Convexity Not Required: Estimation of Smooth Moment Condition Models

Jean-Jacques Forneron* Liang Zhong†

April 19, 2023

Abstract

Generalized and Simulated Method of Moments are often used to estimate structural Economic models. Yet, it is commonly reported that optimization is challenging because the corresponding objective function is non-convex. For smooth problems, this paper shows that convexity is not required: under a global rank condition involving the Jacobian of the sample moments, certain algorithms are globally convergent. These include a gradient-descent and a Gauss-Newton algorithm with appropriate choice of tuning parameters. The results are robust to 1) non-convexity, 2) one-to-one non-linear reparameterizations, and 3) moderate misspecification. In contrast, Newton-Raphson and quasi-Newton methods can fail to converge for the same estimation because of non-convexity. A simple example illustrates a non-convex GMM estimation problem that satisfies the aforementioned rank condition. Empirical applications to random coefficient demand estimation and impulse response matching further illustrate the results.

JEL Classification: C11, C12, C13, C32, C36.
Keywords: Non-linear estimation, over-identification, misspecification.

*Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA. Email: jjmf@bu.edu, Website: http://jjforneron.com.

†Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA. Email: samzl@bu.edu.

The authors would like to thank Jessie Li for suggesting to look at misspecified models and Hiro Kaido, David Lakagos, Bernard Salanié for useful comments.

1 Introduction

The Generalized and Simulated Method of Moments (GMM, SMM) are commonly used to estimate structural Economic models. To find these estimates, modern computer software provides researchers with a large set of free and non-free numerical optimizers, which, after inputting some tuning parameters, return a guess for the parameters of interest. Graduate level Econometric textbooks describe the sampling properties of the estimator but often only dedicate a short section, or paragraph, to the methods used to find the estimator.[1] A number of authors have pointed out the lack of robustness of some of these methods in empirical settings, see for instance McCullough and Vinod (2003), Stokes (2004) or Knittel and Metaxoglou (2014). A simple way to make the estimation more robust is to use a combination of methods in the hope that where one fails, another might succeed.

First, to better understand the issue at hand, a survey of papers published in the American Economic Review between 2016 and 2018 that compute a GMM or related estimator gives some insights into empirical practice.[2] Quantitatively, it highlights the most widely used optimizers and the dimension of a typical estimation problem. An overview of the most popular algorithms and a review of their known local/global convergence properties are given. Qualitatively, authors often comment on the challenge of estimation due to the non-convexity of the sample GMM objective function. Methods like gradient-descent or quasi-Newton are essentially absent from our survey, not too surprisingly as these are convex optimizers.

Second, the main contribution of the paper is to show that *convexity is not required* for some methods to perform well in GMM estimation: some algorithms are globally convergent if the Jacobian of the sample moments satisfies a global rank condition. Since this is perhaps surprising, the following gives some intuition behind the result. Suppose the sample moments $\overline{g}_n(\theta) = \partial_\theta \ell_n(\theta)$ correspond to the gradient of a sample log-likelihood function, say that of a Probit model. Then, their Jacobian $G_n(\theta) = \partial^2_{\theta,\theta'} \ell_n(\theta)$ is the Hessian of the log-likelihood. If the Jacobian has full rank everywhere, then the Hessian is strictly negative definite everywhere; the log-likelihood, to be maximized, is strictly concave. The same need not be true for the GMM objective $Q_n(\theta) = \frac{1}{2}\|\overline{g}_n(\theta)\|^2$, to be minimized, here with identity weighting. Its Hessian $\partial^2_{\theta,\theta'} Q_n(\theta) = G_n(\theta)' G_n(\theta) + (\overline{g}_n(\theta)' \otimes I_d)\partial_\theta \text{vec}(G_n'(\theta))$ can be singular,

[1] Davidson et al. (2004, Ch 6.4), Wooldridge (2010, Ch10), and Hayashi (2011, Ch 7.5) give an overview and references, Cameron and Trivedi (2005, Ch10) gives a comprehensive list of methods. Hansen (2022a) does not cover the topic, Hansen (2022b, Ch12) covers several methods. The handbook chapter by Quandt (1983) gives an in-depth overview of several methods.

[2] The related estimators considered here include the Minimum Distance (MD), Simulated Method of Moments (SMM), and Indirect Inference (II) estimators.

or non-definite, depending on the last term.[3] When this is the case: ℓ_n is concave but Q_n is non-convex, even though they estimate the exact same quantity $\hat{\theta}_n$.

The method of moments can be cast as systems of non-linear equations, $\overline{g}_n(\theta) = 0$, whereas GMM is typically framed as an M-estimation problem, $\min_{\theta\in\Theta} Q_n(\theta)$. The Probit example shows information can be lost when minimizing Q_n with off-the-shelf optimizers. This paper shows that some algorithms are more robust to the non-convexity of Q_n by implicitly solving $\overline{g}_n(\theta) = 0$ rather than minimizing Q_n. Under a rank condition, gradient-descent and Gauss-Newton (GN) are globally convergent, with appropriate tuning. Newton-Raphson and quasi-Newton can be unstable as they require inverting the Hessian of Q_n, which can be singular. The result extends to over-identified and moderately misspecified models by adapting the rank condition appropriately. As one may suspect, the rank condition precludes local optima. Unlike convexity, the rank condition is invariant to smooth one-to-one non-linear reparameterization. When there is a single parameter and moment, the condition is very intuitive: if the moment is strictly increasing (or decreasing) in the parameter, GN is globally convergent. To illustrate: if savings are strictly increasing with risk aversion, then using the savings rate as the moment yields the condition for the coefficient of risk aversion.

A simple MA(1) estimation from Gourieroux and Monfort (1996) illustrates this analytically and numerically. The problem is non-convex, yet the rank condition holds. As predicted, the recommended Gauss-Newton algorithm converges. Newton-Raphson provably diverges, and off-the-shelf optimizers can be unstable. Then, two empirical applications further confirm the predictions. The first application revisits the numerical results of Knittel and Metaxoglou (2014) for estimating random coefficient demand models. The same GN algorithm systematically converges from a wide range of starting values. A basic implementation of GN consists of a loop with *three lines of code*. In contrast, R's more sophisticated built-in optimizers can be inaccurate and often crash without additional error-handling. The second application estimates a small New Keynesian model with endogenous total factor productivity by impulse response matching. Matlab's built-in optimizers have better error-handling so that crashes are less problematic. Nonetheless, these optimizers' performance can be mixed and sensitive to reparameterizations whereas GN performs well for nearly all starting values.

The main takeaway is that non-convexity need not be a deterrent to structural estimation: simple algorithms converge quickly and globally under alternative conditions. Should the rank condition fail, Forneron (2023) builds an algorithm that is globally convergent under

[3] $\otimes$ is the kronecker product, vec vectorizes the matrix into a column vector, $\partial_\theta \text{vec}(G_n'(\theta))$ is the Jacobian of the vectorized Jacobian.

standard econometric assumptions and allows for non-smooth sample moments.

Structure of the paper. First, Section 2 gives a survey of empirical practice and the properties of some commonly used algorithms. Then, Section 3 provides the main results, illustrated in Section 4 with two empirical applications. Appendix A gives the proofs to the main results. Appendix B gives R code to replicate the MA(1) example. Appendix C provides additional simuation and empirical results. Appendix D gives additional details about the methods found in the survey.

2 Commonly used methods and their properties

Before introducing the results, the following provides an overview of empirical practice and the properties of the algorithms that are commonly used.

2.1 A survey of empirical practice

Survey methodology: the survey covers empirical papers published in the American Economic Review (AER) between 2016 and 2018. The focus on this specific outlet is driven by the mandatory data and code policy enacted in 2005. Indeed, since a number of papers provide little or no detail in the paper on the methodology used to compute estimates numerically, it is important to read the replication codes to determine what was implemented. The search function in JSTOR was used to find the papers matching the survey criteria. The database did not include more recent publications at the time of the survey.[4] Table 1 was constructed by reading through the main text, supplemental material, and all available replication codes of the selected papers.

Survey results: Table 1 provides an overview of the quantitative results of the survey. Additional details on the algorithms in the table are given below. There are 23 papers in total, a little over 6 papers per year. Excluding the estimation with 147 parameters, the average estimation has around 10 coefficients, and the median is 6. 3 papers used more than one starting value, and the remaining 20 papers either used the solver default or typed in

[4]The search function in JSTOR allows to search for keywords within the title, abstract, main text, and supplemental material of a paper. Further screening is required to ensure that each paper in the search results actually implements at least one of the estimations considered. The search criteria include keywords: "Method of Moments," "Indirect Inference," "Method of Simulated Moments," "Minimum Distance," and "MM."

Table 1: American Economic Review 2016-2018: GMM and related empirical estimations

Method	# Papers	# Parameters (p)	Data included?
Nelder-Mead	5	2,3,6 (×3),11	0
Simulated Annealing + Nelder-Mead	3	4,8,13	1
Nelder-Mead with multiple starting values	2	?,12	0
Pattern Search	2	6,147	1†
Genetic Algorithm	2	9,14	1
Simulated Annealing	1	4	0
MCMC	1	15	0
Grid Search	1	5	0
No code, no description	6	-	-
Stata/Mata default	5	3,6 (×2),38	3⋆

Legend: # Parameters correspond to the size of the largest specification. Data included? reports if the dataset is included with the replication files. Estimations surveyed include: Generalized Method of Moments (GMM), Minimum Distance (MD), Simulated Method of Moments (SMM), and Indirect Inference. ?: information not available due to the lack of replication codes. ⋆: one of the 3 papers we report to include data requires to download the PSID dataset separately. †: this paper also relies separately on Nelder-Mead, so it is reported under both methods.

a specific value in the replication code. There is generally no information provided on the origin of these specific starting values. Of the papers using multiple starting values, one did not provide replication codes, and the other two used 12 and 50 starting points. Some of the estimations are very time-consuming. For instance, Lise and Robin (2017) use MCMC for estimation (but not inference) and report that each evaluation of the moments takes 45s. In total, their estimation takes more than a week to run in a 96 core cluster environment.

Overall, 10 papers rely on the Nelder-Mead algorithm, alone or in combination with another method, making it the most popular optimizer in this survey. Pattern search, used in 2 papers, belongs to the same family of algorithms as Nelder-Mead. The following gives a brief overview of some methods in Table 1 and their convergence properties. There are few formal results for Genetic algorithms, so they are not discussed.

2.2 Overview of the Algorithms and their properties

The following describes three of the algorithms in Table 1: Nelder-Mead, Grid Search, Multi-Start, and Simulated Annealing. The goal is to give a brief overview of their known convergence properties; further description for each method is given in Appendix D.

Notation: Q_n is a continuous objective function to be minimized over Θ, a convex and compact subset of $\mathbb{R}^p$, $p \geq 1$, $\hat{\theta}_n$ denotes the solution to this minimization problem.

Nelder-Mead. Also called the simplex algorithm, the Nelder and Mead (1965, NM) algorithm comes out as a standard choice for empirical work in our survey. Notably, it was used in Berry et al. (1995, Sec6.5) to estimate the BLP model for the automobile industry. Its main feature is that it can be used even if Q_n is not continuous. It is often referred to as a *local derivative-free* optimizer. It belongs to the direct search family, which includes pattern search seen in Table 1 above.

Despite being widely used, formal convergence results for the simplex algorithm are few. Notably, Lagarias et al. (1998) proved convergence for strictly convex continuous functions for $p = 1$, and a smaller class of functions for $p = 2$ parameters. McKinnon (1998) gave counter-examples for $p = 2$ of smooth, strictly convex functions for which the algorithm converges to a point that is neither a local nor a global optimum, i.e. does not satisfy a first-order condition.[5] Using the algorithm once may not produce consistent estimates in well-behaved problems so it is sometimes combined with a multiple starting value strategy, described below. The TIKTAK Algorithm of Arnoud et al. (2019) builds on NM with multiple starting values. Despite these potential limitations, NM remains popular in empirical work.

Grid-Search. As the name suggests, a grid-search returns the minimizer of Q_n over a finite grid of points. In Economics, it is sometimes used to estimate models where the number of parameters p is not too large. One notable example is Donaldson (2018), who estimates $p = 3$ non-linear coefficients in a gravity model.

Contrary to NM above, grid-search has global convergence guarantees. However, convergence is very slow. Suppose we want the minimizer $\tilde{\theta}_k$ over a grid of k points to satisfy: $Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n) \leq \varepsilon$. Then the search requires at least $k \geq C\varepsilon^{-p}$ grid points where C depends on Q_n and the bounds used for the grid. Suppose $C = 1$, $p = 3$, $\varepsilon = 10^{-2}$, at least $k \geq 10^6$ grid points are needed, which is quite large. If each moment evaluation requires 45s, as in Lise and Robin (2017), this translates into 1.5 years of computation time.

Simulated Annealing. Unlike the methods above, Simulated Annealing (SA) is not a deterministic but a Monte Carlo based optimization method. Along with NM, SA stands out as the standard choice in empirical work. Like the grid-search, SA is guaranteed to converge, with high probability, as the number of iterations increases for an appropriate choice of tuning parameters. The main issue is that tuning parameters for which convergence results have been established result in very slow convergence: $\|\theta_k - \hat{\theta}_n\| \leq O_p(1/\sqrt{\log[k]})$, after

[5] Powell (1973) gives additional counter-examples for the class of direct search algorithms which includes NM and Pattern Search.

k iterations. As a result, SA could - in theory - converge more slowly than a grid-search. Chernozhukov and Hong (2003) consider the frequentist properties of a GMM-based quasi-Bayesian posterior distribution. Draws can be sampled using the random-walk Metropolis-Hastings algorithm which is closely related to SA.

Multiple Starting Values. To accommodate some of the limitations of optimizers, especially the lack of global convergence guarantees, it is common to run a given algorithm with multiple starting values. Setting the starting values is similar to choosing a grid for a grid-search. Andrews (1997) provides a stopping rule which can be used to determine if sufficiently many starting values were used or not. The required number of starting values depends on the objective function Q_n, the choice of the optimizer, and the properties of the sequence used to generate starting values.

3 GMM Estimation without Convexity

The review above highlights some of the challenges involved in minimizing a generic objective function Q_n. The following considers GMM objective functions more specifically and shows that faster convergence is possible, even without global convexity.

