2}$. We have:

$$\Delta_{2,n,\varepsilon}(x_b) = C_4\left(\Gamma_{n,\varepsilon} + \varepsilon^{\psi-1}(c_n n^{-1/2})^2 + a_n x_b^\psi + b_n x_b\right).$$

If $\psi = 1$ and $a_n + b_n \leq \tau\tilde{\gamma}$, then we get:

$$x_{b+1} \leq (1 - \tilde{\gamma} + \tau\tilde{\gamma})x_b + C_4\left(\Gamma_{n,\varepsilon} + \varepsilon^{\psi-1}(c_n n^{-1/2})^2\right) + \Gamma_{2,n,\varepsilon}(\overline{\gamma}/2, \tau) + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi},$$

with probability $1 - (1 + C)/c_n$, as desired. If $\psi < 1$, as in the proof of Theorem 1 there are two cases:

- $x_b \geq a_n^{\frac{1}{1-\psi}} \left(\frac{\tau\tilde{\gamma} - b_n}{C_4}\right)^{\frac{1}{\psi-1}}$, then $a_n x_b^\psi + b_n x_b \leq \tau\tilde{\gamma} x_b$,
- $x_b \leq a_n^{\frac{1}{1-\psi}} \left(\frac{\tau\tilde{\gamma} - b_n}{C_4}\right)^{\frac{1}{\psi-1}}$, then $[a_n^{1+\frac{\psi}{1-\psi}} + b_n a_n^{\frac{1}{1-\psi}}] \max(C_n(\tau), C_n(\tau)^\psi)$, where $C_n(\tau) = \left(\frac{\tau\tilde{\gamma} - b_n}{C_4}\right)^{\frac{1}{\psi-1}}$.

Set $C_n(\tau) = 0$ when $\psi = 1$ to find a common upper-bound:

$$\begin{aligned} x_{b+1} \leq (1 - \tilde{\gamma} + \tau\tilde{\gamma})x_b + C_4\left(\Gamma_{n,\varepsilon} + \varepsilon^{\psi-1}(c_n n^{-1/2})^2\right) + \Gamma_{2,n,\varepsilon}(\overline{\gamma}/2, \tau) + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi} \\ + [a_n^{1+\frac{\psi}{1-\psi}} + b_n a_n^{\frac{1}{1-\psi}}] \max(C_n(\tau), C_n(\tau)^\psi), \end{aligned}$$

with probability $1 - (1 + C)/c_n$. Iterate the above inequality over $j = 0, \ldots$ to find for $b = k_n + j$:

$$\begin{aligned} x_{b+j} \leq (1 - \tilde{\gamma} + \tau\tilde{\gamma})^j x_{k_n} + \frac{1}{(1-\tau)\tilde{\gamma}}\Big[C_4\left(\Gamma_{n,\varepsilon} + \varepsilon^{\psi-1}(c_n n^{-1/2})^2\right) + \Gamma_{2,n,\varepsilon}(\overline{\gamma}/2, \tau) \\ + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi} + [a_n^{1+\frac{\psi}{1-\psi}} + b_n a_n^{\frac{1}{1-\psi}}] \max(C_n(\tau), C_n(\tau)^\psi)\Big], \end{aligned}$$

with probability $1 - (1 + C)/c_n$. Apply Lemma A7 to both sides of the inequality:

$$\begin{aligned} \|\theta_b - \hat{\theta}_n\|_{G'W_nG} \leq (1 - \tilde{\gamma} + \tau\tilde{\gamma})^j \frac{1 + \overline{\gamma}/2}{1 - \overline{\gamma}/2} \underline{r}_{n,g}(\overline{\gamma}/2) \\ + \frac{(1 - \overline{\gamma}/2)^{-1}}{(1-\tau)\tilde{\gamma}}\Big[C_4\left(\Gamma_{n,\varepsilon} + \varepsilon^{\psi-1}(c_n n^{-1/2})^2\right) + \Gamma_{2,n,\varepsilon}(\overline{\gamma}/2, \tau) \\ + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi} + [a_n^{1+\frac{\psi}{1-\psi}} + b_n a_n^{\frac{1}{1-\psi}}] \max(C_n(\tau), C_n(\tau)^\psi)\Big] \\ + 2(1 - \overline{\gamma}/2)^{-1} \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi}, \end{aligned}$$

with probability $1 - (1 + C)/c_n$. The term $\tilde{\Gamma}_{n,\varepsilon}$ can be derived from this equation and some simplifications involving $\Gamma_{2,n,\varepsilon}$. Just as in the proof of Theorem 1 we have for

$$j \geq \frac{\log(\Gamma_{n,\varepsilon}) - \log\left(\frac{1+\overline{\gamma}/2}{1-\overline{\gamma}/2}\underline{r}_{n,g}(\overline{\gamma}/2)\right)}{\log(1 - \tilde{\gamma} + \tau\tilde{\gamma})},$$

that the leading term is less than $\Gamma_{n,\varepsilon} = o(n^{-1/2})$. Also, under the stated assumptions: $\Gamma_{2,n,\varepsilon}(\overline{\gamma}/2, \tau) = o(n^{-1/2})$, $a_n^{1+\frac{\psi}{1-\psi}} = o(n^{-1/2})$, $b_n = o(n^{-1/2})$, and $b_n a_n^{\frac{1}{1-\psi}} = o(n^{-1/2})$ which implies the desired result: $\sqrt{n}\|\theta_b - \hat{\theta}_n\|_{G'W_nG} = o_p(1)$ and concludes the proof. □

Supplement to

"Noisy, Non-Smooth, Non-Convex Estimation of Moments Conditions Models"

Jean-Jacques Forneron*

October 20, 2022

This Supplemental Material consists of Appendices C, D, E, F, G, and H to the main text.

*Department of Economics, Boston University, 270 Bay State Rd, MA 02215 Email: jjmf@bu.edu

Appendix C Proofs for the Preliminary Results

C.1 Preliminary Results for the Population Problem

Proof of Lemma A1. First, we will derive (3). Take θ_b such that $\|\theta_b - \theta^\dagger\| \leq R$, as defined in Lemma 1, and let $G_b = G(\theta_b)$, we have for some intermediate value $\tilde{\theta}_b$ between θ_b and θ_{b+1}:[1]

$$\begin{aligned} g(\theta_{b+1}) - (1-\gamma)g(\theta_b) &= \gamma\left(I_d - G(\tilde{\theta}_b)(G_b'WG_b)^{-1}G_b'W\right)g(\theta_b) \\ &= \gamma\left(I_d - G_b(G_b'WG_b)^{-1}G_b'W\right)g(\theta_b) \qquad \text{(C.22)} \\ &+ \gamma\left(G_b - G(\tilde{\theta}_b)\right)(G_b'WG_b)^{-1}G_b'Wg(\theta_b). \qquad \text{(C.23)} \end{aligned}$$

To bound (C.22), notice that $[I_d - G_b(G_b'WG_b)^{-1}G_b'W]G_b = 0$ so that:

$$\begin{aligned} (C.22) &= \gamma\left(I_d - G_b(G_b'WG_b)^{-1}G_b'W\right)[g(\theta_b) - G_b(\theta_b - \theta^\dagger)] \\ &= \gamma\left(I_d - G_b(G_b'WG_b)^{-1}G_b'W\right)[G(\bar{\theta}_b) - G_b](\theta_b - \theta^\dagger), \end{aligned}$$

for another intermediate value $\bar{\theta}_b$ between θ_b and $\theta^\dagger$. By Lipschitz continuity:

$$\|(C.22)\| \leq \gamma L_G\|\theta_b - \theta^\dagger\|^2 \leq \left(\frac{\gamma L_G}{\underline{\sigma}}\|\theta_b - \theta^\dagger\|\right)\|g(\theta_b)\|,$$

using $\|g(\theta_b)\| = \|G(\bar{\theta}_b)(\theta_b - \theta^\dagger)\| \geq \underline{\sigma}\|\theta_b - \theta^\dagger\|$. To bound (C.23), Lemma 1 implies that $\|\tilde{\theta}_b - \theta^\dagger\| \leq \|\theta_b - \theta^\dagger\|$, then using the Lipschitz continuity of G:

$$\|(C.23)\| \leq \left(\frac{\gamma L_G\kappa_W}{\underline{\sigma}}\|\theta_b - \theta^\dagger\|\right)\|g(\theta_b)\|.$$

Putting together the bounds, we have:

$$\begin{aligned} \|g(\theta_{b+1})\|_W &\leq \left(1 - \gamma + \frac{\gamma\overline{\lambda}_W L_G}{\underline{\sigma}}[\kappa_W + 1]\|\theta_b - \theta^\dagger\|\right)\|g(\theta_b)\|_W \\ &\leq (1-\overline{\gamma})\|g(\theta_b)\|_W, \quad \text{if } \|\theta_b - \theta^\dagger\| \leq \underline{\sigma}\frac{\gamma - \overline{\gamma}}{\gamma\overline{\lambda}_W L_G[\kappa_W + 1]}. \end{aligned}$$

This is the desired result (3). For (4), note that a mean-value expansion $\|g(\theta)\|_W = \|G(\tilde{\theta})(\theta - \theta^\dagger)\|_W$ and the reverse triangular inequality imply:

$$\begin{aligned} \|g(\theta)\|_W &\geq \|G(\theta^\dagger)(\theta - \theta^\dagger)\|_W - \|[G(\theta^\dagger) - G(\tilde{\theta})](\theta - \theta^\dagger)\|_W \\ &\geq \|\theta - \theta^\dagger\|_{G'WG} - \overline{\lambda}_W L_G\|\theta - \theta^\dagger\|^2 \geq \left(1 - \frac{\overline{\lambda}_W^2 L_G}{\underline{\sigma}}\|\theta - \theta^\dagger\|\right)\|\theta - \theta^\dagger\|_{G'WG}, \end{aligned}$$

[1] First notice that: $\theta_{b+1} - \theta_b = -\gamma(B_b'WG_b)^{-1}G_b'Wg(\theta_b)$. Then, the mean-value theorem implies: $g(\theta_{b+1}) - g(\theta_b) = G(\tilde{\theta}_b)(\theta_{b+1} - \theta_b)$.