Let $\overline{g}_n(\theta) = 1/n \sum_{i=1}^n g(\theta; x_i)$ be the sample moments and $G_n(\theta) = \partial_\theta \overline{g}_n(\theta)$ their Jacobian. W_n is a weighting matrix which, for simplicity, does not depend on θ – this excludes continuously-updated estimations. The sample GMM objective function is:

$$Q_n(\theta) = \frac{1}{2} \overline{g}_n(\theta)' W_n \overline{g}_n(\theta).$$

The following gives high level assumptions used to describe the optimization properties.

Assumption 1. *With probability approaching 1: i.* Q_n *has a unique minimum* $\hat{\theta}_n \in$ *interior*(Θ)*, ii.* $\overline{g}_n$ *is twice continuously differentiable, iii.* G_n *is Lipschitz continuous with constant* $L \geq 0$*, and for some* $R_G > 0$ *such that,* $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$ *for all* $\|\theta - \hat{\theta}_n\| \leq R_G$*, iv. The parameters space* Θ *is convex and compact, v.* W_n *is such that* $0 < \underline{\lambda}_W \leq \lambda_{\min}(W_n) \leq \lambda_{\max}(W_n) \leq \overline{\lambda}_W < \infty$.

Primitive conditions for Assumption 1 are given in Appendix A.1. The uniqueness of the arg-minimizer $\hat{\theta}_n$ ensures that the optimization problem has a unique, well-defined solution. Without loss of generality, R_G is such that the closed ball around $\hat{\theta}_n$ of radius R_G is a subset of Θ, with probability approaching 1. Lemma A1 shows that this holds under Assumption

A1. This section will study derivative-based optimizers of the form:

$$\theta_{k+1} = \theta_k - \gamma_k P_k G_n(\theta_k)' W_n \overline{g}_n(\theta_k), \tag{1}$$

for some staring value $\theta_0 \in \Theta$ and a symmetric conditioning matrix P_k. The tuning parameter $\gamma_k \in (0, 1]$ is called the learning rate. It will be assumed to be constant in the following, i.e. $\gamma_k = \gamma$, for simplicity. In practice, adaptive choices of γ_k are common, using a line search for instance. Note that these should satisfy certain conditions to preserve convergence properties, which involve additional tuning parameters (Nocedal and Wright, 2006, Ch3.1). Choices of P_k discussed below correspond to the following algorithms:

1. Gradient-Descent (GD): $P_k = I_d$, for all $k \geq 0$,
2. Newton-Raphson (NR): $P_k = [\partial^2_{\theta,\theta'} Q_n(\theta_k)]^{-1}$, for all $k \geq 0$,
3. quasi-Newton (QN): P_k approximates the Hessian inverse above,
4. Gauss-Newton (GN): $P_k = [G_n(\theta_k)' W_n G_n(\theta_k)]^{-1}$.

The most popular QN software implementation is called BFGS. GD iterations are always well defined, whereas NR requires the Hessian $\partial^2_{\theta,\theta'} Q_n$, and GN the Jacobian G_n to be non-singular.

Assumption 2. *With probability approaching 1: P_k is such that $0 < \underline{\lambda}_P \leq \lambda_{\min}(P_k) \leq \lambda_{\max}(P_k) \leq \overline{\lambda}_P < \infty$.*

Assumption 2 requires P_k to be finite and strictly positive definite. As a result, when Q_n is non-convex, Assumption 2 may not hold for NR and QN without modifications that ensure P_k is non-singular and definite. Some commercial solvers implement modifications described in Nocedal and Wright (2006, Ch3.4). There are, however, a number of additional tuning parameters involved, so numerical stability is not necessarily guaranteed. Conlon and Gortmaker (2020, p1121) recommend using Knitro's Interior/Direct algorithm for BLP estimation – it is designed for problems where the Hessian can be near-singular or non-definite.[6] The following example illustrates that a simple GN algorithm can still deliver reliable results when this occurs.

A pen and pencil example. To build intuition, consider a simple MA(1) process:

$$y_t = e_t - \theta^\dagger e_{t-1}, \quad e_t \overset{iid}{\sim} \mathcal{N}(0, 1), \quad \theta^\dagger \in [-1, 1] = \Theta,$$

[6] See Knitro User Manual, Section 7.2: `https://tomopt.com/docs/knitro/tomlab_knitro008.php`.

for $t = 1, \ldots, n$. $\theta^\dagger$ is the parameter of interest. Set $p \geq 1$, following Gourieroux and Monfort (1996, Ch4.3), $\theta^\dagger$ is estimated by matching coefficients from an auxiliary AR(p) model:

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + u_t.$$

For $p = 1$, $\hat{\beta}_1 \xrightarrow{p} -\theta^\dagger/(1 + \theta^{\dagger 2})$ defines the moment condition:

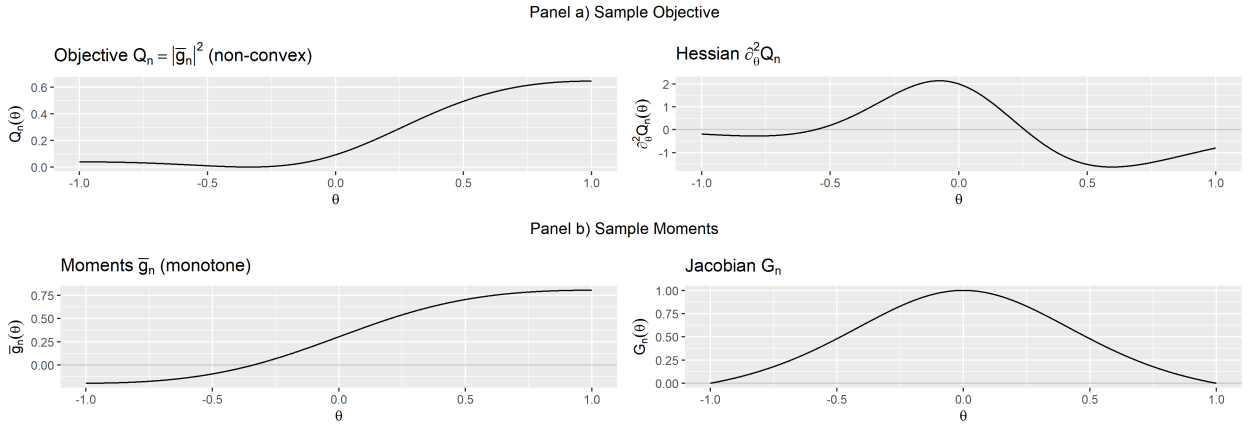
$$\overline{g}_n(\theta) = \hat{\beta}_1 + \frac{\theta}{1 + \theta^2},$$

with Jacobian $G_n(\theta) = (1 - \theta^2)/(1 + \theta^2)^2 > 0$ for any $\theta \in (-1, 1)$ and $G_n(\theta) = 0$ for $\theta \in \{-1, 1\}$. It has full rank on any interval of the form $[-1+\varepsilon, 1-\varepsilon]$, $\varepsilon \in (0, 1)$. However, Figure 1 shows that the Hessian $\partial^2_{\theta,\theta} Q_n(\theta)$ can be positive, negative, or equal to zero depending on the value of θ – the GMM objective Q_n is non-convex. Now notice that:

$$\overline{g}_n(\theta) = \partial_\theta F_n(\theta) \text{ where } F_n(\theta) = \hat{\beta}_1 \theta + \frac{1}{2} \log(1 + \theta^2)$$

which is convex on $[-1, 1]$, strongly convex on any $[-1 + \varepsilon, 1 - \varepsilon]$, $\varepsilon \in (0, 1)$. The two, F_n and Q_n, are minimized at the same solution $\hat{\theta}_n$. From a statistical perspective, Q_n and F_n define identical M-estimates. However, one involves a convex minimization while the other does not. Notice that because the gradient of F_n is $\overline{g}_n$, and its Hessian is G_n, a NR update for F_n coincides with a GN update for $\overline{g}_n$. Implicitly, GN minimizes the convex F_n – whereas NR explicitly minimizes the non-convex Q_n.

Figure 1: MA(1): illustration of non-convexity and the rank condition



Legend: simulated sample of size $n = 200$, $\theta^\dagger = -1/2$, $\overline{g}_n(\theta) = \hat{\beta}_1 - \theta/(1 + \theta^2)$, $W_n = I_d$. The GMM objective (panel a) is non-convex but the sample moments (panel b) satisfy the rank condition.

Table 2 shows the search paths for NR and GN with a fixed $\gamma = 0.1$ as well as R's built-in *optim*'s BFGS implementation and the bound-constrained L-BFGS-B. NR diverges, because the objective is locally concave at $\theta_0 = -0.6$. This is surprising given how close θ_0 is to the true value $\theta^\dagger$. GN converges steadily from the same staring value to $\hat{\theta}_n$. BFGS is more erratic, especially when $\theta_k \simeq -0.5$, i.e. $k = 1$, leading to a search outside the unit circle ($k = 2$), before reaching an area where the iterations are better behaved ($k = 3$ onwards). A natural solution is to introduce bounds using L-BFGS-B. The search, however, remains somewhat erratic as seen in the Table. Compare these to BFGS* and L-BFGS-B* which minimize F_n, instead of Q_n, using the same *optim*. Like GN, they steadily converge to $\hat{\theta}_n$.

Table 2: MA(1): search paths for NR, GN, BFGS, and L-BFGS-B

k	0	1	2	3	4	5	6	7	8	...	99
	$p = 1$										
NR	-0.600	-0.689	-0.722	-0.749	-0.772	-0.793	-0.811	-0.828	-0.843	...	-0.993
GN	-0.600	-0.560	-0.529	-0.504	-0.484	-0.466	-0.451	-0.438	-0.427	...	-0.338
BFGS	-0.600	-0.505	4.425	-0.307	-0.359	-0.338	-0.337	-0.337	-0.337	...	-0.337
L-BFGS-B	-0.600	-0.505	1.000	-0.455	-0.375	-0.318	-0.341	-0.339	-0.338	...	-0.338
BFGS*	-0.600	-0.462	-0.286	-0.345	-0.340	-0.338	-0.338	-0.338	-0.338	...	-0.338
L-BFGS-B*	-0.600	-0.462	-0.286	-0.345	-0.339	-0.338	-0.338	-0.338	-0.338	...	-0.338
	$p = 12$										
NR	0.950	0.956	0.961	0.965	0.969	0.972	0.975	0.978	0.980	...	1.000
GN	0.950	0.890	0.860	0.834	0.810	0.787	0.763	0.740	0.715	...	-0.623
BFGS	0.950	-8.290	-8.279	-8.267	-8.256	-8.244	-8.233	-8.221	-8.209	...	-6.979
L-BFGS-B	0.950	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	...	-1.000

Legend: simulated data with sample size $n = 200$, $\theta^\dagger = -1/2$. For $p = 1$, $\overline{g}_n(\theta) = \hat{\beta}_1 - \theta/(1+\theta^2)$. For $p = 12$, $\overline{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$ where $\beta(\theta)$ is the p-limit of the AR(p) coefficients, evaluated at θ. $W_n = I_d$. The solutions are $\hat{\theta}_n = -0.339$ ($p = 1$) and $\hat{\theta}_n = -0.626$ ($p = 12$). NR = Newton-Raphson, GN = Gauss-Newton. The learning rate is $\gamma = 0.1$ for NR and GN. BFGS = R's *optim*, L-BFGS-B = R's *optim* with bound constraints $\theta \in [-1, 1]$. BFGS* and L-BFGS-B* apply the same optimizers to F_n instead of Q_n. Additional results for GN using a range of values $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ can be found in Appendix C.1, Figures C7, C10.

For $p = 12$, the model becomes over-identified, and the condition for global convergence requires $G_n(\theta_1)' W_n G_n(\theta_2)$ to be non-singular for all pairs $(\theta_1, \theta_2) \in \Theta \times \Theta$. For just-identified models, this amounts to $G_n(\theta)$ non-singular for all $\theta \in \Theta$. Figure C8 in Appendix C illustrates, similar to Figure 1, that Q_n is non-convex and that the rank condition holds for $\Theta = [-1 + \varepsilon, 1 - \varepsilon]$. F_n is not defined because of over-identification. Table 2 shows that NR, BFGS and L-BFGS-B all fail to converge from $\theta_0 = 0.95$, a starting value with negative curvature.[7] Compare with GN, which steadily converges to $\hat{\theta}_n$. Starting closer to the solution,

[7] L-BFGS-B relies on projection descent which maps search directions outside the unit circle to -1 or 1 where $\partial_\theta Q_n(-1) = \partial_\theta Q_n(1) = 0$, a stationary point for (1).

BFGS and L-BFGS-B also fail to converge using $\theta_0 = 0.6$; GN remains accurate (not reported). R code to replicate the results with $p = 12$, $W_n = I_d$ can be found in Appendix B.

3.1 Correctly-specified models

The first set of results concerns models that are correctly specified: the minimizer $\hat{\theta}_n$ is such that $\overline{g}_n(\hat{\theta}_n) = O_p(n^{-1/2})$. The following Proposition shows that for any tuning parameter γ, there exists a neighborhood where (1) is locally convergent.

Proposition 1 (Local Convergence)**.** *If Assumptions 1-2 hold, then for $\gamma \in (0,1)$ small enough, there exist $R_n \geq 0$ and $\tilde{\gamma} \in (0,1)$ such that with probability approaching 1:*

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1-\tilde{\gamma})^{k+1}\|\theta_0 - \hat{\theta}_n\| \tag{2}$$

for any $\|\theta_0 - \hat{\theta}_n\| \leq R_n$. For just-identified models, $\overline{g}_n(\hat{\theta}_n) = 0$, $R_n > 0$ with probability 1. For over-identified models, $\overline{g}_n(\hat{\theta}_n) = O_p(n^{-1/2})$, $R_n > 0$ with probability approaching 1.

A proof specialized to GN, and the general case are given in Appendix A. For GN, $R_n = \min(R_G, \tilde{R}_n)$ is the smallest of R_G and:

$$\tilde{R}_n = (1-\tilde{\gamma}/\gamma)\frac{\underline{\sigma}}{L\sqrt{\kappa_W}} - \frac{1}{\underline{\sigma}\sqrt{\underline{\lambda}_W}}\|\overline{g}_n(\hat{\theta}_n)\|_{W_n},$$

where $\kappa_W = \overline{\lambda}_W/\underline{\lambda}_W$ bounds the condition number of the weighting matrix W_n. Having $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \neq 0$ reduces the area of local convergence. For correctly specified models $\overline{g}_n(\hat{\theta}_n) = O_p(n^{-1/2})$ implies $\tilde{R}_n \xrightarrow{p} \tilde{R} = (1-\tilde{\gamma}/\gamma)\underline{\sigma}\sqrt{\kappa_W}/L > 0$. Note that for GN, Proposition 1 holds for any choice of $\gamma \in (0,1)$. This is typically not the case for other choices of P_k: GD is only locally convergent when γ is sufficiently small. GD and GN iterations require the same inputs G_n and $\overline{g}_n$, but the latter is preferred since it converges more quickly. Because NR and QN iterations require the exact and an approximate Hessian, they are more costly than GD, GN.

3.1.1 Just-Identified Models

The Theorem below proves global convergence for $\gamma \in (0,1)$ sufficiently small – under additional restrictions on G_n.

Theorem 1 (Global Convergence, Just-Identified)**.** *Suppose $\overline{g}_n(\hat{\theta}_n) = 0$, Assumptions 1-2 hold, and with probability approaching 1:*

$$\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0, \, \forall \theta \in \Theta. \tag{3}$$

Then for γ small enough, there exist a $\overline{\gamma} \in (0,1)$, and constants $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$ such that:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})^{k+1} \sqrt{\overline{\lambda}/\underline{\lambda}} \|\theta_0 - \hat{\theta}_n\|, \tag{4}$$

for any starting value $\theta_0 \in \Theta$, with probability approaching 1.

The proof is given in Appendix A. The main steps are to show that for γ sufficiently small, we have: i) $Q_n(\theta_{k+1}) \leq (1-\overline{\gamma})^2 Q_n(\theta_k)$ for some $\overline{\gamma} \in (0,1)$ under condition (3) and Assumptions 1-2. Iterating on this inequality implies convergence of the objective function: $Q_n(\theta_k) \leq (1-\overline{\gamma})^{2k} Q_n(\theta_0)$. The same assumptions also imply the norm equivalence: ii) $\underline{\lambda}\|\theta - \hat{\theta}_n\|^2 \leq Q_n(\theta) \leq \overline{\lambda}\|\theta - \hat{\theta}_n\|^2$ for some $0 < \underline{\lambda} \leq \overline{\lambda} < +\infty$. Together, these two properties imply convergence of θ_k to $\hat{\theta}_n$.

The main takeaway from Theorem 1 is that global convergence can be achieved using any Algorithm which has P_k strictly positive definite for any $k \geq 0$, with an adequate choice of $\gamma \in (0,1)$, if G_n is everywhere non-singular. This assumption does not imply the convexity of Q_n. Note that the choice of γ depends on the choice of algorithm, through P_k. Some methods are associated with faster convergence than others, which is measured by $\overline{\gamma}$. The following discusses the assumptions and the implications for the choice of algorithm.

Discussion of the rank condition. In the scalar case, where both θ and $\overline{g}_n$ are one-dimensional, continuity of G_n and the rank condition (3) implies that $\overline{g}_n$ is strictly monotone, i.e. injective. However, this does not imply that Q_n is convex, as illustrated with the MA(1) example above. Under strict monotonicity, univariate methods such as bisection or golden-search converge at a similar rate but do not extend to multivariate estimations.

In the multivariate case, (3) implies a unique solution since $0 = \overline{g}_n(\theta) = G_n(\tilde{\theta}_n)(\theta - \hat{\theta}_n) \Leftrightarrow \theta = \hat{\theta}_n$, for an intermediate value $\tilde{\theta}_n$. It further implies that Q_n has no local minimum, besides $\hat{\theta}_n$, since $\partial_\theta Q_n(\theta) = G_n(\theta)' W_n \overline{g}_n(\theta) = 0 \Leftrightarrow \overline{g}_n(\theta) = 0 \Leftrightarrow \theta = \hat{\theta}_n$, for just-identified models.

Unlike convexity, the assumption is invariant to one-to-one non-linear reparameterizations. Take any $h(\vartheta) = \theta$ where $\partial_\vartheta h(\vartheta)$ has full rank for all $\vartheta \in h^{-1}(\Theta)$. Consider the reparameterized sample moments $\overline{g}_n \circ h(\vartheta)$. The Jacobian $\partial_\vartheta \overline{g}_n \circ h(\vartheta) = G_n \circ h(\vartheta) \times \partial_\vartheta h(\vartheta)$ has full rank if, and only if, G_n has full rank. Hence, (3) is primitive to the model and the choice of moments since it is not parameterization specific, unlike convexity.

Choice of Algorithm. The conditioning matrix P_k must be finite and strictly positive definite for all $k \geq 0$. This is always the case for GD since $P_k = I_d$ for all k. When G_n has full rank everywhere, GN also satisfies the assumption since $P_k = [G_n(\theta_k)' W_n G_n(\theta_k)]^{-1}$.

However, if Q_n is non-convex, then $\partial^2_{\theta,\theta'} Q_n$ can be singular, and P_k used in NR and QN may not be finite. This can negate the convergence result, as the MA(1) example illustrates.

When the rank condition (3) fails, Assumption 2 does not holds for GN but remains valid for GD. One could apply the Levenberg-Marquardt (LM) algorithm by setting $P_k = (G_n(\theta_k)' W_n G_n(\theta_k) + \lambda I_d)^{-1}$, and Assumption 2 holds for any $\lambda > 0$. Global convergence is not guaranteed, however, since non-global local optima may exist when (3) does not hold.

Theorem 1 is related to solving just-identified non-linear systems of equations, here of the form: $\overline{g}_n(\theta) = 0$. This is less studied than non-linear optimization. Dennis and Schnabel (1996) cast the problem as of minimizing Q_n, using the present notation, and derive global convergence results to a local minimum (Theorems 6.3.3-6.3.4). Deuflhard (2005, Ch3) gives conditions for convergence to a global solution. However, these results require an exact solution $\overline{g}_n(\hat{\theta}_n) = 0$ and, as discussed below, (3) may not suffice under over-identification.

3.1.2 Over-Identified Models

Theorem 2 (Global Convergence, Over-Identified)**.** *Suppose $\overline{g}_n(\hat{\theta}_n) = O_p(n^{-1/2})$, Assumptions 1-2, and for some $\rho \in (0, 1]$, with probability approaching 1:*

$$\sigma_{\min}[G_n(\theta_1)' W_n G_n(\theta_2)] \geq \rho \underline{\sigma}^2 \underline{\lambda}_W > 0, \forall (\theta_1, \theta_2) \in \Theta \times \Theta. \tag{3'}$$

Then for γ small enough, there exist $\overline{\gamma} \in (0, 1)$, $C > 0$, $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$, and $C_n = O_p(1)$ such that with probability approaching 1:

$$\|\theta_k - \hat{\theta}_n\|^2 \leq (1 - \overline{\gamma})^{2k} \frac{\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}{\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}} \|\theta_0 - \hat{\theta}_n\|^2 + C_n \|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n}, \tag{5}$$

for any $\theta_0 \in \Theta$. Given this choice of γ, take R_n from Proposition 1. Since $C_n \|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n} \leq R_n^2/2$ with probability approaching 1, setting $k = k_n + j$, $j \geq 0$ implies:

$$\|\theta_k - \hat{\theta}_n\| \leq (1 - \tilde{\gamma})^j R_n,$$

where $\tilde{\gamma} \in (0, 1)$ is the local rate in Proposition 1 and $k_n \geq \frac{2\log(R_n) - \log 2 - \log(d_{0n})}{2\log(1-\overline{\gamma})}$ with $d_{0n} = 2[\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1}[\|\overline{g}_n(\theta_0)\|^2_{W_n} - \|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n}]$.

The explicit formula for $\underline{\lambda}, \overline{\lambda}, C$, and C_n are given in the proof of the Theorem (Appendix A). As discussed earlier, conditions (3) and (3') are equivalent for just-identified models, when W_n has full rank. Notice that larger values of $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ can degrade convergence. The

results for misspecified models will investigate the robustness of the results when $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ does not vanish in the limit.

Condition (3') extends the monotonicity condition found in the scalar case. When $\overline{g}_n$ and θ are both scalar, $\overline{g}_n$ is strictly increasing (or decreasing) only if $\partial_\theta \overline{g}_n(\theta_1)\partial_\theta \overline{g}_n(\theta_2) \neq 0$ for all θ_1, θ_2; the derivative does not change sign. Here the condition reads as $\partial_\theta \overline{g}_n(\theta_1)' W_n \partial_\theta \overline{g}_n(\theta_2)$ non-singular for θ_1, θ_2. With the continuity of the Jacobian, this yields (3') for some $\rho \in (0, 1]$. Hence (3') reads as a multivariate strict monotonicity condition on $W_n^{1/2}\overline{g}_n(\cdot)$.

To give some intuition why (3) in Theorem 1 alone may not suffice when the model is over-identified, consider the population problem $g(\theta) = 0$. Suppose $G(\theta)$ has full rank for all $\theta \in \Theta$. Suppose θ is a local optimum: $G(\theta)'Wg(\theta) = 0$, then it is a fixed point of (1). Global convergence can only hold if $\theta = \theta^\dagger$ is the only fixed point. Using $g(\theta^\dagger) = 0$, write $G(\theta)'W[g(\theta) - g(\theta^\dagger)] = 0 = G(\theta)'WG(\tilde{\theta})[\theta - \theta^\dagger]$, for some intermediate value $\tilde{\theta}$. This implies that if $G(\theta)'WG(\tilde{\theta})$ has full rank then θ satisfies the first-order condition *only if* $\theta = \theta^\dagger$. Because the matrix $G(\theta)$ is rectangular for over-identified models, (3) is not sufficient to rule out $G(\theta)'WG(\tilde{\theta})$ singular, and thus the existence of local optima.[8] Notice that (3') could be weakened to: $G(\theta)'WG(\tilde{\theta})$ has full rank for all $\theta \in \Theta$ and intermediate values $\tilde{\theta} = \omega\theta + (1-\omega)\theta^\dagger$, $\omega \in (0, 1)$. Likewise, Theorem 2 could be derived assuming (3') only for $\theta_1 \in \Theta$ and $\theta_2 = \omega\theta_1 + (1-\omega)\hat{\theta}_n$, $\omega \in (0, 1)$ since the derivations also involve intermediate values.[9] The following explains why (3') is preferred.

Condition (3') is invariant to one-to-one non-linear reparameterization since $\partial_\vartheta \overline{g}_n \circ h(\vartheta) = G_n \circ h(\vartheta) \times \partial_\vartheta h(\vartheta)$ and (3') becomes $\partial_\vartheta h(\vartheta_1)' G_n \circ h(\vartheta_1)' W_n G_n \circ h(\vartheta_2) \partial_\vartheta h(\vartheta_2)$ which has full rank if $G_n \circ h(\vartheta_1)' W_n G_n \circ h(\vartheta_2)$ has full rank for all $h(\vartheta_1)$, $h(\vartheta_2)$ given that $\partial_\vartheta h(\vartheta_1), \partial_\vartheta h(\vartheta_2)$ are full rank square matrices. The weaker version of (3') discussed above would not be invariant to non-linear reparameterizations since $h(\omega\vartheta + (1-\omega)\hat{\vartheta}_n)$ need not be an intermediate value between $h(\vartheta)$ and $h(\hat{\vartheta}_n)$ for a non-linear $h(\cdot)$.

Condition (3') is not invariant to the choice of W_n. Appendix C illustrates that the condition may hold for some W_n (Figure C8) but not for another (Figure C9). This is perhaps not surprising as changes in W_n can alter the null space of $G_n(\theta_1)'W_n$.

Another pen and pencil example. Take $y_i \sim \mathcal{N}(\mu, \sigma^2)$, the parameters of interest are $\theta = (\mu, \sigma^2)$. Compute the sample moments $\hat{\mu}_n = (\hat{\mu}_{n1}, \hat{\mu}_{n2}, \hat{\mu}_{n4})'$, where $\hat{\mu}_{n1} = \overline{y}_n$, $\hat{\mu}_{n2} = \hat{\sigma}_n^2$, and $\hat{\mu}_{n4} = 1/n \sum_{i=1}^n (y_i - \overline{y}_n)^4$, and let: $\overline{g}_n(\theta) = \hat{\mu}_n - (\mu, \sigma^2, 3\sigma^4)$. Set $W_n = I_d$ and take $\theta^\dagger = (0, 1)$. A quick numerical computation reveals the population objective function is

[8] Take $G(\theta_1)' = (1, 0)$ and $G(\theta_2)' = (0, 1)$, both have full rank and yet $G(\theta_1)'G(\theta_2) = 0$ is singular.

[9] More specifically, the condition is used to bound (A.5) which involves θ_k and an intermediate value $\tilde{\theta}_k$.

non-convex: the eigenvalues of $\partial^2_{\theta,\theta'}Q(\theta)$ are $(74, 2)$ at $\theta = (0, 1)$ and $(2, -7)$ at $\theta = (0, 1/2)$ – the Hessian is positive definite at the true value but not everywhere. For starting values such that $\partial^2_{\theta,\theta'}Q_n(\theta)$ is (near)-singular, NR and QN iterations can be erratic, as in the MA(1) example. Nonetheless, condition (3') holds since:

$$G_n(\theta)' = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & -6\sigma^2 \end{pmatrix}, \quad G_n(\theta_1)'G_n(\theta_2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 + 36\sigma_1^2\sigma_2^2 \end{pmatrix}$$

is positive definite for any two $\theta_1 = (\mu_1, \sigma_1^2)$, $\theta_2 = (\mu_2, \sigma_2^2)$ with $\sigma_1, \sigma_2 \geq 0$. In this simple example, the Hessian of Q_n can be singular, yet condition (3') holds globally.

Much like convexity, condition (3') can be challenging to verify analytically for more complex complex models. Still, it can be evaluated numerically on a grid, as illustrated for the MA(1) model in Figure C8, Appendix C. Unlike just-identified models, condition (3') does not immediately exclude local optima in finite samples. Still, the following Proposition shows that all local optima θ_n are asymptotically valid estimators, i.e. $\sqrt{n}(\theta_n - \hat{\theta}_n) = o_p(1)$.