pick $\|\theta - \theta^\dagger\| \leq \frac{\overline{\gamma}}{2}\frac{\underline{\sigma}}{\overline{\lambda}_W^2 L_G}$ to get the lower bound in (4). Repeat the same steps with the triangular inequality to get the upper bound. The $\overline{r}_g$ can then be explicitly derived from the inequalities above using the norm equivalence between $\|\cdot\|$ and $\|\cdot\|_{G'WG}$. Finally, for (5) the global identification condition, Assumption 1v, implies

$$\inf_{\|\theta-\theta^\dagger\|_{G'WG}\geq\overline{r}_g} \|g(\theta)\|_W \geq \inf_{\|\theta-\theta^\dagger\|\geq\underline{\sigma}\sqrt{\underline{\lambda}_W}\overline{r}_g} \|g(\theta)\|_W = \delta(\underline{\sigma}\sqrt{\underline{\lambda}_W}\overline{r}_g),$$

and set $\underline{\delta} = \delta(\underline{\sigma}\sqrt{\underline{\lambda}_W}\overline{r}_g) > 0$. The local norm equivalence (4) implies for $\|\theta - \theta^\dagger\|_{G'WG} \leq \underline{r}_g \leq \overline{r}_g$ that we have $\|g(\theta)\|_W \leq (1+\overline{\gamma}/2)(1-\overline{\gamma})\underline{r}_g \leq \underline{\delta}$ if $\underline{r}_g \leq \min(\frac{\underline{\delta}}{(1+\overline{\gamma}/2)(1-\overline{\gamma})}, \overline{r}_g)$ which yields the desired result and concludes the proof. □

C.2 Preliminary Results for the Finite-Sample Problem

Proof of Lemma A2

i. By construction $\overline{g}_{n,\varepsilon}(\theta) = \int_Z \overline{g}_n(\theta + \varepsilon Z)\phi(Z)dZ$, applying the change of variable $u = \theta + \varepsilon Z$, we get the identity: $\int_Z \overline{g}_n(\theta + \varepsilon Z)\phi(Z)dZ = \varepsilon \int_u \overline{g}_n(u)\phi(\frac{u-\theta}{\varepsilon})du$,

ii. Applying Leibniz's rule to the first identity, we get: $\partial_\theta \overline{g}_{n,\varepsilon}(\theta) = -\int_u \overline{g}_n(u)\phi'(\frac{u-\theta}{\varepsilon})du$. Apply the change of variable $u = \theta + \varepsilon Z$ to get the second identity: $\partial_\theta \overline{g}_{n,\varepsilon}(\theta) = -\frac{1}{\varepsilon}\int_Z \overline{g}_n(\theta + \varepsilon Z)\phi'(Z)dZ$,

iii. The first part follows from the same derivations as above, the second follows from Liebniz's rule applied directly to $G_\varepsilon(\theta)$.

□

Proof of Lemma A3

i. We can write: $G_\varepsilon(\theta) = \int_Z G(\theta + \varepsilon Z)\phi(Z)dZ$. Now pick any two $(\theta_1, \theta_2) \in \Theta$, we have:

$$\|G_\varepsilon(\theta_1) - G_\varepsilon(\theta_2)\| \leq \int_Z \|G(\theta_1 + \varepsilon Z) - G(\theta_2 + \varepsilon Z)\|\phi(Z)dZ \leq L_G\|\theta_1 - \theta_2\|.$$

ii. Using the definition of G_ε and G, we have:

$$\|G_\varepsilon(\theta) - G(\theta)\| = \|\int_Z [G(\theta + \varepsilon Z) - G(\theta)]\phi(Z)dZ\| \leq \varepsilon L_G \int_Z \|Z\|\phi(Z)dZ = \varepsilon L_G M_{1,Z}.$$

□

Proof of Lemma A4. The first inequality is a consequence of Theorem 2.14.2 in van der Vaart and Wellner (1996).

a. Inequality a. follows from Markov's inequality and the first inequality.

b. For inequality b., note that for any $\|\theta_1 - \theta_2\| \leq \delta$ we have:

$$\begin{aligned}
&\|[\overline{g}_{n,\varepsilon}(\theta_1) - g_\varepsilon(\theta_1)] - [\overline{g}_{n,\varepsilon}(\theta_2) - g_\varepsilon(\theta_2)]\| \\
&\quad = \| \int_Z \left([\overline{g}_n(\theta_1 + \varepsilon Z) - g(\theta_1 + \varepsilon Z)] - [\overline{g}_n(\theta_2 + \varepsilon Z) - g(\theta_2 + \varepsilon Z)]\right) \phi(Z) dZ \| \\
&\quad \leq \int_Z \sup_{\|\theta_1 - \theta_2\| \leq \delta} \|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\| \phi(Z) dZ \\
&\quad = \sup_{\|\theta_1 - \theta_2\| \leq \delta} \|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\|,
\end{aligned}$$

since $\int \phi(Z) dZ = 1$. Then inequality a. yields the desired result.

c. For inequality c., note that for any $\|\theta_1 - \theta_2\| \leq \delta$ we have:

$$\begin{aligned}
&\|[\overline{G}_{n,\varepsilon}(\theta_1) - G_\varepsilon(\theta_1)] - [\overline{G}_{n,\varepsilon}(\theta_2) - G_\varepsilon(\theta_2)]\| \\
&\quad = \| -\frac{1}{\varepsilon} \int_Z \left([\overline{g}_n(\theta_1 + \varepsilon Z) - g(\theta_1 + \varepsilon Z)] - [\overline{g}_n(\theta_2 + \varepsilon Z) - g(\theta_2 + \varepsilon Z)]\right) \phi'(Z) dZ \| \\
&\quad \leq \frac{1}{\varepsilon} \int_Z \sup_{\|\theta_1 - \theta_2\| \leq \delta} \|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\| \times \|\phi'(Z)\| dZ \\
&\quad = \sup_{\|\theta_1 - \theta_2\| \leq \delta} \|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\| \frac{M_{1,Z}}{\varepsilon},
\end{aligned}$$

where $M_{1,Z} = \int \|\phi'(Z)\| dZ$, the result then follows from inequality a.

For the final statement, simply note that all three derivations require bounding the same supremum, i.e. the same event. □

Proof of Lemma A5. Note that $\overline{G}_{n,\varepsilon}(\theta) = \overline{G}_{n,\varepsilon}(\theta) - G(\theta) + G(\theta)$, using Weyl's inequality for singular values (see Problem III.6.5 in Bhatia, 2013), we have:

$$\sigma_{\min}[G(\theta)] \leq \sigma_{\min}[\overline{G}_{n,\varepsilon}(\theta)] + \sigma_{\min}[\overline{G}_{n,\varepsilon}(\theta) - G(\theta)],$$

which implies:

$$\sigma_{\min}[\overline{G}_{n,\varepsilon}(\theta)] \geq \underline{\sigma} - \sigma_{\min}[\overline{G}_{n,\varepsilon}(\theta) - G(\theta)],$$

since $\sigma_{\min}[G(\theta)] \geq \underline{\sigma}$ for $\|\theta - \theta^\dagger\| \leq R_G$. Note that $\sigma_{\min}[\overline{G}_{n,\varepsilon}(\theta) - G(\theta)] \leq \|\overline{G}_{n,\varepsilon}(\theta) - G(\theta)\|$. By the triangular inequality and Lemmas A3, A4:

$$\begin{aligned}\|\overline{G}_{n,\varepsilon}(\theta) - G(\theta)\| &\leq \|\overline{G}_{n,\varepsilon}(\theta) - G_\varepsilon(\theta)\| + \|G_\varepsilon(\theta) - G(\theta)\| \\ &\leq \frac{c_n}{\varepsilon\sqrt{n}} C_\Theta L_g M_{1,Z} R_G^\psi + \varepsilon L_G M_{1,Z},\end{aligned}$$

with probability $1 - C/c_n$. Putting everything together yields the desired result. □

Proof of Lemma A6: The result holds for $\delta = 0$, so we can work with $\delta > 0$. Take any two θ_1, θ_2; we have:

$$\begin{aligned}\|[\overline{g}_n(\theta_1) - \overline{g}_n(\theta_2)] - G_{n,\varepsilon}(\theta_1)(\theta_1 - \theta_2)\| \leq{}& \|[\overline{g}_n(\theta_1) - \overline{g}_n(\theta_2)] - G(\theta_1)(\theta_1 - \theta_2)\| && \text{(C.24)} \\ &+ \|G_\varepsilon(\theta_1) - G(\theta_1)\| \times \|\theta_1 - \theta_2\| && \text{(C.25)} \\ &+ \|G_{n,\varepsilon}(\theta_1) - G_\varepsilon(\theta_1)\| \times \|\theta_1 - \theta_2\|. && \text{(C.26)}\end{aligned}$$

Using Lemma A4 and the Lipschitz continuity of G, we have:

$$\|(C.24)\| \leq L_g C_\Theta c_n n^{-1/2} \|\theta_1 - \theta_2\|^\psi + L_G \|\theta_1 - \theta_2\|^2,$$

with probability $1 - C/c_n$. Also, we have $\|(C.25)\| \leq \varepsilon L_G M_{1,Z} \|\theta_1 - \theta_2\|$. Using the integral representation, we have - after a mean-zero adjustement:

$$G_{n,\varepsilon}(\theta_1) - G_\varepsilon(\theta_1) = \frac{1}{\varepsilon} \int ([\overline{g}_n(\theta_1 + \varepsilon Z) - \overline{g}_n(\theta_1)] - [g(\theta_1 + \varepsilon Z) - g(\theta_1)]) Z \phi(Z) dZ,$$

from which we deduce, using Lemma A4 again, that:

$$\|(C.26)\| \leq L_g C_\Theta \left[\int \|Z\|^{1+\psi} \phi(Z) dZ \right] c_n n^{-1/2} \varepsilon^{\psi - 1} \|\theta_1 - \theta_2\|,$$

with probability $1 - C/c_n$. Combine the three bounds together to get the desired result. Note that $\int \|Z\|^{1+\psi} \phi(Z) dZ \leq \int \|Z\|^2 \phi(Z) dZ = M_{2,Z}$ □

Proof of Lemma A7: The proof will only focus on the lower bound, since the derivations are similar for the upper bound. Using the reverse triangular inequality:

$$\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n} \geq \|\theta - \hat{\theta}_n\|_{G' W_n G} - \|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n) - G(\theta^\dagger)(\theta - \hat{\theta}_n)\|_{W_n}.$$