Proposition 2 (Local optima are asymptotically valid estimators)**.** *Suppose Assumption 1 and (3') hold. If $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} = O_p(n^{-1/2})$, then for any θ_n such that $G_n(\theta_n)'W_n\overline{g}_n(\theta_n) = 0$:*

$$\|\theta_n - \hat{\theta}_n\| \leq \overline{C}\|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n} = O_p(n^{-1}),$$

where $\overline{C} > 0$ only depends on G_n and W_n.

Proposition 1 implies local convergence to $\hat{\theta}_n$. And yet, Proposition 2 alone does not exclude the presence of local optima θ_n near $\hat{\theta}_n$ to which (1) could converge. These two results are not at odds, however. Proposition 2 and local convexity imply: $\theta_n = \hat{\theta}_n$ with probability approaching 1. The following sketches the argument. Take θ_n, a local optimum: $\partial_\theta Q_n(\theta_n) = 0$ and $\|\tilde{\theta}_n - \hat{\theta}_n\| \leq O_p(n^{-1})$, by Proposition 2. Then $0 = \partial_\theta Q_n(\theta_n)(\theta_n - \hat{\theta}_n) = (\theta_n - \hat{\theta}_n)'\partial^2_{\theta,\theta'}Q_n(\tilde{\theta}_n)(\theta_n - \hat{\theta}_n) \geq \lambda_{\min}[\partial^2_{\theta,\theta'}Q_n(\tilde{\theta}_n)]\|\theta_n - \hat{\theta}_n\|^2$ for an intermediate value $\tilde{\theta}_n$. Also, $\partial^2_{\theta,\theta'}Q_n(\tilde{\theta}_n) = \partial^2_{\theta,\theta'}Q_n(\hat{\theta}_n) + o_p(1) = G_n(\hat{\theta}_n)'W_nG_n(\hat{\theta}_n) + o_p(1)$, by continuity, for correctly specified models. This implies $\lambda_{\min}[\partial^2_{\theta,\theta'}Q_n(\tilde{\theta}_n)] > 0$, strictly, with probability approaching 1. As desired, we have that $\|\theta_n - \hat{\theta}_n\| = 0$ with probability approaching 1.

3.2 Misspecified models

So far, the results imply that fast global convergence is feasible under a rank condition for correctly specified models. In applications, misspecification can be a concern so that

understanding the robustness of the results above to non-negligible deviations from this baseline is empirically relevant. Recently, Hansen and Lee (2021) studied the properties of iterated GMM procedures. Here the focus is on computing a GMM estimate. The following considers "moderate" amounts of misspecification in the sense that:

$$\text{plim}_{n\to\infty} Q_n(\hat{\theta}_n) := \varphi^2 \geq 0$$

exists and can be non-zero in the limit. When $\varphi > 0$, the degree of misspecification is non-negligible asymptotically and, with optimal weighting, the J-statistic $n\|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n} \to \infty$ can diverge. However, φ cannot be too large for the local and global convergence results to hold as shown below. For simplicity, only Gauss-Newton will be considered in the results. Also, since G_n cannot be full rank at $\theta = \hat{\theta}_n$ when the model is both just-identified and misspecified,[10] the results presented here solely consider over-identified models.

For correctly specified models, an over-identification test can diagnose global convergence (Andrews, 1997, Sec3.3). For misspecified models, such test would frequently reject in large samples. Then the issue is that, when the test rejects, either 1) the estimates are not valid, or 2) the model fits the data poorly in some dimension(s). When Q_n is globally convex, the estimates are the global solution if, and only if, they satisfy the first and second order optimality conditions. Without convexity, this only guarantees a local optimum. Here, checking the rank condition (3'), as in Figure C7, helps determine whether global convergence is at risk, or not, which is the main concern when $Q_n(\hat{\theta}_n)$ is significantly non-zero.

Proposition 3 (Local Convergence, Misspecified)**.** *Suppose Assumptions 1-2 hold, and φ is such that:*

$$0 \leq \varphi < \frac{\sigma^2 \underline{\lambda}_W}{L \overline{\lambda}_W^{1/2}}. \tag{6}$$

For any $\gamma \in (0,1)$, there exists $\tilde{\gamma} \in (0,\gamma)$, such that, with probability approaching 1, for some $R_n > 0$, strictly positive, any $\|\theta_0 - \hat{\theta}_n\| \leq R_n$, and all $k \geq 0$:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1-\tilde{\gamma})^{k+1}\|\theta_0 - \hat{\theta}_n\|. \tag{2'}$$

Also, $plim_{n\to\infty} R_n = R > 0$.

[10]The solution $\hat{\theta}_n$ is s.t. $G_n(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) = 0$, misspecification implies $\overline{g}_n(\hat{\theta}_n) \neq 0$, and since W_n has full rank, it must be that $G_n(\hat{\theta}_n)$ is singular for just-identified models. For over-identified models, $\overline{g}_n(\hat{\theta}_n)$ is in the null space of $G_n(\hat{\theta}_n)' W_n$, which allows $G_n(\hat{\theta}_n)$ to be full rank.

R_n in Proposition 3 takes the same form as in Proposition 1, (6) ensures that the corresponding R is strictly positive in the limit; the neighborhood of convergence is non-negligible asymptotically. Under identity weighting, $W_n = I_d$, (6) only depends on $\underline{\sigma}$ and L. For linear models, $L = 0$ implies that any $\varphi \in [0, \infty)$ is feasible. For non-linear models, a larger $L > 0$ requires a smaller φ: increased non-linearity requires milder misspecification. Smaller values of $\underline{\sigma}$, which measures local identification, also require a smaller φ. In Proposition 1 with GN, any rate of convergence $\overline{\gamma} \in (0, \gamma)$ can be used. When the model is misspecified, larger values of $\varphi \geq 0$ require $\tilde{\gamma} \in (0, \gamma)$ to be sufficiently small.

Theorem 3 (Global Convergence, Misspecified)**.** *Suppose Assumptions 1-2 and (3') hold. If φ is such that:*

$$0 \leq \varphi < \frac{\underline{\sigma}^2 \underline{\lambda}_W}{2L\overline{\lambda}_W^{1/2}}. \tag{6'}$$

Then for γ small enough, there exist $\overline{\gamma} \in (0,1)$ and $C_n = O_p(1)$, $0 < C, \underline{\lambda}, \overline{\lambda} < \infty$, which do not depend on φ, such that with probability approaching 1:

$$\|\theta_k - \hat{\theta}_n\|^2 \leq (1-\overline{\gamma})^{2k} \frac{\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}{\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}} \|\theta_0 - \hat{\theta}_n\|^2 + C_n \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2. \tag{5'}$$

Let $\Delta = 1/2[\underline{\sigma}^2 \underline{\lambda}_W - 2L\overline{\lambda}_W^{1/2}\varphi] > 0$. Suppose $\gamma \in (0,1)$ and $\varphi \geq 0$ are such that:

$$\frac{\Delta\gamma^2 c_2 + 2\gamma \bar{c}_3^2/\underline{c}_1}{[\gamma \underline{c}_1/2 - \gamma^2 c_2]\Delta^2} \varphi^2 < \left((1-\varepsilon)\frac{\underline{\sigma}}{L\sqrt{\kappa_W}} - \frac{\varphi}{\underline{\sigma}\sqrt{\underline{\lambda}_W}} \right)^2, \tag{7}$$

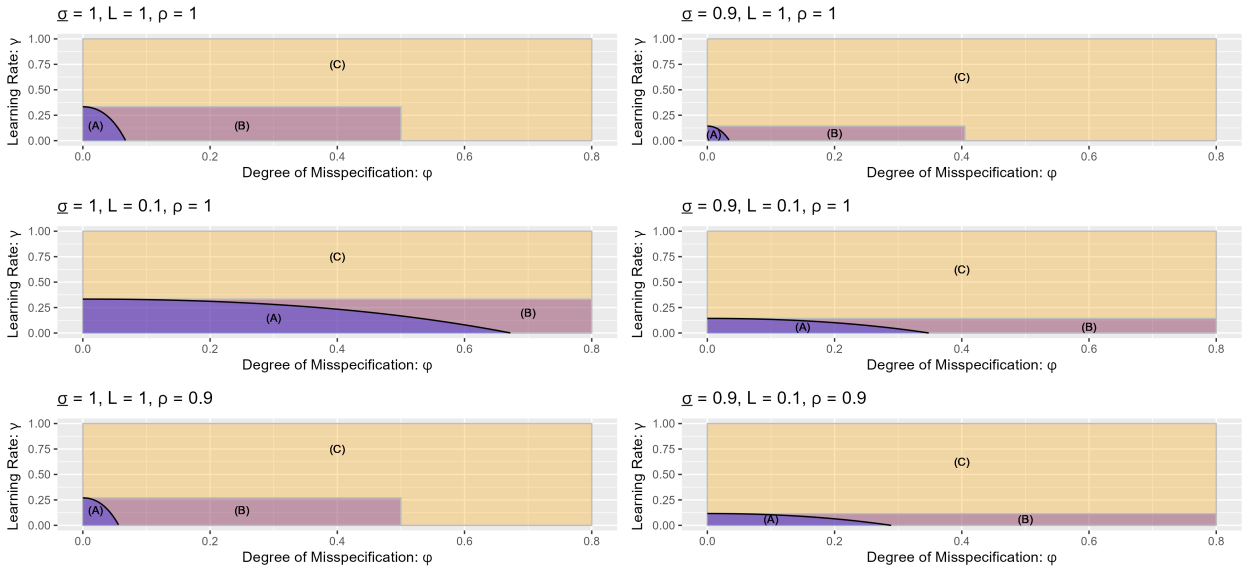
for some $\varepsilon \in (0,1)$, where $\underline{c}_1 = 2/3\rho^2([\underline{\sigma}/\overline{\sigma}]^2 \kappa_W^{-1})^2$, $c_2 = L_Q(\overline{\sigma}\overline{\lambda}_W^{1/2}/[\underline{\sigma}^2\underline{\lambda}_W])^2$, $\bar{c}_3 = 2\overline{\sigma}\overline{\lambda}_W^{1/2}$, $\kappa_W = \overline{\lambda}_W/\underline{\lambda}_W$, and L_Q is the Lipschitz constant of $\partial_\theta Q_n$. Take any $\tilde{\gamma} \in (0, \varepsilon\gamma)$ and R_n from Proposition 3, set $k = k_n + j$, $j \geq 0$, then with probability approaching 1:

$$\|\theta_k - \hat{\theta}_n\| \leq (1-\tilde{\gamma})^j R_n,$$

where $k_n \geq \frac{2\log(R_n) + \log(\delta) - \log(d_{0n})}{2\log(1-\overline{\gamma})}$, with d_{0n} as in Theorem 2, for some small enough $\delta \in (0,1)$.

Theorem 3 above shows that for "moderate" amounts of misspecification, GN remains globally convergent for an appropriate choice of tuning parameter $\gamma \in (0,1)$. Three conditions are required for the result to hold: (6') bounds φ from above; (5') bounds γ from above – independently of φ. Condition (7) restricts the pair (γ, φ) to prove convergence to

Figure 2: Illustration of the conditions for (γ, φ) in Theorem 3



Legend: (A) all 3 conditions hold, (B) condition (7) does not hold, (C) all conditions fail. Black curve: upper bound for (7). Figure drawn using $\overline{\sigma} = 1, W_n = I_d, L_Q = 1$, and $\varepsilon = 10^{-4}$.

the global solution. Figure 2 illustrates the 3 conditions and the effect $\underline{\sigma}$, L, and ρ have on the feasible set for (γ, φ). Everything else held equal, a smaller value of $\underline{\sigma}$ or ρ shrinks the feasible set. A smaller L, implies less non-linearity, offsets these two and expands the feasible set for φ. Generally, larger values of φ restrict the learning rate γ to be smaller.

4 Empirical Applications

4.1 Estimation of a Random Coefficient Demand Model Revisited

The following revisits the results for random coefficient demand estimation in Knittel and Metaxoglou (2014) with the cereal data from Nevo (2001).[11] This is a non-linear instrumental variable regression with sample moment conditions: $\overline{g}_n(\theta, \beta) = \frac{1}{n} \sum_{j,t} z_{jt}[\delta_{jt}(\theta) - x'_{tj}\beta]$, where z_{jt} are the instruments, x_{jt} the linear regressors in market j at time period t. The 8

[11] It available in the R package BLPestimatoR (Brunner et al., 2017). The data consist of 2,256 observations for 24 products (brands) in 47 cities over two quarters, in 94 markets. The specification is identical to Nevo's, with cereal brand dummies, price, sugar content (sugar), a mushy dummy indicating whether the cereal gets soggy in milk (mushy), and 20 IV variables.

parameters of interest are the random coefficients θ,[12] which enter δ_{jt}, recovered from market shares s_{jt} using the fixed point algorithm of Berry et al. (1995).[13] The 25 linear coefficients β are nuisance parameters concentrated out by two-stage least squares for each θ.

Table 3: Demand for Cereal: performance comparison

		STDEV				INCOME				objs	# of crashes
		const.	price	sugar	mushy	const.	price	sugar	mushy		
TRUE	est	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	-
	se	0.11	0.76	0.01	0.15	0.56	3.06	0.02	0.26	-	
GN	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
BFGS	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	30
	std	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	
NM	avg	0.32	0.35	-0.08	-0.88	3.94	-2.64	-0.10	1.26	628.44	4
	std	1.37	8.91	0.09	3.08	3.63	10.74	0.23	5.13	772.23	
SA	avg	0.87	-0.58	-0.72	-0.00	0.01	0.33	1.64	-1.16	$1.46 \cdot 10^5$	3
	std	7.68	8.66	3.58	7.88	6.67	6.97	3.65	7.92	$2.36 \cdot 10^5$	
SA+NM	avg	0.43	-0.88	-0.06	-0.84	4.15	-2.18	-0.15	0.71	506.44	3
	std	0.61	9.45	0.12	2.25	3.56	11.48	0.19	5.06	1250.65	

Legend: Comparison for 50 starting values in $[-10, 10] \times \cdots \times [-10, 10]$. Avg, Std: sample average and standard deviation of optimizer outputs. TRUE: full sample estimate (est) and standard errors (se). Objs: avg and std of minimized objective value. # of crashes: optimization terminated because the objective function returned an error. GN run with $\gamma = 0.1$ for $k = 150$ iterations for all starting values. Additional results for GN using a range of values $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ can be found in Appendix C.2, Table C5.

Table 3 and Figure 3 compare the performance of quasi-Newton (BFGS), Nelder-Mead (NM), Simulated-Annealing (SA), and Nelder-Mead after Simulated-Annealing (SA+NM), using R's default optimizer *optim*, with Gauss-Newton (GN) for 50 different starting values.[14] As reported in Knittel and Metaxoglou (2014), optimization can crash often.[15] Crashes could be avoided using error handling (try-catch statements). However, this may not be enough to produce accurate estimates as the next application will illustrate.[16] Only GN and BFGS systematically produce accurate estimates, but BFGS crashes 60% of the time. Derivative-free optimizers, NM, SA, and SA+NM, can produce inaccurate estimates.

[12] 8 parameters are the unobserved standard deviation and the income coefficient on the constant term, price, sugar, and mushy.

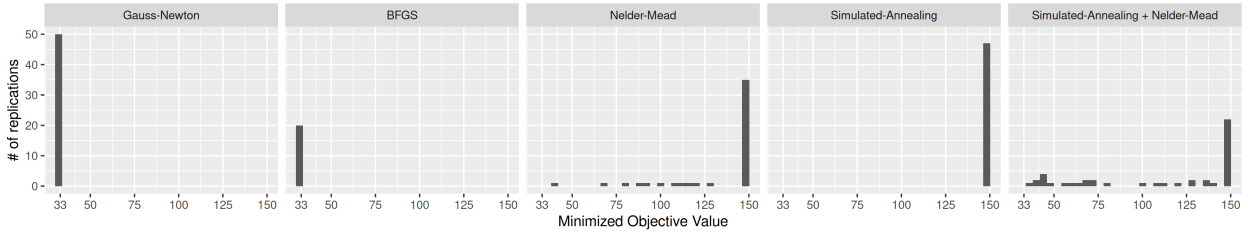
[13] The maximum number of iterations is set to 20000, the tolerance level for convergence to 10^{-12}.

[14] The solution of the contraction mapping is not well defined for all values in Θ, so we use the first 50 values produced by the Sobol sequence such that δ_{jt} is finite for all j, t.

[15] The optimizers will crash when the fixed point algorithms fail to return finite values. This is typically the case when the search direction was poorly chosen at the previous iteration.

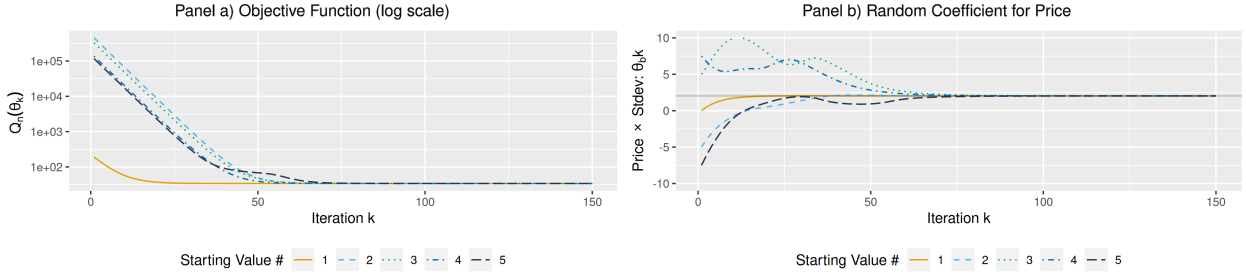
[16] Conlon and Gortmaker (2020) illustration that modifications to the fixed point algorithm and specific optimizer implementations to handle near-singularity of the Hessian can also improve performance for BFGS.

Figure 3: Demand for Cereal: distribution of minimized objective values



Legend: Comparison for 50 starting values. Minimized objective values for non-crashed optimizations. Objective values are truncated from above at $Q_n(\theta) = 150$.

Figure 4: Demand for Cereal: Gauss-Newton iterations for 5 starting values



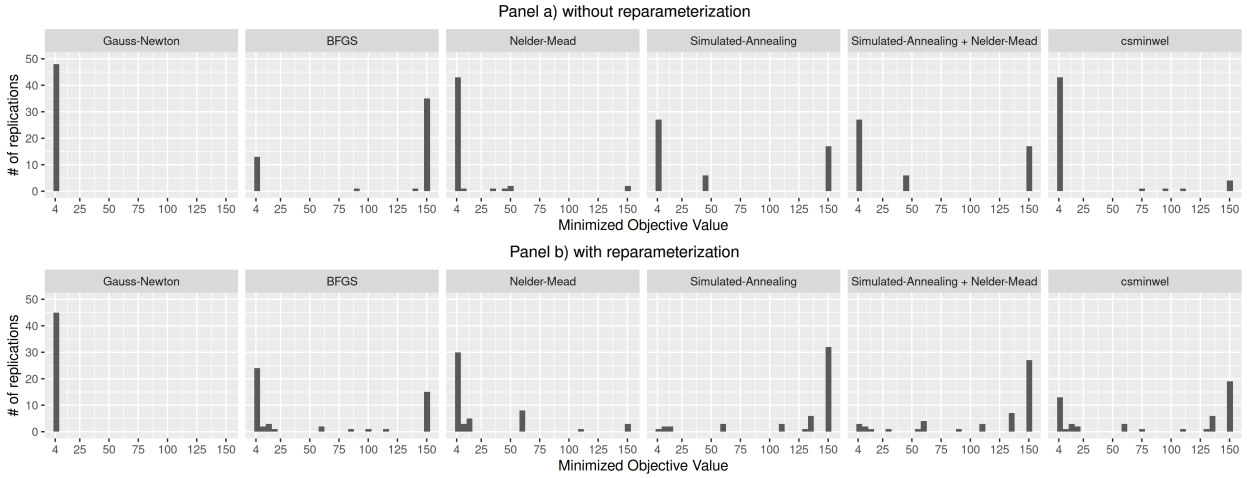
Legend: 150 GN iterations for 5 starting values in $[-10, 10] \times \cdots \times [-10, 10]$. Panel b) horizontal grey line = full sample estimate.

Figure 4, illustrates the convergence of GN for the first 5 starting values. In line with the predictions of Theorem 2, though Q_n is non-convex, GN iterations steadily converge to the solution. This type of "Gauss-Newton regression" is related to Salanié and Wolak (2022) who compute a two-stage least-squares estimate of a linear approximation of the BLP random coefficient model.

4.2 Innovation, Productivity, and Monetary Policy

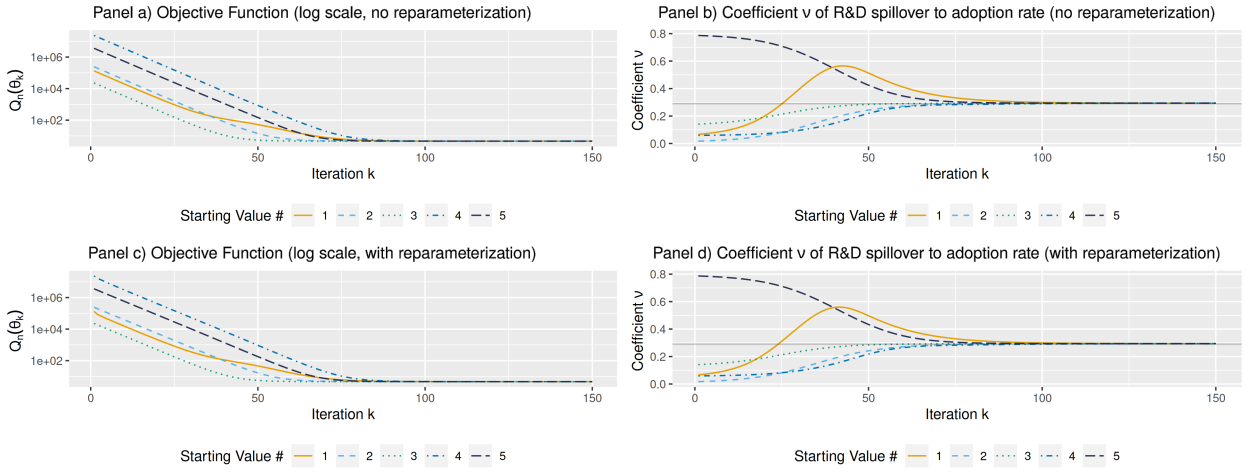
The second application revisits Moran and Queralto (2018)'s estimation of a model with endogenous total factor productivity (TFP) growth (see Moran and Queralto, 2018, Sec2, for details about the model). They estimate parameters related to Research and Development (R&D) by matching the impulse response function (IRF) of an identified R&D shock to R&D and TFP in a small-scale Vector Auto-Regression (VAR) estimated on U.S. data.

Figure 5: Impulse Response Matching: distribution of minimized objective values



Legend: Comparison for 50 starting values. Minimized objective values for non-crashed optimizations. Objective values are truncated from above at $Q_n(\theta) = 150$.

Figure 6: Impulse Response Matching: Gauss-Newton iterations for 5 starting values



Legend: 150 GN iterations for 5 non-crashing starting values. Panels a,c) value of the objective function at each iteration, Panels b,d) coefficient η at each iteration; horizontal grey line = full sample estimate.

Table 4: Impulse Response Matching: performance comparison

		η	ν	ρ_s	σ_s	objs	# of crashes
TRUE	est	0.30	0.29	0.39	0.17	4.65	-
WITHOUT REPARAMETERIZATION							
GN	avg	0.30	0.29	0.39	0.17	4.65	2
	std	0.00	0.00	0.00	0.00	0.00	
BFGS	avg	0.12	0.10	-0.34	5.42	$2.23 \cdot 10^4$	0
	std	0.56	0.20	0.47	4.77	$2.31 \cdot 10^4$	
CSMINWEL	avg	0.36	-0.00	0.27	0.15	46.42	0
	std	0.24	1.54	0.33	0.19	183.74	
NM	avg	0.47	-5.27	0.43	0.16	14.81	0
	std	0.54	37.28	0.11	0.05	34.32	
SA	avg	1.39	-2.08	0.48	0.09	75.21	0
	std	2.23	3.59	0.19	0.09	91.35	
SA+NM	avg	0.97	-84.27	0.41	0.09	66.53	0
	std	2.01	124.00	0.22	0.09	79.78	
WITH REPARAMETERIZATION							
GN	avg	0.30	0.29	0.39	0.17	4.65	5
	std	0.00	0.00	0.00	0.00	0.00	
BFGS	avg	0.37	0.21	0.07	0.14	104.08	0
	std	0.32	0.14	0.65	0.06	136.63	
CSMINWEL	avg	0.62	0.20	0.07	0.14	133.76	0
	std	0.39	0.22	0.76	0.08	123.32	
NM	avg	0.48	0.26	0.37	0.39	$1.29 \cdot 10^3$	0
	std	0.33	0.16	0.34	1.68	$8.92 \cdot 10^3$	
SA	avg	0.60	0.21	0.44	1.26	$6.93 \cdot 10^3$	0
	std	0.46	0.30	0.74	3.62	$2.06 \cdot 10^4$	
SA+NM	avg	0.61	0.21	0.43	1.08	$5.41 \cdot 10^3$	2
	std	0.45	0.29	0.71	3.33	$1.76 \cdot 10^4$	
lower bound		0.05	0.01	-0.95	0.01	-	-
upper bound		0.99	0.90	0.95	12	-	-

Legend: Comparison for 50 starting values. TRUE: full sample estimate (est). Objs: avg and std of minimized objective value. # of crashes: optimization terminated because objective returned error. Lower/upper bound used for the reparameterization. GN run with $\gamma = 0.1$ for $k = 150$ iterations for all starting values. Standard errors were not computed in the original study. Additional results for GN, using a range of values $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ can be found in Appendix C.3, Tables C6, C7

The parameters of interest are $\theta = (\eta, \nu, \rho_s, \sigma_s)$ which measure, respectively, the elasticity of technology creation to R&D, R&D spillover to adoption, the persistence coefficient and size of impulse to the R&D wedge. The sample moments are $\overline{g}_n(\theta) = \hat{\psi}_n - \psi(\theta)$, $\hat{\psi}_n$ and $\psi(\theta)$ are the sample and predicted IRFs, respectively. The latter is computed using Dynare in Matlab.

To minimize Q_n, the authors use Sims's CSMINWEL[17] algorithm with a reparameterization which bounds the coefficients.[18] This type of reparameterization is commonly used, however its yields a near-singular Jacobian near the boundary which can affect both local and global convergence, according to the results.

In the original paper, the authors intialize the estimation at $\theta_0 = (\eta_0, \nu_0, \rho_{s0}, \sigma_{s0}) = (0.20, 0.20, 0.30, 0.10)$, very close to $\hat{\theta}_n$. Here, 50 starting values are generated, within the bounds in Table 4. The model is estimated using CSMINWEL and the same set of optimizers used in the previous replication. Table 4 reports the results with and without the non-linear reparameterization. Similar to the MA(1) model with $p = 12$, without the reparameterization, several optimizers return values outside the parameter bounds which motivates the contraints in these cases. GN correctly estimates the parameters for all starting values but crashes twice for starting values for which both η and ν are close to their lower bounds where the Jacobian is nearly singular. With the reparameterization, GN crashed more often, five times in total, but is otherwise accurate. The other two gradient-based optimizers, BFGS and CSMINWEL, never crash because of better error handling in Matlab. They do encounter a number values where the model cannot be solved in Dynare, which then returns an error. This suggests they produce inaccurate search directions. They produce valid estimates less often than GN. Figure 5 illustrates that CSMINWEL is sensitive to reparameterization. Likewise, derivative-free methods can be inaccurate, as illustrated in Table 4 and Figure 5; some crashes occur despite Matlab's error handling. Finally, Figure 6 shows 5 optimization paths for which GN does not crash with and without the reparameterization. They are nearly identical. Tables C6, C7 in Appendix C.3 gives additional results for larger values of $\gamma \in (0, 1]$, plus results with error handling and the global step from Forneron (2023).