Using Lemma A4 and the Lipschitz continuity of G, we have:

$$\begin{aligned}&\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n) - G(\theta^\dagger)(\theta - \hat{\theta}_n)\|_{W_n} \\ &\leq \overline{\lambda}_W \left(L_g C_\Theta c_n n^{-1/2} \|\theta - \hat{\theta}_n\|^\psi + L_G \|\theta - \hat{\theta}_n\| \left[\|\theta - \hat{\theta}_n\| + C_a c_n n^{-1/2} \right] \right),\end{aligned}$$

with probability $1-(1+C)/c_n$. There are two cases. If $\|\theta-\hat{\theta}_n\| \leq c_n n^{-1/2}$, then:

$$\begin{aligned}
&\|\overline{g}_n(\theta)-\overline{g}_n(\hat{\theta}_n)-G(\theta^\dagger)(\theta-\hat{\theta}_n)\|_{W_n} \\
&\leq \underline{\sigma}^{-2}\kappa_W L_G[\|\theta-\hat{\theta}_n\|+C_a c_n n^{-1/2}]\|\theta-\hat{\theta}_n\|_{G'W_nG}+\overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi} \\
&\leq \eta\|\theta-\hat{\theta}_n\|_{G'W_nG}+\overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi},
\end{aligned}$$

if $\|\theta-\hat{\theta}_n\| \leq R_n(\eta)$. Otherwise, when $\|\theta-\hat{\theta}_n\| \geq c_n n^{-1/2}$ then

$$\begin{aligned}
&\|\overline{g}_n(\theta)-\overline{g}_n(\hat{\theta}_n)-G(\theta^\dagger)(\theta-\hat{\theta}_n)\|_{W_n} \\
&\leq \underline{\sigma}^{-2}\kappa_W L_G[\|\theta-\hat{\theta}_n\|+C_a c_n n^{-1/2}+L_g C_\Theta (c_n n^{-1/2})^{\psi}]\|\theta-\hat{\theta}_n\|_{G'W_nG} \\
&\leq \eta\|\theta-\hat{\theta}_n\|_{G'W_nG},
\end{aligned}$$

if $\|\theta-\hat{\theta}_n\| \leq R_n(\eta)$. Combining the bounds for a given value of η yields the desired results. □

Proof of Lemma A8 For any $r>0$,

$$\|\theta-\hat{\theta}_n\|_{G'W_nG} \geq r \Rightarrow \|\theta-\theta^\dagger\| \geq \frac{r}{\overline{\sigma}\overline{\lambda}_W^{1/2}}-\|\theta^\dagger-\hat{\theta}_n\|.$$

Pick $r=\underline{\sigma}\underline{\lambda}_W^{1/2}R_n(\eta)$, this implies:

$$\|\theta-\theta^\dagger\| \geq \frac{R_n(\eta)}{\sqrt{\kappa_G\kappa_W}}-\|\theta^\dagger-\hat{\theta}_n\| \geq \frac{R_n(\eta)}{\sqrt{\kappa_G\kappa_W}}-C_a c_n n^{-1/2} := \overline{r}_{n,g}(\eta),$$

with probability $1-1/c_n$, where C_a is defined in the proof of Lemma 3. Notice that $\overline{r}_{n,g}(\eta) \to \overline{r}_g(\eta) > 0$ as $n \to \infty$. Making use of (reverse) triangular inequalities, we have with probability $1-(1+C)/c_n$:

$$\begin{aligned}
\inf_{\|\theta-\hat{\theta}_n\|_{G'W_nG} \geq \underline{\sigma}\underline{\lambda}_W^{1/2}R_n(\eta)} \|\overline{g}_n(\theta)\|_{W_n} &\geq \inf_{\|\theta-\theta^\dagger\| \geq \overline{r}_{n,g}(\eta)} \|\overline{g}_n(\theta)\|_{W_n} \\
&\geq \frac{\delta(\overline{r}_{n,g}(\eta))}{\sqrt{\kappa_W}}-\overline{\lambda}_W^{1/2}c_n n^{-1/2}\left(L_g C_\Theta \mathrm{diam}(\Theta)^\psi+\lambda_{\max}(\Sigma)^{1/2}p\right) \\
&:= \underline{r}_{n,g}(\eta),
\end{aligned}$$

where the last terms are derived using Lemma A4 and the bound for $\|\overline{g}_n(\theta^\dagger)\|$ in the proof of Lemma 3. By continuity of δ, $\underline{r}_{n,g}(\eta) \to \underline{r}_g(\eta) > 0$ as $n \to \infty$ and $c_n n^{-1/2} \to 0$. □

Proof of Lemma A9: Take θ_b as described in the Lemma, recall that we have $\theta_{b+1} - \theta_b = -\gamma(G_b' W_n G_b)^{-1} G_b' W_n \overline{g}_n(\theta_b)$, where $G_b = G_{n,\varepsilon}(\theta_b)$. The main idea will be to use the identity:

$$\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 = \|\overline{g}_n(\theta)\|_{W_n}^2 - \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 - 2\left(\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\right)' W_n \overline{g}_n(\hat{\theta}_n), \quad \text{(C.27)}$$

in tandem with the following inequality:

$$\|\overline{g}_n(\theta_{b+1}) - \overline{g}_n(\hat{\theta}_n)\|_{W_n} \leq (1-\gamma)\|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)\|_{W_n} \quad \text{(C.28)}$$

$$+ \|\overline{g}_n(\theta_{b+1}) - \overline{g}_n(\hat{\theta}_n) - (1-\gamma)[\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)]\|_{W_n}. \quad \text{(C.29)}$$

From Lemma A6, we have for some $C_{(A.1)} > 0$:

$$\|\overline{g}_n(\theta_{b+1}) - \overline{g}_n(\theta_b) - G_{n,\varepsilon}(\theta_b)(\theta_{b+1} - \theta_b)\|_{W_n} \leq \overline{\lambda}_W L_G[\|\theta_{b+1} - \theta_b\|^2 + M_{2,Z}\varepsilon\|\theta_{b+1} - \theta_b\|] + \overline{\lambda}_W C_{(A.1)} c_n n^{-1/2}[\|\theta_{b+1} - \theta_b\|^{\psi} + \varepsilon^{\psi-1}\|\theta_{b+1} - \theta_b\|],$$

with probability $1 - C/c_n$. Note that for $\|\theta_b - \hat{\theta}_n\| \leq R_{n,\varepsilon}$, we have:

$$\|\theta_{b+1} - \theta_b\| \leq \underline{\sigma}_{n,\varepsilon}^{-1}\sqrt{\kappa_W}[\|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)\|_{W_n} + \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}],$$

with probability $1 - C/c_n$. We also have $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \leq C_a c_n n^{-1/2}$, with probability $1 - 1/c_n$. Now notice that:

$$\|(C.29)\|_{W_n} = \|\overline{g}_n(\theta_{b+1}) - \overline{g}_n(\theta_b) + \gamma[\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)]\|_{W_n}$$
$$\leq \|\overline{g}_n(\theta_{b+1}) - \overline{g}_n(\theta_b) - G_{n,\varepsilon}(\theta_{b+1} - \theta_b)\|_{W_n} + \gamma\|(I - P_b)[\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)]\|_{W_n} + \gamma\|P_b\overline{g}_n(\hat{\theta}_n)\|_{W_n},$$

where $P_b = G_b(G_b' W_n G_b)^{-1} G_b' W_n$ is an orthogonal projection matrix. By orthogonality:

$$\begin{aligned}\|(I - P_b)[\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)]\|_{W_n} &= \|(I - P_b)[\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n) - G_b(\theta_b - \hat{\theta}_n)]\|_{W_n} \\ &\leq \|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n) - G_b(\theta_b - \hat{\theta}_n)\|_{W_n} \\ &\leq \overline{\lambda}_W L_G(\|\theta_b - \hat{\theta}_n\|^2 + \varepsilon M_{2,Z}\|\theta_b - \hat{\theta}_n\|) \\ &+ \overline{\lambda}_W C_{(A.1)} c_n n^{-1/2}[\|\theta_b - \hat{\theta}_n\|^{\psi} + \varepsilon^{\psi-1}\|\theta_b - \hat{\theta}_n\|],\end{aligned}$$

with probability $1 - (1+C)/c_n$. Then, using Lemma A4 we have:

$$\|\overline{G}_{n,\varepsilon}(\theta_b) - \overline{G}_{n,\varepsilon}(\hat{\theta}_n)\| \leq L_g C_{\Theta} M_{1,Z} c_n n^{-1/2}\varepsilon^{-1}\|\theta_b - \hat{\theta}_n\|^{\psi} + L_G\|\theta_b - \hat{\theta}_n\|,$$

with probability $1 - (1+C)/c_n$. Together with Lemma 3, this implies that:

$$\|P_g\overline{g}_n(\hat{\theta}_n)\|_{W_n} \leq \underline{\sigma}_{n,\varepsilon}^{-2}\sqrt{\kappa_W}[\Gamma_{n,\varepsilon} + C_a c_n n^{-1/2}(L_g C_{\Theta} M_{1,Z} c_n n^{-1/2}\varepsilon^{-1}\|\theta_b - \hat{\theta}_n\|^{\psi} + L_G\|\theta_b - \hat{\theta}_n\|)],$$

with probability $1-(1+C)/c_n$. Lemma A7 implies that for $\eta \in (0,1)$ and $\|\theta_b - \hat{\theta}_n\| \leq R_n(\eta)$, we have with probability $1-(1+C)/c_n$:

$$\|\theta_b - \hat{\theta}_n\|_{G'W_nG} \leq \frac{\underline{\sigma}_{n,\varepsilon}^{-1}\underline{\lambda}_W^{-1/2}}{1-\eta}\left(\|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)\|_{W_n} + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi}\right).$$