5 Conclusion

Non-convexity of the GMM objective function is an important challenge for structural estimation, and the survey highlights how practitioners approach this issue. This paper considers an alternative condition under which there are globally convergent algorithms. The results are robust to non-convexity, one-to-one non-linear reparameterizations, and moderate misspecification. Econometric theory emphasizes the role of the weighting matrix W_n on the

[17] Details about CSMINWEL and code can be found at: `http://sims.princeton.edu/yftp/optimize/`.

[18] The replication uses the mapping $\theta_j = \underline{\theta}_j + \frac{\overline{\theta}_j - \underline{\theta}_j}{1+\exp(-\vartheta_j)}$, where each ϑ_j is unconstrained. The original study relied on $\theta_j = 1/2(\overline{\theta}_j + \underline{\theta}_j) + 1/2(\overline{\theta}_j - \underline{\theta}_j)\frac{\vartheta_j}{\sqrt{1+\vartheta_j^2}}$, which we found to make optimizers very unstable.

statistical efficiency of the estimator $\hat{\theta}_n$, Hansen and Lee (2021) showed it can alter the pseudo-true value of the parameter under misspecification. Here, the rank condition (3') may or may not hold, depending on W_n. The condition number κ_W also affects local convergence. This highlights another role for the weighting matrix: it may facilitate or hinder the estimation itself. Two empirical applications illustrate the performance of the preferred Gauss-Newton algorithm.

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.

Arnoud, A., F. Guvenen, and T. Kleineberg (2019): "Benchmarking Global Optimizers," *NBER Working Paper*.

Bélisle, C. J. (1992): "Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$," *Journal of Applied Probability*, 29, 885–895.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841.

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

Brunner, D., F. Heiss, A. Romahn, and C. Weiser (2017): *Reliable estimation of random coefficient logit demand models*, 267, DICE Discussion Paper.

Cameron, A. C. and P. K. Trivedi (2005): *Microeconometrics: methods and applications*, Cambridge University Press.

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

Colacito, R., M. Croce, S. Ho, and P. Howard (2018): "BKK the EZ way: International long-run growth news and capital flows," *American Economic Review*, 108, 3416–49.

Conlon, C. and J. Gortmaker (2020): "Best practices for differentiated products demand estimation with pyblp," *The RAND Journal of Economics*, 51, 1108–1161.

Davidson, R., J. G. MacKinnon, et al. (2004): *Econometric theory and methods*, vol. 5, Oxford University Press New York.

Dennis, J. E. and R. B. Schnabel (1996): *Numerical methods for unconstrained optimization and nonlinear equations*, SIAM.

Deuflhard, P. (2005): *Newton methods for nonlinear problems: affine invariance and*

adaptive algorithms, vol. 35, Springer Science & Business Media.

Donaldson, D. (2018): "Railroads of the Raj: Estimating the impact of transportation infrastructure," *American Economic Review*, 108, 899–934.

Fang, K.-T. and Y. Wang (1993): *Number-theoretic methods in statistics*, vol. 51, CRC Press.

Forneron, J.-J. (2023): "Noisy, Non-Smooth, Non-Convex Estimation of Moment Condition Models," *arXiv preprint arXiv:2301.07196*.

Gourieroux, C. and A. Monfort (1996): *Simulation-based econometric methods*, Oxford university press.

Hansen, B. E. (2022a): *Econometrics*, Princeton University Press.

——— (2022b): *Probability and Statistics for Economists*, Princeton University Press.

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

Hayashi, F. (2011): *Econometrics*, Princeton University Press.

Jennrich, R. I. (1969): "Asymptotic properties of non-linear least squares estimators," *The Annals of Mathematical Statistics*, 40, 633–643.

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

Lagarias, J. C., J. A. Reeds, M. H. Wright, and P. E. Wright (1998): "Convergence properties of the Nelder–Mead simplex method in low dimensions," *SIAM Journal on optimization*, 9, 112–147.

Lemieux, C. (2009): *Monte Carlo and Quasi-Monte Carlo Sampling*, Springer Series in Statistics, Springer New York.

Lise, J. and J.-M. Robin (2017): "The Macrodynamics of Sorting between Workers and Firms," *American Economic Review*, 107, 1104–35.

McCullough, B. D. and H. D. Vinod (2003): "Verifying the solution from a nonlinear solver: A case study," *American Economic Review*, 93, 873–892.

McKinnon, K. I. (1998): "Convergence of the Nelder–Mead Simplex method to a nonstationary Point," *SIAM Journal on optimization*, 9, 148–158.

Moran, P. and A. Queralto (2018): "Innovation, productivity, and monetary policy," *Journal of Monetary Economics*, 93, 24–41.

Nash, J. C. (1990): *Compact numerical methods for computers: linear algebra and function minimisation*, Routledge.

Nelder, J. A. and R. Mead (1965): "A simplex method for function minimization," *The*

computer journal, 7, 308–313.

Nevo, A. (2001): "Measuring market power in the ready-to-eat cereal industry," *Econometrica*, 69, 307–342.

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.

Powell, M. J. (1973): "On search directions for minimization algorithms," *Mathematical programming*, 4, 193–201.

Quandt, R. E. (1983): "Computational problems and methods," *Handbook of econometrics*, 1, 699–764.

Salanié, B. and F. A. Wolak (2022): "Fast, Detail-free, and Approximately Correct: Estimating Mixed Demand Systems," .

Spall, J. C. (2005): *Introduction to stochastic search and optimization: estimation, simulation, and control*, John Wiley & Sons.

Stokes, H. H. (2004): "On the advantage of using two or more econometric software systems to solve the same problem," *Journal of Economic and Social Measurement*, 29, 307–320.

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

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*, MIT press.

Appendix A Proofs for the Main Results

A.1 Primitive Conditions for Assumption 1

In the following we will use the notation: $\overline{g}_n(\theta) = 1/n \sum_{i=1}^n g(\theta; x_i)$, $g(\theta) = \mathbb{E}[\overline{g}_n(\theta)]$, $G(\theta; x_i) = \partial_\theta g(\theta; x_i)$, $G_n(\theta) = 1/n \sum_{i=1}^n G(\theta; x_i)$, $G(\theta) = \mathbb{E}[G_n(\theta)]$, $Q_n(\theta) = \overline{g}_n(\theta)' W_n \overline{g}_n(\theta)$, and $Q(\theta) = g(\theta)' W g(\theta)$. W_n and W are symmetric. With probability approaching 1 will be abbreviated as wpa1. $\mathcal{B}_R(\theta^\dagger)$ is a closed ball of radius R, centered around $\theta^\dagger$.

Assumption A1. *Suppose the observations x_i are iid and, for some $\delta \in (0,1)$: i. $Q(\theta) = \|g(\theta)\|_W^2$ has a unique minimum $\theta^\dagger \in$ interior(Θ) such that $Q(\theta^\dagger) = 0$, $\mathbb{E}[\|g(\theta; x_i)\|^2] < \infty$, for all $\theta \in \Theta$, ii. $g(\theta; x_i)$ and $g(\theta)$ are twice continuously differentiable on Θ, iii. $\mathbb{E}[\|G(\theta; x_i)\|^2] < \infty$, for all $\theta \in \Theta$, there exist a $\bar{L}(\cdot) \geq 0$ such that for any $(\theta_1, \theta_2) \in \Theta^2$, $\|G(\theta_1; x_i) - G(\theta_2; x_i)\| \leq \bar{L}(x_i)\|\theta_1 - \theta_2\|$, where $\mathbb{E}[|\bar{L}(x_i)|^2] < \infty$, and $\mathbb{E}[\bar{L}(x_i)] \leq (1-\delta)L$, for some $R_G > 0$ such that $\mathcal{B}_{(1+\delta)R_G}(\theta^\dagger) \subseteq \Theta$, $\sigma_{\min}[G(\theta)] \geq (1+\delta)\underline{\sigma} > 0$ for all $\|\theta - \theta^\dagger\| \leq (1+\delta)R_G$, iv. The parameters space Θ is convex and compact, and v. $W_n \xrightarrow{p} W$, $0 < (1+\delta)\underline{\lambda}_W \leq \lambda_{\min}(W) \leq \lambda_{\max}(W)(1-\delta)\overline{\lambda}_W < \infty$.*

Remarks. The condition $Q(\theta^\dagger) = 0$ is only used to prove that the arg-minimizer $\hat{\theta}_n$ is unique wpa1. It could be relaxed to allow for moderate misspecification, at the cost of longer derivations and several bounds on the amount of misspecification φ, as in Theorem 3. The condition that x_i are iid can also be weakened to allow for non-identically distributed dependent observations by appropriately adjusting the moment conditions in A1i, iii which are used to derive uniform laws of large numbers for $\overline{g}_n$ and G_n.

Lemma A1. *Assumption A1 implies Assumption 1.*

Proof of Lemma A1. Assumption 1ii, iv follow from A1ii, iv. Use Weyl's perturbation inequality for singular values (Bhatia, 2013, Problem III.6.5) to find $\lambda_{\min}(W_n) \geq \lambda_{\min}(W) - \sigma_{\max}(W_n - W) \geq (1+\delta)\underline{\lambda}_W - o_p(1) \geq \underline{\lambda}_W$, wpa 1. Likewise, $\lambda_{\max}(W_n) \leq \overline{\lambda}_W$, wpa1. This yields Assumption 1v.

Assumption A1iii and compactness imply uniform convergence of the sample Jacobian $\sup_{\theta\in\Theta} \|G_n(\theta) - G(\theta)\| = o_p(1)$, see Jennrich (1969). We also have uniform convergence for the same moments. Condition ii implies $\overline{g}_n(\theta) - g(\theta) = o_p(1)$, for all θ. Notice that $\|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\| = \|[G_n(\tilde{\theta}) - G(\tilde{\theta})](\theta_1 - \theta_2)\| \leq [\sup_{\theta\in\Theta} \|G_n(\theta) - G(\theta)\|]\|\theta_1 - \theta_2\|$, where the sup is a $o_p(1)$ by uniform convergence of G_n. Using a finite cover and arguments similar to Jennrich (1969), this implies uniform convergence: $\sup_{\theta\in\Theta} \|\overline{g}_n(\theta) - g(\theta)\| = o_p(1)$.

Then, uniform convergence of $\overline{g}_n$ and $W_n \xrightarrow{p} W$ imply uniform converge of Q_n to Q. Continuity and the global identification condition A1i. imply $\hat{\theta}_n \xrightarrow{p} \theta^\dagger$ (Newey and McFadden, 1994, Th2.1). This implies that $\|\theta - \hat{\theta}_n\| \leq R_G \Rightarrow \|\theta - \theta^\dagger\| \leq R_G + o_p(1) \leq (1+\delta)R_G$, wpa 1, i.e. $\mathcal{B}_{R_G}(\hat{\theta}_n) \subseteq \mathcal{B}_{(1+\delta)R_G}(\theta^\dagger) \subseteq \Theta$. This implies $\hat{\theta}_n \in \text{interior}(\Theta)$, wpa1. Then, for the same θ, $\sigma_{\min}[G(\theta)] \geq (1+\delta)\underline{\sigma}$, wpa1. Apply Weyl's inequality for singular values to find that, uniformly in θ: $\sigma_{\min}[G_n(\theta)] \geq \sigma_{\min}[G_n(\theta)] - \sigma_{\max}[G(\theta) - G_n(\theta)] \geq (1+\delta)\underline{\sigma} - o_p(1) \geq \underline{\sigma} > 0$, wpa 1. Take any two θ_1, θ_2 in Θ, $\|G_n(\theta_1) - G_n(\theta_2)\| \leq 1/n \sum_{i=1}^n \bar{L}(x_i)\|\theta_1 - \theta_2\| \leq [(1+\delta)L + o_p(1)]\|\theta_1 - \theta_2\| \leq L\|\theta_1 - \theta_2\|$, wpa1, using a law of large numbers for $\bar{L}(x_i)$. This yields all the conditions in Assumption 1iii.

Remains to show that the sample arg-minimizer $\hat{\theta}_n$ is unique, wpa1, so that Assumption 1i also holds. Having shown that Assumption 1ii-v holds, we can use it in the following. Using the same steps as in the beginning of the proof of Theorem 2:

$$(\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n})\|\theta - \hat{\theta}_n\|^2 \leq 2[Q_n(\theta) - Q_n(\hat{\theta}_n)],$$

wpa1, but here only locally, i.e. uniformly in $\|\theta - \hat{\theta}_n\| \leq R_G$. Any approximate minimizer is in this neighborhood wpa1, by consistency, so we only need to check the unicity of an exact minimizer within this neighborhood.

We have $\underline{\lambda} = \min_{\|\theta - \hat{\theta}_n\| \leq R_G} \lambda_{\min}(G_n(\theta)' W_n G_n(\theta)) \geq \underline{\sigma}^2 \underline{\lambda}_W$, wpa1, and $C := 2\sqrt{\lambda_{\max}(W_n)}L$, bounded wpa1 as well. Now $Q_n(\hat{\theta}_n) = o_p(1)$ by uniform convergence and A1i, $Q(\theta^\dagger) = 0$, likewise $(\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}) \geq 1/2\underline{\lambda} > 0$, wpa1. Take $\hat{\theta}_n$ to be an exact minimizer, and any $\theta \neq \hat{\theta}_n$ such that $\|\theta - \hat{\theta}_n\| \leq R_G$, then $Q_n(\theta) \geq Q_n(\hat{\theta}_n) + 1/4\underline{\lambda}\|\theta - \hat{\theta}_n\|^2 > Q_n(\hat{\theta}_n)$, wpa1. This implies that the arg-minimizer $\hat{\theta}_n$ is unique wpa1 and concludes the proof. □

A.2 Proofs for Section 3.1

Proof of Proposition 1 (Gauss-Newton). Take $\theta_k \in \Theta$, the update (1) can be re-written as:

$$\begin{aligned} \theta_{k+1} - \hat{\theta}_n = &\Big(I_d - \gamma P_k G_n(\theta_k)' W_n G_n(\theta_k)\Big)(\theta_k - \hat{\theta}_n) \\ &- \gamma P_k G_n(\theta_k)' W_n [\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)]. \end{aligned} \tag{A.1}$$

For GN, $P_k G_n(\theta_k)' W_n G_n(\theta_k) = I_d$ so that we have:

$$\begin{aligned} \theta_{k+1} - \hat{\theta}_n = &(1-\gamma)(\theta_k - \hat{\theta}_n) \\ &- \gamma P_k G_n(\theta_k)' W_n [\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)] \\ &- \gamma P_k [G_n(\theta_k) - G_n(\hat{\theta}_n)]' W_n \overline{g}_n(\hat{\theta}_n), \end{aligned} \tag{A.1'}$$

using the first-order condition $G_n(\hat{\theta}_n)'W_n\overline{g}_n(\hat{\theta}_n) = 0$. From Assumption 1, there exists $R_G > 0$ such that: $\underline{\sigma} \leq \sigma_{\min}[G_n(\theta_k)]$ for any $\|\theta_k - \hat{\theta}_n\| \leq R_G$, which implies that P_k is well defined and bounded. Since G_n is Lipschitz continuous with constant $L \geq 0$:

$$\|P_k G_n(\theta_k)'W_n[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)]\| \leq \underline{\sigma}^{-1}\sqrt{\overline{\lambda}_W/\underline{\lambda}_W}L\|\theta_k - \hat{\theta}_n\|^2,$$

We also have:

$$\|P_k[G_n(\theta_k) - G_n(\hat{\theta}_n)]'W_n\overline{g}_n(\hat{\theta}_n)\| \leq \underline{\sigma}^{-2}(\sqrt{\overline{\lambda}_W}/\underline{\lambda}_W)L\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\|\theta_k - \hat{\theta}_n\|.$$

Combine these two inequalities into (A.1') to find:

$$\begin{aligned}&\|\theta_{k+1} - \hat{\theta}_n\| \\ &\leq \left(1 - \gamma + \gamma\left[\underline{\sigma}^{-1}\sqrt{\overline{\lambda}_W/\underline{\lambda}_W}L\|\theta_k - \hat{\theta}_n\| + \underline{\sigma}^{-2}(\sqrt{\overline{\lambda}_W}/\underline{\lambda}_W)L\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right]\right)\|\theta_k - \hat{\theta}_n\|.\end{aligned} \tag{A.1"}$$

Now take any $\tilde{\gamma} \in (0, \gamma)$, let:

$$\tilde{R}_n = \frac{\gamma - \tilde{\gamma}}{\gamma}\left[L^{-1}\underline{\sigma}\sqrt{\underline{\lambda}_W/\overline{\lambda}_W}\right] - (\underline{\sigma}^{-1}/\sqrt{\underline{\lambda}_W})\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}.$$

Let $R_n = \min(\tilde{R}_n, R_G)$, for any $\|\theta_k - \hat{\theta}_n\| \leq R_n$, we have $\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq R_n$. By recursion, we then have for any $\|\theta_0 - \hat{\theta}_n\| \leq R_n$:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1 - \tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1 - \tilde{\gamma})^{k+1}\|\theta_0 - \hat{\theta}_n\|,$$

as stated in (2). □

Proof of Proposition 1 (General Case). Take $\theta_k \in \Theta$, the update (1) can be re-written as:

$$\begin{aligned}\theta_{k+1} - \hat{\theta}_n =& \Big(I_d - \gamma P_k G_n(\theta_k)'W_n G_n(\theta_k)\Big)(\theta_k - \hat{\theta}_n) \\ &- \gamma P_k G_n(\theta_k)'W_n[\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)].\end{aligned} \tag{A.1}$$

Taking norms on both sides this identity yields:

$$\begin{aligned}\|\theta_{b+1} - \hat{\theta}_n\| \leq& \sigma_{\max}\Big[I_d - \gamma P_k G_n(\theta_k)'W_n G_n(\theta_k)\Big]\|\theta_b - \hat{\theta}_n\| \\ &+ \gamma\|P_k G_n(\theta_k)'W_n[\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)]\|,\end{aligned} \tag{A.1'}$$

where $\sigma_{\max}$ returns the largest singular value. We will now bound each of these two terms. First, note that $\sigma_{\max}[I_d - \gamma P_k G_n(\theta_k)' W_n G_n(\theta_k)] = \sigma_{\max}[I_d - \gamma P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] = \max_{j=1,\dots,d} |\lambda_j[I_d - \gamma P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}]|$, where λ_j are the eigenvalues. Because this is a difference of Hermitian matrices, Weyl's perturbation inequality (Bhatia, 2013, Corollary III.2.2) implies the following bounds:

$$\begin{aligned} 1 - \gamma\lambda_{\max}[P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] &\leq \lambda_{\min}[I_d - \gamma P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] \\ &\leq \lambda_{\max}[I_d - \gamma P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] \\ &\leq 1 - \gamma\lambda_{\min}[P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}]. \end{aligned}$$

Let $\overline{\sigma} = \max_{\theta\in\Theta} \sigma_{\max}[G_n(\theta)]$, suppose $0 < \gamma < [\overline{\lambda}_P \overline{\lambda}_W \overline{\sigma}^2]^{-1}$, we then have:

$$0 \leq 1 - \gamma\lambda_{\max}[P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] \leq 1 - \gamma\lambda_{\min}[P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}],$$

so that we are only concerned with the upper bound. From Assumption 1, $\|\theta - \hat{\theta}_n\| \leq R_G \Rightarrow \sigma_{\min}[G_n(\theta)] \geq \underline{\sigma}$. Combine with the bound for γ to find:

$$0 \leq \sigma_{\max}[I_d - \gamma P_k G_n(\theta_k)' W_n G_n(\theta_k)] \leq 1 - \gamma\underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2 < 1,$$

for any choice of $\gamma \in (0, [\overline{\lambda}_P \overline{\lambda}_W \overline{\sigma}^2]^{-1})$. For the second term in (A.1), using the identity $G_n(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) = 0$ and the mean value Theorem, we have for some intermediate value $\tilde{\theta}_k$:

$$\begin{aligned} P_k G_n(\theta_k)' W_n[\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)] =& P_k G_n(\theta_k)' W_n[G_n(\tilde{\theta}_k) - G_n(\theta_k)](\theta_k - \hat{\theta}_n) \\ &+ P_k[G_n(\theta_k) - G_n(\hat{\theta}_n)]' W_n \overline{g}_n(\hat{\theta}_n). \end{aligned}$$

Since G_n is Lipschitz continuous with constant $L \geq 0$:

$$\begin{aligned} \|(A.1')\| &\leq (1 - \gamma\underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2)\|\theta_b - \hat{\theta}_n\| + \gamma\overline{\lambda}_P \overline{\lambda}_W \overline{\sigma} L \|\theta_b - \hat{\theta}_n\|^2 + \gamma\overline{\lambda}_P \overline{\lambda}_W^{1/2} L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \|\theta_b - \hat{\theta}_n\| \\ &= \left(1 - \gamma\underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2 + \gamma\left[\overline{\lambda}_P \overline{\lambda}_W \overline{\sigma} L \|\theta_b - \hat{\theta}_n\| + \overline{\lambda}_P \overline{\lambda}_W^{1/2} L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right]\right) \|\theta_b - \hat{\theta}_n\|. \end{aligned}$$

Let $c_1 = \overline{\lambda}_P \overline{\lambda}_W \overline{\sigma} L$, $c_2 = \overline{\lambda}_P \overline{\lambda}_W^{1/2} L$, pick $\tilde{\gamma} \in (0, \gamma\underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2)$, and assume:

$$\|\theta_k - \hat{\theta}_n\| \leq \frac{\gamma\underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2 - \tilde{\gamma}}{\gamma c_1} - \frac{c_2}{c_1} \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} := \tilde{R}_n. \tag{A.2}$$

Take $R_n = \min(R_G, \tilde{R}_n)$, $\|\theta_k - \hat{\theta}_n\| \leq R_n$ implies that, by construction:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1 - \tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1 - \overline{\gamma})^{k+1}\|\theta_0 - \hat{\theta}_n\|,$$

by recursion, if $\|\theta_0 - \hat{\theta}_n\| \leq R_n$. □

Proof of Theorem 1 In the just-identified case, we will repeatedly use the identities $\overline{g}_n(\hat{\theta}_n) = 0$ and $Q_n(\hat{\theta}_n) = \frac{1}{2}\overline{g}_n(\hat{\theta}_n)'W_n\overline{g}_n(\hat{\theta}_n) = 0$. Take any $\theta \in \Theta$, by the Mean Value Theorem there is some $\tilde{\theta}_n$ between θ and $\hat{\theta}_n$ such that $\overline{g}_n(\theta) = G_n(\tilde{\theta}_n)(\theta - \hat{\theta}_n)$ which implies:

$$\left(\frac{1}{2}\min_{\theta\in\theta}\lambda_{\min}[G_n(\theta)'W_nG_n(\theta)]\right)\|\theta - \hat{\theta}_n\|^2 \leq Q_n(\theta) \leq \left(\frac{1}{2}\max_{\theta\in\theta}\lambda_{\max}[G_n(\theta)'W_nG_n(\theta)]\right)\|\theta - \hat{\theta}_n\|^2,$$

and, using the full rank assumption, let $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$, denote respectively the min and the max. This yields an equivalence between the two distances $\|\theta - \hat{\theta}_n\|$ and $Q_n(\theta)$. Apply the Mean Value Theorem to Q_n, for some $\tilde{\theta}_k$ between θ_k and θ_{k+1} in (1):

$$Q_n(\theta_{k+1}) = Q_n(\theta_k) + \partial_\theta Q_n(\theta_k)'(\theta_{k+1} - \theta_k) + [\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)]'(\theta_{k+1} - \theta_k), \qquad \text{(A.3)}$$

where $\partial_\theta Q_n(\theta_k) = G_n(\theta_k)'W_n\overline{g}_n(\theta_k)$ and $\theta_{k+1} - \theta_k = -\gamma P_k G_n(\theta_k)'W_n\overline{g}_n(\theta_k)$. This yields a first inequality:

$$\begin{aligned}\partial_\theta Q_n(\theta_k)'(\theta_{k+1} - \theta_k) &= -\gamma\overline{g}_n(\theta_k)'W_n^{1/2}\left(W_n^{1/2}G_n(\theta_k)P_kG_n(\theta_k)'W_n^{1/2}\right)W_n^{1/2}\overline{g}_n(\theta_k) \\ &\leq -\gamma c_1\|\overline{g}_n(\theta_k)\|_{W_n}^2,\end{aligned}$$

where $c_1 := \min_{\theta\in\Theta}\lambda_{\min}[W_n^{1/2}G_n(\theta)P_kG_n(\theta)'W_n^{1/2}] > 0$, using the full rank assumption, and $\|\overline{g}_n(\theta_k)\|_{W_n}^2 = Q_n(\theta_k)$. For the second inequality, since $\overline{g}_n$ is twice continuously differentiable and Θ is compact, $\partial_\theta Q_n$ is Lipschitz continuous with constant $L_Q \geq 0$. This implies:

$$\begin{aligned}\|[\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)]'(\theta_{k+1} - \theta_k)\| \leq L_Q\|\tilde{\theta}_k - \theta_k\| \times \|\theta_{k+1} - \theta_k\| &\leq L_Q\|\theta_{k+1} - \theta_k\|^2 \\ &\leq \gamma^2 c_2\|\overline{g}_n(\theta_k)\|_{W_n}^2,\end{aligned}$$

since $\|\theta_{k+1} - \theta_k\| = \gamma\|P_kG_n(\theta_k)'W_n\overline{g}_n(\theta_k)\|$ and $\|\tilde{\theta}_k - \theta_k\| \leq \|\theta_{k+1} - \theta_k\|$, setting $c_2 := L_Q\max_{\theta\in\Theta}\sigma_{\max}^2[P_kG_n(\theta)'W_n^{1/2}] < +\infty$. Combine the two inequalities into (A.3) to find:

$$Q_n(\theta_{k+1}) \leq (1 - \gamma c_1 + \gamma^2 c_2)Q_n(\theta_k),$$

for any $\theta_k \in \Theta$. The polynomial $P(\gamma) = 1 - \gamma c_1 + \gamma^2 c_2$ is such that $P(0) = 1, d_\gamma P(0) < 0$ which implies $P(\gamma) \in (0,1)$ strictly for any $\gamma > 0$ sufficiently small. Take any such γ and let $(1-\overline{\gamma})^2 = P(\gamma) \in (0,1)$ for some $\overline{\gamma} \in (0,1)$, by construction. Take any $\theta_0 \in \Theta$, by recursion:

$$Q_n(\theta_{k+1}) \leq (1-\overline{\gamma})^2 Q_n(\theta_k) \leq \cdots \leq (1-\overline{\gamma})^{2(k+1)} Q_n(\theta_0).$$

Now apply the distance equivalence derived earlier to get the desired result:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})^{k+1} \sqrt{\overline{\lambda}/\underline{\lambda}} \|\theta_0 - \hat{\theta}_n\|.$$

□

Proof of Theorem 2. The general layout of the proof is similar to the just-identified case. Differences arise because $Q_n(\hat{\theta}_n) \neq 0$ in general and $G_n(\theta)$ only has rank d_θ which is less than the dimension of $\overline{g}_n$ so that several parts of the proof do not apply anymore. First:

$$Q_n(\theta) - Q_n(\hat{\theta}_n) = \frac{1}{2}\left[\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\right]' W_n \left[\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\right] + \left[\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\right]' W_n \overline{g}_n(\hat{\theta}_n).$$