Let $x_b = \|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n)\|_{G'W_nG}$. After grouping terms appropriately, the above inequalities imply that (C.28)-(C.29) can be re-written, for $\|\theta_b - \hat{\theta}_n\| \leq R_n(\eta)$, as:

$$x_{b+1} \leq (1-\gamma)x_b + 3L_G\underline{\sigma}_{n,\varepsilon}^{-1}\sqrt{\kappa_W}\left(\overline{\lambda}_W + \frac{\gamma\underline{\sigma}_{n,\varepsilon}^{-1}\sqrt{\kappa_W}}{(1-\eta)^2}\right)x_b^2 + \overline{\lambda}_W L_G\underline{\sigma}_{n,\varepsilon}^{-1}\sqrt{\kappa_W}\left(1+\frac{\gamma}{1-\eta}\right)\varepsilon x_b$$
$$+ C_{(C.29)}\frac{\underline{\sigma}_{n,\varepsilon}^{-3}}{(1-\eta)^2}\left(\Gamma_{n,\varepsilon} + \varepsilon^{\psi-1}(c_n n^{-1/2})^2 + [c_n n^{-1/2} + (c_n n^{-1/2})^2\varepsilon^{-1}]x_b^\psi + (c_n n^{-1/2})x_b\right),$$

with probability $1-(1+C)/c_n$ when assuming (without loss of generality) that $\overline{\lambda}_W \geq 1$ and $\underline{\sigma}_{n,\varepsilon} \leq 1$. Lemma A7 also implies that:

$$x_b^2 \leq [(1+\eta)\|\theta_b - \hat{\theta}_n\|_{G'W_nG}]x_b + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi} x_b,$$

with probability $1-(1+C)/c_n$. Let $R_{2,n,\varepsilon}(\eta) > 0$ be such that:

$$3(1+\eta)L_G\underline{\sigma}_{n,\varepsilon}^{-1}\sqrt{\kappa_W}\left(\overline{\lambda}_W + \frac{\gamma\underline{\sigma}_{n,\varepsilon}^{-1}\sqrt{\kappa_W}}{(1-\eta)^2}\right)R_{2,n,\varepsilon}(\eta) + \overline{\lambda}_W L_G\underline{\sigma}_{n,\varepsilon}^{-1}\sqrt{\kappa_W}\left(1+\frac{\gamma}{1-\eta}\right) \leq \overline{\gamma} - \gamma,$$

then for $\|\theta_b - \hat{\theta}_n\|_{G'W_nG} \leq R_{2,n,\varepsilon}$ (which is bounded below), we have:

$$x_{b+1} \leq (1-\overline{\gamma})x_b + C_{(A.3)}\frac{\underline{\sigma}_{n,\varepsilon}^{-3}}{(1-\eta)^2}\left(\Gamma_{n,\varepsilon} + \varepsilon^{\psi-1}(c_n n^{-1/2})^2 + [c_n n^{-1/2} + (c_n n^{-1/2})^2\varepsilon^{-1}]x_b^\psi + (c_n n^{-1/2})x_b\right),$$

with probability $1-(1+C)/c_n$ which implies the desired result. □

Proof of Lemma A10: Start with the identity:

$$\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 = \|\overline{g}_n(\theta)\|_{W_n}^2 - \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 - 2\left(\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\right)' W_n\overline{g}_n(\hat{\theta}_n). \quad \text{(C.30)}$$

which holds for any $\theta \in \Theta$. Now use Lemmas 3 and A6 to bound the last term:

$$\begin{aligned}
&\|\left(\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\right)' W_n\overline{g}_n(\hat{\theta}_n)\| \\
&\leq \|\theta - \hat{\theta}_n\| \times \|G_{n,\varepsilon}(\hat{\theta}_n)' W_n\overline{g}_n(\hat{\theta}_n)\| + \overline{\lambda}_W C_a c_n n^{-1/2}\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n) - G_{n,\varepsilon}(\hat{\theta}_n)(\theta - \hat{\theta}_n)\| \\
&\leq \Gamma_{n,\varepsilon}\|\theta - \hat{\theta}_n\| + \overline{\lambda}_W C_a c_n n^{-1/2}\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n) - G_{n,\varepsilon}(\hat{\theta}_n)(\theta - \hat{\theta}_n)\|,
\end{aligned}$$

with probability $1-(1+C)/c_n$. Lemma A6 implies that with probability $1-C/c_n$:

$$\|\overline{g}_n(\theta)-\overline{g}_n(\hat{\theta}_n)-G_{n,\varepsilon}(\hat{\theta}_n)(\theta-\hat{\theta}_n)\| \leq L_g C_\Theta \|\theta-\hat{\theta}_n\|^2 + L_G M_{1,Z}\varepsilon\|\theta-\hat{\theta}_n\| + L_g C_\Theta c_n n^{-1/2}[\|\theta-\hat{\theta}_n\|^\psi + \varepsilon^{\psi-1}\|\theta-\hat{\theta}_n\|].$$

Since $\|\theta-\hat{\theta}_n\| \leq R_n(\eta)$ by assumption, Lemma A7 implies that:

$$\|\theta-\hat{\theta}_n\| \leq \frac{\underline{\sigma}_{n,\varepsilon}^{-1}\underline{\lambda}_W^{-1/2}}{1-\eta}\left(\|\overline{g}_n(\theta)-\overline{g}_n(\hat{\theta}_n)\|_{W_n} + \overline{\lambda}_W L_g C_\Theta (c_n n^{-1/2})^{1+\psi}\right),$$

with probability $1-(1+C)/c_n$. Now group the bounds together to find for $x = \|\overline{g}_n(\theta)-\overline{g}_n(\hat{\theta}_n)\|_{W_n}$:

$$\begin{aligned}
&\|\left(\overline{g}_n(\theta)-\overline{g}_n(\hat{\theta}_n)\right)' W_n \overline{g}_n(\hat{\theta}_n)\| \\
&\leq C_{(C.30)}\frac{\underline{\sigma}_{n,\varepsilon}^{-2}}{(1-\eta)^2}\Big(c_n n^{-1/2}x^2 + [\Gamma_{n,\varepsilon} + (c_n n^{-1/2})^2\varepsilon^{\psi-1}]x + (c_n n^{-1/2})^2 x^\psi \\
&+ (c_n n^{-1/2})^{1+\psi}\Gamma_{n,\varepsilon} + (c_n n^{-1/2})^{2+\psi+\psi^2} + \varepsilon(c_n n^{-1/2})^{2+\psi} + (c_n n^{-1/2})^{3+\psi}\varepsilon^{\psi-1}\Big),
\end{aligned}$$

with probability $1-(1+C)/c_n$. Using the same x as above, we have:[2]

$$\begin{aligned}
&\left|\|\overline{g}_n(\theta)\|_{W_n}^2 - \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 - x^2\right| \leq (2\tau-\tau^2)x^2 \\
&+ C_{(C.31)}\Big[\frac{\underline{\sigma}_{n,\varepsilon}^{-4}}{(1-\eta)^4}\frac{1}{\overline{\gamma}-[\overline{\gamma}/2]^2}\left(\Gamma_{n,\varepsilon}^2 + (c_n n^{-1/2})^4\varepsilon^{2\psi-2} + (c_n n^{-1/2})^{2+2\psi}\right) \qquad \text{(C.31)} \\
&+ \frac{\underline{\sigma}_{n,\varepsilon}^{-2}}{(1-\eta)^2}\left((c_n n^{-1/2})^{1+\psi}\Gamma_{n,\varepsilon} + (c_n n^{-1/2})^{2+\psi+\psi^2} + \varepsilon(c_n n^{-1/2})^{2+\psi} + (c_n n^{-1/2})^{3+\psi}\varepsilon^{\psi-1}\right)\Big] \qquad \text{(C.32)}
\end{aligned}$$

with probability $1-(1+C)/c_n$ if $2C_{(C.30)}\underline{\sigma}_{n,\varepsilon}^{-2}(1-\eta)^{-2}c_n n^{-1/2} \leq (2\tau-\tau^2)/3$, which holds for $c_n n^{-1/2}$ sufficiently small. Set $\Gamma_{2,n,\varepsilon}^2(\eta,\tau) = (C.31)+(C.32)$. Putting everything together, we get:

$$\begin{aligned}
(1-\tau)^2\|\overline{g}_n(\theta)-\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 - \Gamma_{2,n,\varepsilon}^2(\eta,\tau) &\leq \|\overline{g}_n(\theta)\|_{W_n}^2 - \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 \\
&\leq (1+\tau)^2\|\overline{g}_n(\theta)-\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 + \Gamma_{2,n,\varepsilon}^2(\eta,\tau),
\end{aligned}$$

with probability $1-(1+C)/c_n$, for $\|\theta-\hat{\theta}_n\| \leq R_n(\eta)$, or $\|\theta-\hat{\theta}_n\|_{G'W_nG} \leq \underline{\sigma}\underline{\lambda}_W^{1/2}R_n(\eta)$. The result that $n\Gamma_{2,n,\varepsilon}^2(\eta,\tau) = o(1)$ follows from $\sqrt{n}\Gamma_{n,\varepsilon} = o(1)$ under the restrictions on ε. □

[2]Here we will use $\min[1-(1-\tau)^2,(1+\tau)^2-1] \geq 2\tau-\tau^2$ to derive the $(1-\tau)^2$ lower and a $(1+\tau)^2$ upper bounds we want in the Lemma.

Appendix D Preliminary Results for Section 4

Lemma D11. *Suppose Assumptions 1-2 hold. For $L \geq 1$, define:*

$$\hat{G}_L(\theta_b) = \frac{1}{L}\sum_{\ell=0}^{L-1}\frac{1}{\varepsilon}[\overline{g}_n(\theta_{b-\ell}+\varepsilon Z_{b-\ell}) - \overline{g}_n(\theta_{b-\ell})]Z'_{b-\ell}, \quad \tilde{G}_L(\theta_b) = \frac{1}{L}\sum_{\ell=0}^{L-1} G(\theta_{b-\ell})Z_{b-\ell}Z'_{b-\ell}$$

Take c_n and $b_{\max} \geq 1$. Let $t_n = \log(c_n) + \log(b_{\max}+L+1)$, we have:

$$\sup_{0\leq b\leq b_{\max}} \|\hat{G}_L(\theta_b) - \tilde{G}_L(\theta_b)\| \leq \left(L_g C_\Theta c_n n^{-1/2}\varepsilon^{\psi-1} + \varepsilon L_G\right)\left[d_\theta + 2d_\theta t_n + \sqrt{2d_\theta t_n}\right]^{3/2}, \quad \text{(D.33)}$$

with probability $1-(2+C)/c_n$.

Lemma D12. *Suppose Assumptions 1-2 hold. For $L \geq 1$, define:*

$$\tilde{G}_L(\theta_b) = \frac{1}{L}\sum_{\ell=0}^{L-1} G(\theta_{b-\ell})Z_{b-\ell}Z'_{b-\ell}, \quad G_L(\theta_b) = \frac{1}{L}\sum_{\ell=0}^{L-1} G(\theta_{b-\ell})$$

Take c_n and $b_{\max} \geq 1$, let $d = d_\theta + d_g$. Let $t_n = [\log(c_n) + d\log(9) + \log(b_{\max}+1)]/L$ $\tilde{t}_n = C_Z[\log(c_n) + \log(b_{\max}+L+1) + \log(2)]L^{-1/2}$, for a constant C_Z which only depends on d_θ, $\mu_{d_\theta} = \mathbb{E}(\|ZZ' - I_{d_\theta}\|)$ where $Z \sim \mathcal{N}(0, I_{d_\theta})$ and $\overline{\sigma}_n = \sigma_{\max}[G(\theta^\dagger)] + L_G C_a c_n n^{-1/2}$, we have:

$$\sigma_{\max}[\overline{G}_L(\theta_b) - G_L(\theta_b)] \leq 4\overline{\sigma}_n\left(t_n + \sqrt{t_n}\right) + L_G\left(\max_{-L+1\leq\ell\leq 0}\|\theta_{b-\ell} - \hat{\theta}_n\|\right)[\mu_{d_\theta} + \tilde{t}_n], \quad \text{(D.34)}$$

with probability $1-3/c_n$.

Appendix E Proofs for the Preliminary Results for Section 4

Proof of Lemma D11: Take $b \geq 0$ and $\ell \geq 0$. For any $Z_{b-\ell} \in \mathbb{R}^{d_\theta}$ we have:

$$\|\frac{1}{\varepsilon}[\overline{g}_n(\theta_{b-\ell} + \varepsilon Z_{b-\ell}) - \overline{g}_n(\theta_{b-\ell}) - \varepsilon G(\theta_{b-\ell})Z_{b-\ell}]Z'_{b-\ell}\| \leq L_g C_\Theta c_n n^{-1/2}\varepsilon^{\psi-1}\|Z_{b-\ell}\|^{1+\psi} + \varepsilon L_G\|Z_{b-\ell}\|^3,$$

uniformly in $\theta_{b-\ell} \in \Theta$, with probability $1 - (1+C)/c_n$. The inequality relies on Lemma A4 and the Liptschitz-continuity assumption. Note that $\|Z_{b-\ell}\|^2 \sim \chi^2_{d_\theta}$ so that Lemma 1 in Laurent and Massart (2000) implies that, for any $t > 0$:

$$\mathbb{P}(\|Z_{b-\ell}\|^2 \geq d_\theta + 2d_\theta t + \sqrt{2d_\theta t}) \leq \exp(-t).$$

Pick $t_n = \log(b_{\max} + L + 1) + \log(c_n)$, we have:

$$\sup_{-L+1 \leq \ell \leq b_{\max}} \|Z_\ell\|^2 \leq d_\theta + 2d_\theta t_n + \sqrt{2d_\theta t_n},$$

with probability $1 - 1/c_n$. Since $d_\theta + 2d_\theta t_n + \sqrt{2d_\theta t_n} \geq 1$, this yields the bound:

$$\begin{aligned}\sup_{0 \leq b \leq b_{\max}, -L+1 \leq \ell \leq 0} &\|\frac{1}{\varepsilon}[\overline{g}_n(\theta_{b-\ell} + \varepsilon Z_{b-\ell}) - \overline{g}_n(\theta_{b-\ell}) - \varepsilon G(\theta_{b-\ell})Z_{b-\ell}]Z'_{b-\ell}\| \\ &\leq \left(L_g C_\Theta c_n n^{-1/2}\varepsilon^{\psi-1} + \varepsilon L_G\right)\left[d_\theta + 2d_\theta t_n + \sqrt{2d_\theta t_n}\right]^{3/2},\end{aligned}$$

with probability $1 - (2+C)/c_n$. Bound the average with the sup and we get the desired result (D.33). □

Proof of Lemma D12: The result will be derived using an ε-net argument (see Vershynin, 2018, Section 4.2) combined with an exponential tail inequality. Notice that:

$$\overline{G}_L(\theta_b) - G_L(\theta_b) = \frac{1}{L}\sum_{\ell=0}^{L-1} G(\theta_{b-\ell})[Z_{b-\ell}Z'_{b-\ell} - I_{d_\theta}],$$

where by Gaussianity of $Z_{b-\ell}$, the $Z_{b-\ell}Z'_{b-\ell}$ are iid Wishart distributed. Using the triangular inequality on the spectral norm and the Lipschitz-continuity of the Jacobian, we have:

$$\begin{aligned}\sigma_{\max}[\overline{G}_L(\theta_b) - G_L(\theta_b)] \leq \sigma_{\max}&\left[\frac{1}{L}\sum_{\ell=0}^{L-1} G(\hat{\theta}_n)[Z_{b-\ell}Z'_{b-\ell} - I_{d_\theta}]\right] \\ &+ \frac{1}{L}\sum_{\ell=0}^{L-1} L_G\|\theta_{b-\ell} - \hat{\theta}_n\| \times \|Z_{b-\ell}Z'_{b-\ell} - I_{d_\theta}\|.\end{aligned}$$

The following derives the bounds for these two terms. Let $S^{d_\theta-1}$ and S^{d_g-1} be respectively the unit sphere in $\mathbb{R}^{d_\theta}$ and $\mathbb{R}^{d_g}$, i.e. $x \in S^{d_\theta-1}$ implies $x \in \mathbb{R}^{d_\theta}$ and $\|x\| = 1$. For any given $b \in \{0, \ldots, b_{\max}\}$ and matrix A with dimensions $d_g \times d_\theta$, the largest singular value satisfies:

$$\sigma_{\max}[A] = \sup_{(x,y)\in S^{d_\theta-1}\times S^{d_g-1}} y'Ax.$$

Pick any two $(x, y) \in S^{d_\theta-1} \times S^{d_g-1}$, $b \geq 0, \ell \geq 0$ and a $t \in \mathbb{R}$ such that $ty'x < 1/2$:

$$\begin{aligned}
\mathbb{E}\left(\exp[ty'G(\hat{\theta}_n)[Z_{b-\ell}Z'_{b-\ell} - I_d]]\right) &= \mathbb{E}\left(\exp[\text{trace}(txy'G(\hat{\theta}_n)Z_{b-\ell}Z'_{b-\ell})\right)\exp(-ty'G(\hat{\theta}_n)x) \\
&= [\det(I - 2txy'G(\hat{\theta}_n))]^{-1/2}\exp(-ty'G(\hat{\theta}_n)G(\hat{\theta}_n)x) \\
&= [1 - 2ty'G(\hat{\theta}_n)x]^{-1/2}\exp(-ty'G(\hat{\theta}_n)x),
\end{aligned}$$

where the second equality follows from the formula for the moment generating function of a Wishart distribution (Muirhead, 1982, Section 8), the third equality follows from Sylvester's determinant identity. Note that $y'G(\hat{\theta}_n)x \leq \sigma_{\max}[G(\hat{\theta}_n)]$ for any unitary x, y. Let $\overline{\sigma}_n = \sigma_{\max}[G(\hat{\theta}_n)]$. Using an inequality from the proof of Lemma 1 in Laurent and Massart (2000), we have for $u = ty'G(\hat{\theta}_n)x \leq \overline{\sigma}_n t < 1/2$, $t \geq 0$:

$$\log\left[\mathbb{E}\left(\exp[ty'G(\hat{\theta}_n)[Z_{b-\ell}Z'_{b-\ell} - I_d]]\right)\right] \leq \frac{u^2}{1-2u} \leq \frac{t^2\overline{\sigma}_n^2}{1-2t\overline{\sigma}_n},$$

summing over $\ell = 0$ to $L-1$, yields:

$$\sum_{\ell=0}^{L-1}\log\left[\mathbb{E}\left(\exp[ty'G(\hat{\theta}_n)[Z_{b-\ell}Z'_{b-\ell} - I_d]]\right)\right] \leq \frac{t^2L\overline{\sigma}_n^2}{1-2t\overline{\sigma}_n} = \frac{t^2 2L\overline{\sigma}_n^2}{2(1-2t\overline{\sigma}_n)}.$$

From Birgé and Massart (1998), this implies that following inequality for any $t > 0$:

$$\mathbb{P}\left(\sum_{\ell=0}^{L-1} y'G(\hat{\theta}_n)x \geq 2\overline{\sigma}_n[t + \sqrt{Lt}]\right) \leq \exp(-t).$$

Now this will be combined with an ε-net argument. Using Vershynin (2018), Problem 4.4.3, for any ε-nets N_1, N_2 of the spheres $S^{d_\theta-1}$ and S^{d_g-1} we have when $\varepsilon < 1/2$:

$$\sigma_{\max}[A] \leq \sup_{(x,y)\in N_1\times N_2}\frac{1}{1-2\varepsilon}y'Ax.$$

The cardinality of N_1 and N_2 is at most $(2/\varepsilon+1)^{d_\theta}$ and $(2/\varepsilon+1)^{d_g}$, respectively (Vershynin, 2018, Corollary 4.2.13). Pick $\varepsilon = 1/4$, the cardinality of $N_1 \times N_2$ is at most 9^d with $d = d_\theta + d_g$. Then, for $0 \leq b \leq b_{\max}$, we have:

$$\mathbb{P}\left(\sup_{0\leq b\leq b_{\max}}\left[\sup_{(x,y)\in S^{d_\theta-1}\times S^{d_g-1}}\sum_{\ell=0}^{L-1} y'G(\hat{\theta}_n)x\right] \geq 4\overline{\sigma}_n[t_n + \sqrt{t_n}]\right) \leq (b_{\max}+1)9^d\exp(-Lt_n) = \frac{1}{c_n},$$

for $t_n = [\log(c_n) + d\log(9) + \log(b_{\max}+1)]/L$. This yields the first bound:

$$\sup_{0\le b\le b_{\max}} \sigma_{\max}\left[\frac{1}{L}\sum_{\ell=0}^{L-1} G(\hat\theta_n)[Z_{b-\ell}Z'_{b-\ell} - I_{d_\theta}]\right] \le 4\overline{\sigma}_n\left(t_n + \sqrt{t_n}\right),$$

with probability $1 - 1/c_n$. Note that the probability is conditional on the sample data. We also have $\overline{\sigma}_n \le \sigma_{\max}(G(\theta^\dagger)) + L_G\|\hat\theta_n - \theta^\dagger\| \le \overline{\sigma} + L_G C_a c_n n^{-1/2}$ with probability $1 - 1/c_n$. We can set $\overline{\sigma}_n = \overline{\sigma} + L_G C_a c_n n^{-1/2}$ in the bound and we get the same result with unconditional probability $1 - 2/c_n$.

To bound $\sigma_{\max}(Z_{b-\ell}Z'_{b-\ell} - I_{d_\theta}) = \|Z_{b-\ell}Z'_{b-\ell} - I_{d_\theta}\|$, follow the same steps as above replacing $G(\hat\theta_n)$ with I_{d_θ} and taking $y \in S^{d_\theta - 1}$, and we get for $t > 0$:

$$\mathbb{P}\left(\|Z_{b-\ell}Z'_{b-\ell} - I_{d_\theta}\| \ge 4(t + \sqrt{t})\right) \le \exp(-t),$$

from which we deduce that the distribution of $\|Z_{b-\ell}Z'_{b-\ell} - I_{d_\theta}\|$ is sub-exponential. Let $\mu_{d_\theta} = \mathbb{E}(\|Z_{b-\ell}Z'_{b-\ell} - I_d\|)$, apply Bernstein's inequality to find:[3]

$$\mathbb{P}\left(\left|\sum_{\ell=0}^{L-1}[\|Z_{b-\ell}Z'_{b-\ell} - I_d\| - \mu_{d_\theta}]\right| > t\right) \le 2\exp\left(-c\min\left(\frac{t^2}{LC_Z^2}, \frac{t}{C_Z}\right)\right),$$

where c is an absolute constant and C_Z is the Orlicz norm of $\|Z_{b-\ell}Z'_{b-\ell} - I_d\| - \mu_{d_\theta}$. Take:

$$\tilde t_n = C_Z\frac{\log(c_n) + \log(2) + \log(b_{\max} + L + 1)}{\sqrt{L}},$$

then with probability $1 - 1/c_n$, we have:

$$\sup_{-L+1\le \ell\le b_{\max}} \frac{1}{L}\sum_{\ell=0}^{L-1}\|Z_{b-\ell}Z'_{b-\ell} - I_d\| \le \mu_{d_\theta} + \tilde t_n,$$

uniformly in $b \in \{0,\dots,b_{\max}\}$. Now apply this to get:

$$\frac{1}{L}\sum_{\ell=0}^{L-1}[\|\theta_{b-\ell} - \hat\theta_n\| \times \|Z_{b-\ell}Z'_{b-\ell} - I_d\|] \le L_G\left(\max_{0\le\ell\le L-1}\|\theta_{b-\ell} - \hat\theta_n\|\right)\left(\mu_{d_\theta} + \tilde t_n\right),$$

uniformly in $b \in \{0,\dots,b_{\max}\}$, with probability $1 - 1/c_n$. Pick the same $t_n = \log(c_n) + \log(b_{\max} + L + 1) + d\log(9)$ for both bounds and we get:

$$\sigma_{\max}[\overline{G}_L(\theta_b) - G_L(\theta_b)] \le 4\overline{\sigma}_n\left(t_n + \sqrt{t_n}\right) + L_G\left(\max_{0\le\ell\le L-1}\|\theta_{b-\ell} - \hat\theta_n\|\right)\left(\mu_{d_\theta} + \tilde t_n\right),$$

with probability $1 - 3/c_n$. This is the desired result (D.34). □

[3]See e.g. Vershynin (2018), Theorem 2.8.1.

Appendix F Proofs for Section 4

Proof of Proposition 2: Take θ_{b+1} as described in (1'). As in the proof of Proposition 1, we have when $\|\theta_b - \hat{\theta}_n\| \leq R_{n,G}$:

$$\begin{aligned}
&\|\theta_{b+1} - \hat{\theta}_n - (1-\gamma)(\theta_b - \hat{\theta}_n) - \alpha(\theta_b - \theta_{b-1})\| \\
&\leq (B.9) + (B.10) \leq (B.17) = \gamma \frac{\sqrt{\kappa_W} L_G}{\underline{\sigma}_{n,\varepsilon}} \left(\|\theta_b - \hat{\theta}_n\| + M_{1,Z}\varepsilon \right) \|\theta_b - \hat{\theta}_n\| + \gamma \Delta_{n,\varepsilon},
\end{aligned}$$

with probability $1-(1+C)/c_n$, with the same $\Delta_{n,\varepsilon}$ used in Proposition 1. The left-hand-side can be re-written in companion form as:

$$\|\theta_{b+1} - \hat{\theta}_n - (1-\gamma)(\theta_b - \hat{\theta}_n) - \alpha(\theta_b - \theta_{b-1})\| = \|(\boldsymbol{\theta_{b+1}} - \boldsymbol{\hat{\theta}_n}) - A(\gamma,\alpha)(\boldsymbol{\theta_b} - \boldsymbol{\hat{\theta}_n})\|.$$

Now, apply the reverse triangular inequality and plug-in the above inequality to get:

$$\begin{aligned}
\|\boldsymbol{\theta_{b+1}} - \boldsymbol{\hat{\theta}_n}\| \leq \lambda_{\max}[A(\gamma,\alpha)] \|\boldsymbol{\theta_b} - \boldsymbol{\hat{\theta}_n}\| \\
+ \gamma \frac{\sqrt{\kappa_W} L_G}{\underline{\sigma}_{n,\varepsilon}} \left(\|\theta_b - \hat{\theta}_n\| + M_{1,Z}\varepsilon \right) \|\theta_b - \hat{\theta}_n\| + \gamma \Delta_{n,\varepsilon},
\end{aligned}$$

with probability $1-(1+C)/c_n$, with $\lambda_{\max}[A(\gamma,\alpha)] = 1-\gamma(\alpha)$ by definition. Now pick $\|\theta_b - \hat{\theta}_n\| \leq R_{n,\varepsilon}(\alpha)$, we have:

$$\|\boldsymbol{\theta_{b+1}} - \boldsymbol{\hat{\theta}_n}\| \leq (1-\overline{\gamma}) \|\boldsymbol{\theta_b} - \boldsymbol{\hat{\theta}_n}\| + \gamma \Delta_{n,\varepsilon},$$

with probability $1-(1+C)/c_n$ which concludes the proof. □

Proof of Proposition 3: Similar to Propositions 1 and 2 we can write with $\hat{G}_b = \hat{G}_L(\theta_b)$:

$$\|\theta_{b+1} - \hat{\theta}_n - (1-\gamma)(\theta_b - \hat{\theta}_n)\| \leq \gamma \sqrt{\kappa_W} \sigma_{\min}[\hat{G}_b]^{-1} \|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n) - G(\hat{\theta}_n)(\theta_b - \hat{\theta}_n)\| \tag{F.35}$$

$$+ \gamma \sigma_{\min}[\hat{G}_b]^{-2} \underline{\lambda}_W^{-1} \|\hat{G}_b' W_n \overline{g}_n(\hat{\theta}_n)\| \tag{F.36}$$

$$+ \gamma \sqrt{\kappa_W} \sigma_{\min}[\hat{G}_b]^{-1} \|\hat{G}_b - G(\hat{\theta}_n)\| \times \|\theta_b - \hat{\theta}_n\|. \tag{F.37}$$

First, a lower bound for the least singular value of $\hat{G}_b$ is required. Weyl's inequality implies $\sigma_{\min}(\hat{G}_b) \geq \sigma_{\min}[G(\theta^\dagger)] - \|G(\theta^\dagger) - \hat{G}_b\|$. Using the triangular inequality, Lipschitz-continuity of G, and Lemmas 3, D12 and D11:

$$\begin{aligned}
\|G(\theta^\dagger) - \hat{G}_b\| &\leq \|G_L(\theta_b) - G(\hat{\theta}_n)\| + \|G(\hat{\theta}_n) - G(\theta^\dagger)\| + \|\hat{G}_b - \tilde{G}_b\| + \|G_L(\theta_b) - \tilde{G}_b\| \\
&\leq L_G \big(\max_{-L+1 \leq \ell \leq 0} \|\theta_{b-\ell} - \hat{\theta}_n\| \big) + L_G C_a c_n n^{-1/2} + \|(D.33)\| + \|(D.34)\|,
\end{aligned}$$

where (D.33), (D.34) are given in Lemmas D12 and D11. They imply that

$$\|(D.33)\| + \|(D.34)\| \leq C_{(D.33)} \left(c_n n^{-1/2}\varepsilon^{\psi-1} + \varepsilon\right) \delta_n^{3/2} + C_{(D.34)}(1 + c_n n^{-1/2})\delta_n L^{-1/2} + L_G \mu_{d_\theta}(\max_{-L+1\leq \ell \leq 0} \|\theta_{b-\ell} - \hat{\theta}_n\|),$$

with probability $1 - (5 + C)/c_n$ when each $\|\theta_{b-\ell} - \hat{\theta}_n\| \leq R_G$, and $\delta_n = \log(c_n) + \log(b_{\max} + L + 1) \geq 1$. Pick $0 < R_{G,2} \leq R_G$ such that:

$$\sigma_{\min}[G(\theta^\dagger)] - L_G(1 + \mu_{d_\theta})R_{G,2} = \underline{\sigma}/2.$$

Note that $R_{G,2}$ only depends on G and d_θ. Putting everything together, for $\max_{-L+1\leq \ell \leq 0} \|\theta_{b-\ell} - \hat{\theta}_n\| \leq R_{G,2}$, there is a constant $C_{\sigma,2} > 0$ such that:

$$\sigma_{\min}(\hat{G}_b) \geq \underline{\sigma}/2 - C_{\sigma,2} \left(c_n n^{-1/2}\varepsilon^{\psi-1} + \varepsilon + L^{-1/2}\right) \delta_n^{3/2} := \hat{\underline{\sigma}}_{n,\varepsilon}, \tag{F.38}$$

with probability $1 - (5 + C)/c_n$. This allows to handle the singular value appearing several times in (F.36). Next, the bound for (F.35) can be derived using the same steps used in the proof of Proposition 1:

$$\|\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n) - G(\hat{\theta}_n)(\theta_b - \hat{\theta}_n)\| \leq L_G\|\theta_b - \hat{\theta}_n\|^2 + L_g C_\Theta c_n n^{-1/2}\|\theta_b - \hat{\theta}_n\|^\psi,$$

with probability $1 - (1 + C)/c_n$. Next, to bound (F.36) consider:

$$\|\hat{G}_b' W_n \overline{g}_n(\hat{\theta}_n)\| \leq \overline{\lambda}_W \|\hat{G}_b - G(\theta^\dagger)\| \times \|\overline{g}_n(\hat{\theta}_n)\| \tag{F.39}$$

$$+ \|G(\theta^\dagger)' W_n \overline{g}_n(\hat{\theta}_n)\| \tag{F.40}$$

From the proof of Lemma 3, we have:

$$\|(F.39)\| \leq \overline{\lambda}_W \|\hat{G}_b - G(\theta^\dagger)\| C_c c_n n^{-1/2}, \quad \|(F.40)\| \leq C_{(B.4)}(c_n n^{-1/2})^{1+\psi},$$

with probability $1-(1+C)/c_n$. Using the derivations for the singular value above, we further get:

$$\|(F.39)\| \leq C_{(F.39)} c_n n^{-1/2} \left[\left(c_n n^{-1/2}\varepsilon^{\psi-1} + \varepsilon + L^{-1/2}\right) \delta_n^{3/2} + \max_{-L+1\leq \ell \leq 0} \|\theta_{b-\ell} - \hat{\theta}_n\|\right],$$

with probability $1 - (5 + C)/c_n$. Finally, for (F.37), we can use:

$$\|\hat{G}_b - G(\hat{\theta}_n)\| \leq L_G \mu_{d_\theta}(\max_{-L+1\leq \ell \leq 0} \|\theta_{b-\ell} - \hat{\theta}_n\|) + C_{(F.37)} \left(c_n n^{-1/2}\varepsilon^{\psi-1} + \varepsilon + L^{-1/2}\right) \delta_n^{3/2},$$

with probability $1 - (5 + C)/c_n$.

Now take $\mathcal{E}_b = (\max_{-L+1\leq \ell \leq 0} \|\theta_{b-\ell} - \hat{\theta}_n\|) \leq R_{G,2} - C_a c_n n^{-1/2}$, so that $\|\theta_{b-\ell} - \theta^\dagger\| \leq R_G$ for all $\ell \in \{-L+1, \ldots, 0\}$. Combine the bounds to find:

$$\|\theta_{b+1} - \hat{\theta}_n\| \leq \left(1 - \gamma + \gamma \hat{\underline{\sigma}}_{n,\varepsilon}^{-1} \sqrt{\kappa_W} L_G \left[\|\theta_b - \hat{\theta}_n\| + \mu_{d_\theta} \mathcal{E}_b\right]\right) \|\theta_b - \hat{\theta}_n\| + \frac{\gamma}{\hat{\underline{\sigma}}_{n,\varepsilon}^2} \hat{\Delta}_{n,\varepsilon}(\|\theta_b - \hat{\theta}_n\|, \mathcal{E}_b),$$

with probability $1 - (5 + C)/c_n$. The last term is given by:

$$\hat{\Delta}_{n,\varepsilon}(\|\theta_b - \hat{\theta}_n\|, \mathcal{E}_b) = C_3 \left(\hat{\Gamma}_{n,\varepsilon} + \delta_n^{3/2}[c_n n^{-1/2} \varepsilon^{\psi-1} + \varepsilon + L^{-1/2}]\mathcal{E}_b + c_n n^{-1/2} \|\theta_b - \hat{\theta}_n\|^\psi\right),$$

using: $\hat{\Gamma}_{n,\varepsilon} = c_n n^{-1/2} \left[(c_n n^{-1/2})^\psi + \delta_n^{3/2}(c_n n^{-1/2} \varepsilon^{\psi-1} + \varepsilon + L^{-1/2})\right]$. Noting that $\|\theta_b - \hat{\theta}_n\| \leq \mathcal{E}_b$, this implies that

$$\mathcal{E}_b \leq \frac{\overline{\gamma} - \gamma}{\gamma} \frac{\hat{\underline{\sigma}}_{n,\varepsilon}}{L_G \sqrt{\kappa_W}(1 + \mu_{d_\theta})} \Rightarrow \|\theta_{b+1} - \hat{\theta}_n\| \leq (1 - \overline{\gamma})\|\theta_b - \hat{\theta}_n\| + \frac{\gamma}{\hat{\underline{\sigma}}_{n,\varepsilon}^2} \hat{\Delta}_{n,\varepsilon}(\|\theta_b - \hat{\theta}_n\|, \mathcal{E}_b),$$

with probability $1 - (5 + C)/c_n$, which is the desired contraction result. □

Appendix G Additional Results for Section 5

G.1 Aiyagari Model

Additional implementation details Consumers face an exogenous income flow and choose consumption c_t to maximize the expected intertemporal utility under a borrowing constraint:

$$\max \mathbb{E}_t \left(\sum_{j=0}^{\infty} \beta^j u(c_{t+j}) \right) \text{ s.t. } a_{t+j+1} + c_{t+j} \leq y_{t+j} + (1+r)a_{t+j}, a_{t+j} \geq \underline{a},$$

where $a_{t+j} \geq 0$ are future saving at period $t+j$ for $j = 0, 1, \ldots$, $r \geq 0$ is the interest rate, and $\underline{a} \leq 0$ is the borrowing constraint. The utility function is $u(c) = (c^\gamma - 1)/(1-\gamma)$, where $\gamma > 0$ measures risk-aversion. Finally, income y_{t+j} follows an AR(1) process in logarithms:

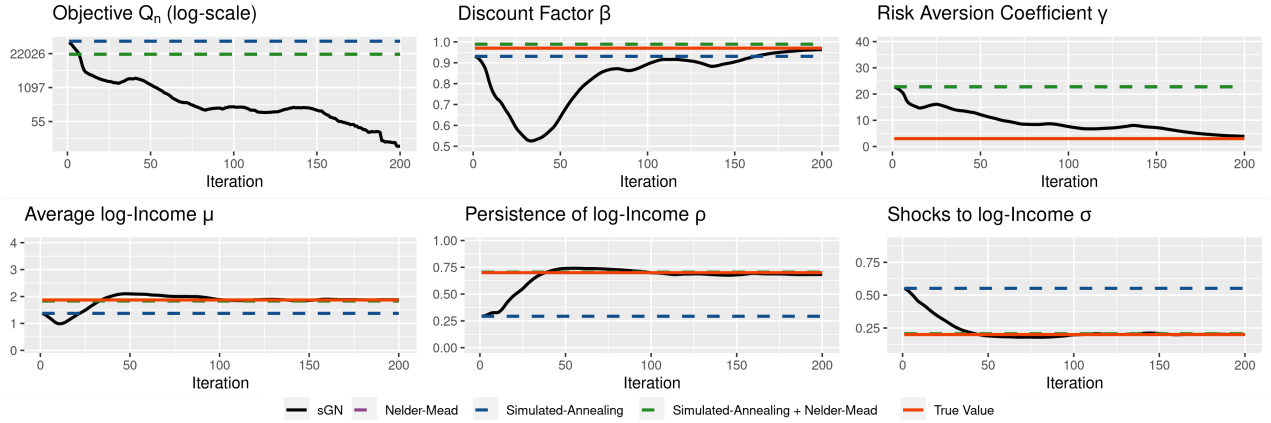
$$\log(y_{t+1}) = \mu + \rho[\log(y_t) - \mu] + \sigma e_t, \quad e_t \overset{iid}{\sim} \mathcal{N}(0,1).$$

To solve the model, the income process is discretized using the quadrature method of Tauchen and Hussey (1991) on a 15 point grid. Then the policy function is computed using value function iterations over a discretized state-space of 300 points.[4] A panel of consumers is generated (a_{it}, y_{it}) with $i = 1, \ldots n$, $n = 10000$, $t = 1, 2$.[5] The estimation matches the $\tau = 0.2, 0.3, \ldots, 0.7, 0.8$ level quantiles of the asset distribution (y_{i2}) between sample and simulated data, as well as the OLS AR(1) estimates from regressing $\log(y_{i2})$ on $\log(y_{i1})$. Interest rates are fixed to a 5% annual rate. The true value is $\theta^\dagger = (\beta^\dagger, \gamma^\dagger, \mu^\dagger, \rho^\dagger, \sigma^\dagger) = (0.97, 3, \log(6.5), 0.7, 0.2)$.

[4] There are many other solution methods for this type of model. Value function iterations can be applied to wide range of models but also make estimation very difficult. As such, it makes for a good benchmark to evaluate Algorithm 1.

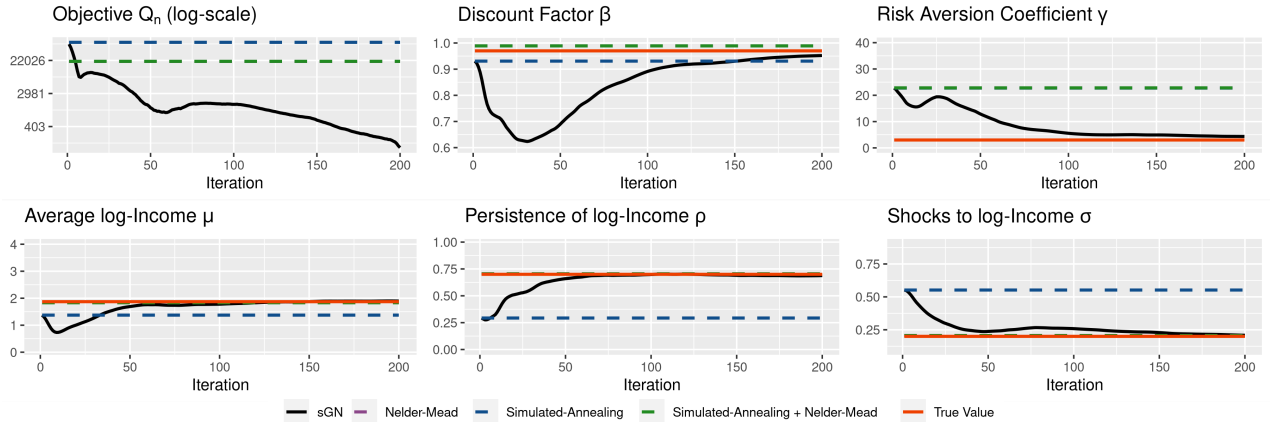
[5] A burn-in of 248 time periods is discarded for each individual to approximate the ergodic distribution.

Figure G6: Aiyagari Model: local, global optimizers and sGN ($\varepsilon = 0.2$)



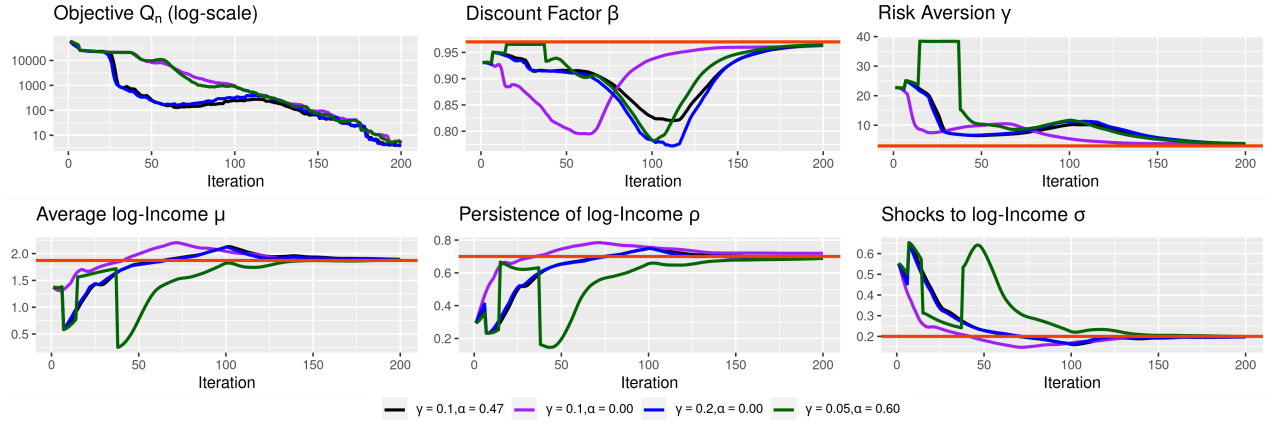
Legend: $n = 10000$, $T = 2$. $\gamma = 0.1$, $\alpha = 0.47$. sGN (black): Algorithm 1. Simulated-Annealing (dashed blue): 5000 iterations from θ_0. Simulated-Annealing + Nelder-Mead (dashed green): run Nelder-Mead after 5000 Simulated-Annealing iterations.

Figure G7: Aiyagari Model: local, global optimizers and sGN ($\varepsilon = 0.5$)



Legend: $n = 10000$, $T = 2$. $\gamma = 0.1$, $\alpha = 0.47$. sGN (black): Algorithm 1. Simulated-Annealing (dashed blue): 5000 iterations from θ_0. Simulated-Annealing + Nelder-Mead (dashed green): run Nelder-Mead after 5000 Simulated-Annealing iterations.

Figure G8: Aiyagari Model: sGN with different choices of tuning parameters ($\varepsilon = 0.1$)



Legend: $\varepsilon = 0.1$ for all 4 sets of tuning parameters. Orange line = True Value.

Results with other tuning parameters $\varepsilon, \gamma, \alpha$

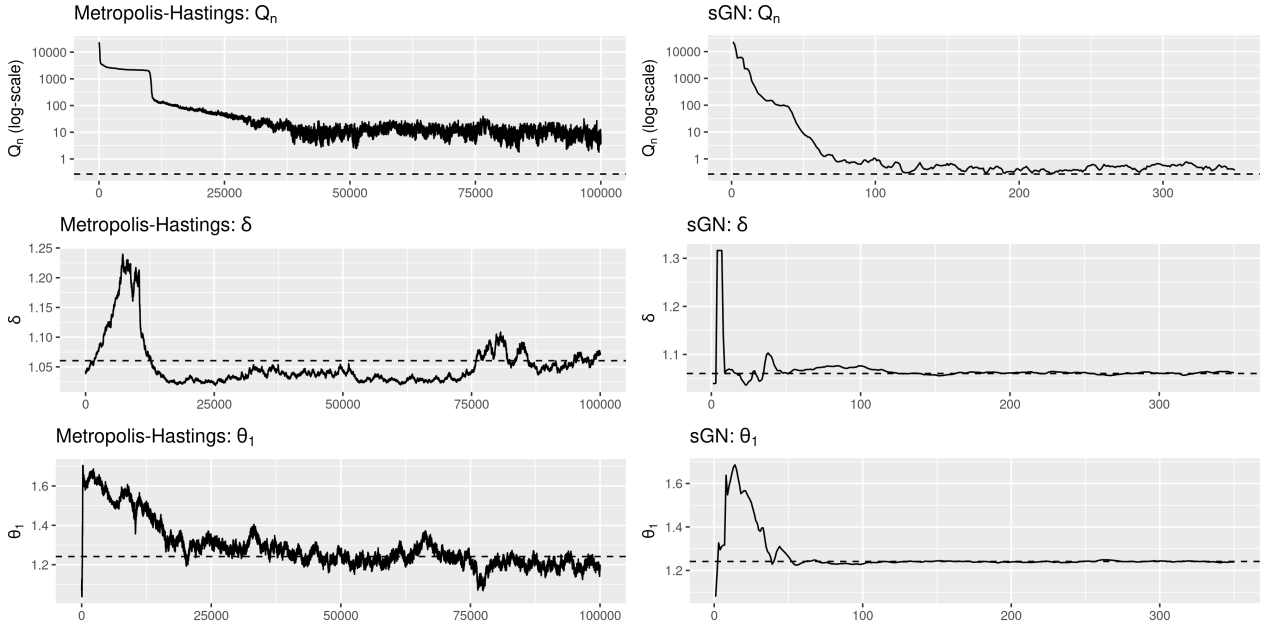
G.2 Interdependent Durations

Table G5: Interdependent Duration Estimates: Honoré and de Paula (2018) and sGN

	Coefficients for Wives				Coefficients for Husbands			
	Honoré & de Paula		sGN		Honoré & de Paula		sGN	
δ	1.052	1.064	1.060	1.064	1.052	1.064	1.060	1.064
	(0.039)	(0.042)	(0.039)	(0.037)	(0.039)	(0.042)	(0.039)	(0.037)
θ_1	1.244	1.244	1.241	1.233	1.169	1.218	1.181	1.192
	(0.054)	(0.054)	(0.055)	(0.050)	(0.043)	(0.058)	(0.043)	(0.040)
$\geq$ 62 yrs-old	10.640	13.446	10.203	12.254	31.532	39.824	33.330	35.371
	(5.916)	(5.694)	(7.818)	(5.692)	(11.356)	(11.372)	(8.131)	(7.672)
$\geq$ 65 yrs-old	10.036	12.326	10.480	11.974	25.696	29.254	25.203	26.240
	(11.555)	(7.495)	(10.067)	(10.897)	(9.497)	(11.229)	(13.215)	(14.289)
Constant	−5.786	−5.790	−5.777	−5.695	−5.587	−5.449	−5.644	−5.240
	(0.225)	(0.276)	(0.226)	(0.250)	(0.231)	(0.266)	(0.189)	(0.207)
Age Diff.	−0.074	−0.075	−0.076	−0.073	0.021	0.025	0.021	0.025
	(0.016)	(0.016)	(0.016)	(0.016)	(0.008)	(0.007)	(0.008)	(0.008)
Non-Hisp. Black		−0.149		−0.150		−0.203		−0.177
		(0.153)		(0.156)		(0.155)		(0.162)
Other race		−0.649		−0.653		−0.151		−0.162
		(0.337)		(0.308)		(0.287)		(0.289)
Hispanic		−0.490		−0.493		−0.626		−0.615
		(0.192)		(0.189)		(0.180)		(0.178)
High school or GED		0.052		0.015		−0.109		−0.098
		(0.158)		(0.158)		(0.118)		(0.128)
Some college		−0.131		−0.145		−0.357		−0.331
		(0.169)		(0.170)		(0.133)		(0.144)
College or above		−0.052		−0.085		−0.522		−0.493
		(0.189)		(0.192)		(0.128)		(0.135)
NE		0.002		−0.039		0.060		0.039
		(0.146)		(0.149)		(0.122)		(0.133)
SO		0.065		0.032		−0.219		−0.243
		(0.115)		(0.115)		(0.106)		(0.111)
WE		0.217		0.181		0.066		0.051
		(0.145)		(0.150)		(0.121)		(0.131)
τ	0.526	0.429	0.424	0.429	-	-	-	-
	(0.399)	(0.371)	(0.296)	(0.300)	-	-	-	-
Starting Obj. Value	93.70	89.77	2.10^4	5.10^4	-	-	-	-
Final Obj. Value	0.470	0.758	0.271	0.342	-	-	-	-
Number of Coef.	12	30	12	30	-	-	-	-
Number of Obs.	1227	1227	1227	1227	-	-	-	-
Computation Time	3h25m	5h34m	11min	11min	-	-	-	-

Legend: sGN: $\varepsilon = 10^{-2}$, $\gamma = 0.1$, $\alpha = 0.47$, $B = 350$ iterations in total. Husbands: - same as wives. Coefficients for wives and husbands are estimated jointly.

Figure G9: Interdependent Duration Estimates: MCMC and sGN



Legend: sGN: $\varepsilon = 10^{-2}$, $\gamma = 0.1$, $\alpha = 0.47$, $B = 350$ iterations in total. MCMC: 100000 iterations, same starting value, random-walk tuned to target $\approx 38\%$ acceptance rate around the solution $\hat{\theta}_n$.

Appendix H Comparison with a multi-start Algorithm

The following discusses the differences between the global convergence results in Lemma 2 and global convergence using multiple starting values for the local optimizer in Lemma 1. Recall from Lemma 1 that for any $\|\theta_0 - \theta^\dagger\| \leq R$, we have $\|\theta_b - \theta^\dagger\| \leq (1-\overline{\gamma})^b R$. Using covering arguments, running the local algorithm with $\tilde{k} \geq 1$ different starting values $(\theta^\ell)_{0\leq\ell\leq\tilde{k}-1}$ results in convergence if $\|\theta^\ell - \theta^\dagger\| \leq R$ for some $\ell \in \{0,\ldots,\tilde{k}-1\}$. Using covering arguments, this implies that $\tilde{k} \geq R^{-d_\theta}\text{vol}(\Theta)/\text{vol}(B)$. Because $R \geq \underline{r}_g$, using the lower bounds for $\tilde{k}$ and the k required in Lemma 2 we have $\tilde{k} \leq k$ for the same local step.

As for Algorithm 1, the $\tilde{k}$ needed in practice depends on the choice of covering sequence, the parameter space, and the moments. It is not known to the researcher so that it is often recommended to pick a large $\tilde{k}$, though some papers use only a handful of starting values in practice. This contrasts with Algorithm 1 where the user only needs to input a stopping criteria, it does not require to set k explicitly as the Algorithm transitions from global to local convergence without any input from the user (cf. Lemma 2, Theorem 2). This implies that the $\tilde{k}$ used in practice could be greater than k, if the user chooses a very large $\tilde{k}$ to guarantee convergence. If the multi-start algorithm is run for $\tilde{b}$ iterations for each starting value. The cost of running the algorithm is at least $\tilde{k} \times \tilde{b}$. In comparison, the combined local and global steps are run for $b = k + j$ iterations, but each iteration involves both the local and global steps instead of just the local step. The combined local and global steps are less costly when the user would set $\tilde{k} > k$ to be very large to ensure convergence. Also note, that other choices of local minimizers could be associated with a different R, potentially smaller than $\underline{r}_g$, and a potentially slower rate of convergence. This could be the case when using a multi-start approach with the Nelder-Mead algorithm for the local search, for instance.