The leading term equals $\frac{1}{2}(\theta - \hat{\theta}_n)' G_n(\tilde{\theta}_n)' W_n G_n(\tilde{\theta}_n)(\theta - \hat{\theta}_n)$ which can be bounded above and below using the same approach as before. For the last term, use the first-order condition $G_n(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) = 0$ to get, using $\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n) = [G(\tilde{\theta}_n) - G_n(\hat{\theta}_n) + G_n(\hat{\theta}_n)](\theta - \hat{\theta}_n)$:

$$\begin{aligned}\|[\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)]' W_n \overline{g}_n(\hat{\theta}_n)\| &\leq \sqrt{\lambda_{\max}(W_n)} \|\theta - \hat{\theta}_n\| \times \|G_n(\hat{\theta}_n) - G_n(\tilde{\theta}_n)\| \times \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \\ &\leq \sqrt{\lambda_{\max}(W_n)} L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \|\theta - \hat{\theta}_n\|^2,\end{aligned}$$

where $L \geq 0$ is the Lipschitz constant of G_n. Let $0 < \underline{\lambda} = \min_{\theta \in \Theta} \lambda_{\min}[G_n(\theta)' W_n G_n(\theta)] \leq \overline{\lambda} = \max_{\theta \in \Theta} \lambda_{\max}[G_n(\theta)' W_n G_n(\theta)] < \infty$, apply the triangular inequality and its reverse to find the relation:

$$\left(\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right) \|\theta - \hat{\theta}_n\|^2 \leq 2[Q_n(\theta) - Q_n(\hat{\theta}_n)] \leq \left(\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right) \|\theta - \hat{\theta}_n\|^2, \tag{A.4}$$

where $C := 2\sqrt{\lambda_{\max}(W_n)} L \geq 0$ is finite. As in the just-identified case, we can write:

$$Q_n(\theta_{k+1}) = Q_n(\theta_k) + \partial_\theta Q_n(\theta_k)'(\theta_{k+1} - \theta_k) + [\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)]'(\theta_{k+1} - \theta_k),$$

and bound each of the last two terms. As before, we have:

$$\partial_\theta Q_n(\theta_k)'(\theta_{k+1}-\theta_k) = -\gamma \overline{g}_n(\theta_k)' W_n G_n(\theta_k) P_k G_n(\theta_k)' W_n \overline{g}_n(\theta_k),$$

however $\dim(W_n^{1/2}\overline{g}_n(\theta_k)) > \text{rank}[W_n^{1/2} G_n(\theta_k) P_k G_n(\theta_k)' W_n^{1/2}]$, the model being over-identified. Hence, we only have $\partial_\theta Q_n(\theta_k)'(\theta_{k+1}-\theta_k) \leq 0$ which, unlike the just-identified case, does not imply a strict contraction. Nevertheless, we have $\overline{g}_n(\theta_k) = G_n(\tilde{\theta}_k)(\theta_k - \hat{\theta}_n) + \overline{g}_n(\hat{\theta}_n)$ where $\dim(\theta_k - \hat{\theta}_n)$ equals the rank of the matrix above. Let $A_k = W_n^{1/2} G_n(\theta_k) P_k G_n(\theta_k)' W_n^{1/2}$:

$$\begin{aligned}
&-\gamma \overline{g}_n(\theta_k)' W_n G_n(\theta_k) P_k G_n(\theta_k)' W_n \overline{g}_n(\theta_k) \\
&\qquad = -\gamma \left[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n)\right]' W_n^{1/2} A_k W_n^{1/2} \left[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n)\right] \qquad \text{(A.5)} \\
&\qquad\quad - \gamma \overline{g}_n(\hat{\theta}_n)' W_n^{1/2} A_k W_n^{1/2} \overline{g}_n(\hat{\theta}_n) \qquad \text{(A.6)} \\
&\qquad\quad - 2\gamma \overline{g}_n(\hat{\theta}_n)' W_n^{1/2} A_k W_n^{1/2} \left[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n)\right]. \qquad \text{(A.7)}
\end{aligned}$$

Now, bound these terms one at a time:

$$\begin{aligned}
(A.5) &= -\gamma(\theta_k - \hat{\theta}_n)' G_n(\tilde{\theta}_k)' W_n^{1/2} A_k W_n^{1/2} G_n(\tilde{\theta}_k)(\theta_k - \hat{\theta}_n) \\
&\leq -\gamma \|\theta_k - \hat{\theta}_n\|^2 \min_{\theta \in \Theta} \lambda_{\min}[G_n(\theta)' W_n^{1/2} A_k W_n^{1/2} G_n(\theta)] \\
&\leq -\gamma c_{1n}[Q_n(\theta_k) - Q_n(\hat{\theta}_n)],
\end{aligned}$$

where the last inequality comes from (A.4) above, and

$$c_{1n} := \min_{\theta \in \Theta} \lambda_{\min}[G_n(\theta)' W_n^{1/2} A_k W_n^{1/2} G_n(\theta)](\overline{\lambda}/2 + C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n})^{-1} > 0,$$

is bounded below by a strictly positive value with probability approaching 1, using Assumption 2 and condition (3').[19] To see why condition (3') is critical, notice that $G_n(\theta)' W_n^{1/2} A_k W_n^{1/2} G_n(\theta) = G_n(\theta)' W_n G_n(\theta_k) P_k G_n(\theta_k)' W_n G_n(\theta)$ is symmetric and has full rank if both $G_n(\theta)' W_n G_n(\theta_k)$ and P_k have full rank. Both condition (3') and Assumption 2 need to hold for c_{1n} to be non-zero. Then $(A.6) \leq -\gamma Q_n(\hat{\theta}_n)\lambda_{\min}(A_k) \leq 0$. For the remaining term, apply the Cauchy-Schwarz inequality, the Mean Value Theorem, and (A.4) to find the last bound:

$$\|(A.7)\| \leq 2\gamma \sqrt{Q_n(\hat{\theta}_n)} \frac{\max_{\theta \in \Theta} \sigma_{\max}[A_k W_k^{1/2} G_n(\theta)]}{[\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{1/2}} [Q_n(\theta_k) - Q_n(\hat{\theta}_n)]^{1/2}$$

[19] An explicit lower bound is given in the proof of Theorem 3.

Let $c_{3n} := 2\max_{\theta\in\Theta}\sigma_{\max}[A_k W_k^{1/2} G_n(\theta)][\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1/2}$. As in the just-identified case, $\partial_\theta Q_n$ is Lipschitz continuous with constant $L_Q \geq 0$, which yields the same inequality as in the proof of Theorem 1:

$$\|[\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)]'(\theta_{k+1} - \theta_k)\| \leq \gamma^2 L_Q \max_{\theta\in\Theta} \sigma^2_{\max}[P_k G_n(\theta)' W_n^{1/2}] Q_n(\theta_k).$$

Let $c_2 := L_Q \max_{\theta\in\Theta}\sigma^2_{\max}[P_k G_n(\theta)' W_n^{1/2}] \geq 0$. Combine all the inequalities above to get:

$$\begin{aligned} Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq{}& (1 - \gamma c_{1n} + \gamma^2 c_2)[Q_n(\theta_k) - Q_n(\hat{\theta}_n)] \\ &+ \gamma^2 c_2 Q_n(\hat{\theta}_n) + \gamma c_{3n}\sqrt{Q_n(\hat{\theta}_n)}\left[Q_n(\theta_k) - Q_n(\hat{\theta}_n)\right]^{1/2}. \end{aligned}$$

Because of the square root on $Q_n(\theta_k) - Q_n(\hat{\theta}_n)$, this is a non-linear recursion. To derive explicit convergence results, we will bound it by a linear recursion using:

i. If $[Q_n(\theta_k) - Q_n(\hat{\theta}_n)]^{1/2} \geq 2c_{3n}/c_{1n}\sqrt{Q_n(\hat{\theta}_n)}$, then:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq (1 - \gamma\frac{c_{1n}}{2} + \gamma^2 c_2)[Q_n(\theta_k) - Q_n(\hat{\theta}_n)] + \gamma^2 c_2 Q_n(\hat{\theta}_n).$$

ii. Otherwise:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq (1 - \gamma c_{1n} + \gamma^2 c_2)[Q_n(\theta_k) - Q_n(\hat{\theta}_n)] + \left(\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}\right) Q_n(\hat{\theta}_n).$$

A majorization of these two bounds implies:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq \left(1 - \gamma\frac{c_{1n}}{2} + \gamma^2 c_2\right)\left[Q_n(\theta_k) - Q_n(\hat{\theta}_n)\right] + \left(\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}\right) Q_n(\hat{\theta}_n). \tag{A.8}$$

Let $P_n(\gamma) = (1 - \gamma\frac{c_{1n}}{2} + \gamma^2 c_2)$. Then, using the same arguments as in the just-identified case, for $\gamma > 0$ sufficiently small, we have $P_n(\gamma) = (1 - \overline{\gamma})^2 \in (0, 1)$, i.e. (1) is a strict contraction globally. Iterate on the recursion to find:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq (1 - \overline{\gamma})^{2(k+1)}[Q_n(\theta_0) - Q_n(\hat{\theta}_n)] + \frac{\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}}{1 - (1 - \overline{\gamma})^2} Q_n(\hat{\theta}_n).$$

Apply the distance equivalence to find:

$$\|\theta_{k+1} - \hat{\theta}_n\|^2 \leq (1-\overline{\gamma})^{2(k+1)} \left(\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)^{-1} [Q_n(\theta_0) - Q_n(\hat{\theta}_n)]$$
$$+ \left(\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)^{-1} \frac{\gamma^2 c_2 + 2\gamma \frac{c_{3n}^2}{c_{1n}}}{1-(1-\overline{\gamma})^2} Q_n(\hat{\theta}_n).$$

For the choice of $\gamma > 0$ which yields the result, there exists a $R_n > 0$ for which Proposition 1 holds. Then in large samples we have:

$$\left(\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)^{-1} \frac{\gamma^2 c_2 + 2\gamma \frac{c_{3n}^2}{c_{1n}}}{1-(1-\overline{\gamma})^2} Q_n(\hat{\theta}_n) \leq R_n^2/2,$$

with increasing probability. For k large enough:

$$(1-\overline{\gamma})^{2(k)} \left(\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)^{-1} [Q_n(\theta_0) - Q_n(\hat{\theta}_n)] \leq R_n^2/2,$$

as well.[20] Then, with increasing probability, for this choice of k we have:

$$\|\theta_k - \hat{\theta}_n\| \leq R_n,$$

apply Proposition 1 for another $j \geq 0$ iterations to find:

$$\|\theta_{k+j} - \hat{\theta}_n\| \leq (1-\tilde{\gamma})^j R_n,$$

where $\tilde{\gamma}$ is the convergence rate in Proposition 1 which need not the same as the global rate derived above. This concludes the proof. □

Proof of Proposition 2 Take θ such that $G_n(\theta)' W_n \overline{g}_n(\theta) = 0$, then for some intermediate value $\tilde{\theta}_n$:

$$G_n(\theta)' W_n G_n(\tilde{\theta}_n)(\theta - \hat{\theta}_n) = -G_n(\theta)' W_n \overline{g}_n(\hat{\theta}_n).$$

Take norms on both sides to find:

$$\|\theta - \hat{\theta}_n\| \leq \frac{\max_{\theta \in \Theta} \sigma_{\max}[G_n(\theta) W_n^{1/2}]}{\min_{\theta_1, \theta_2 \in \Theta} \sigma_{\min}[G_n(\theta_1)' W_n G_n(\theta_2)]} \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} := C_1 \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}.$$

[20] Let $d_{0n} = [\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1}[Q_n(\theta_0) - Q_n(\hat{\theta}_n)]$, pick $k \geq \frac{2\log R_n - \log 2 - \log d_{0n}}{2\log(1-\overline{\gamma})}$.

This implies that $\|\theta - \hat{\theta}_n\| \leq C_1 \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ which is not quite the desired result. Using $G_n(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) = 0$, we can further write:

$$G_n(\theta)' W_n G_n(\tilde{\theta}_n)(\theta - \hat{\theta}_n) = -[G_n(\theta) - G_n(\hat{\theta}_n)]' W_n \overline{g}_n(\hat{\theta}_n),$$

where now $\|[G_n(\theta) - G_n(\hat{\theta}_n)]' W_n^{1/2}\| \leq L\sqrt{\overline{\lambda}_W} C_1 \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$, using the Lipschitz continuity of G, the bound for $\|\theta - \hat{\theta}_n\|$ and the eigenvalue bound for W_n. Now we have:

$$\|\theta - \hat{\theta}_n\| \leq \frac{L\sqrt{\overline{\lambda}_W} C_1}{\min_{\theta_1, \theta_2 \in \Theta} \sigma_{\min}[G_n(\theta_1)' W_n G_n(\theta_2)]} \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 := C \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2.$$

which is the desired result. □

A.3 Proofs for Section 3.2

Proof of Proposition 3 (Gauss-Newton): Following the proof of Proposition 1:

$$\begin{aligned} &\|\theta_{k+1} - \hat{\theta}_n\| \\ &\leq \left(1 - \gamma + \gamma \left[\underline{\sigma}^{-1} \sqrt{\overline{\lambda}_W / \underline{\lambda}_W} L \|\theta_k - \hat{\theta}_n\| + \underline{\sigma}^{-2} (\sqrt{\overline{\lambda}_W} / \underline{\lambda}_W) L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right]\right) \|\theta_k - \hat{\theta}_n\|. \end{aligned} \tag{A.1"}$$

Take $\tilde{\gamma} \in (0, \gamma)$, and $\tilde{R}_n$ such that:

$$\tilde{R}_n = \frac{\gamma - \tilde{\gamma}}{\gamma} \left[L^{-1} \underline{\sigma} \sqrt{\underline{\lambda}_W / \overline{\lambda}_W}\right] - (\underline{\sigma}^{-1} / \sqrt{\underline{\lambda}_W}) \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}.$$

We have

$$\text{plim}_{n \to \infty} \tilde{R}_n = \tilde{R} = \frac{\gamma - \tilde{\gamma}}{\gamma} [L^{-1} \underline{\sigma} \sqrt{\underline{\lambda}_W / \overline{\lambda}_W}] - (\underline{\sigma}^{-1} / \sqrt{\underline{\lambda}_W}) \varphi > 0 \Leftrightarrow \varphi < [1 - \tilde{\gamma}/\gamma] \frac{\underline{\sigma}^2 \underline{\lambda}_W}{L\sqrt{\overline{\lambda}_W}}.$$

Under the stated Assumptions, for any $\varphi \geq 0$ such that $\varphi < \frac{\underline{\sigma}^2 \underline{\lambda}_W}{L\sqrt{\overline{\lambda}_W}}$, there exists $\tilde{\gamma} \in (0, \gamma)$, sufficiently small such that the above strict inequality holds. Then, $\tilde{R}_n \geq (1 - \varepsilon)\tilde{R} > 0$ with probability approaching 1 for any $\varepsilon \in (0, 1)$. Let $R_n = \min(\tilde{R}_n, R_G)$, take $\|\theta_0 - \hat{\theta}_n\| \leq R_n$, by recursion:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1 - \tilde{\gamma}) \|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1 - \tilde{\gamma})^{k+1} \|\theta_0 - \hat{\theta}_n\|,$$

with probability approaching 1, for all $k \geq 0$. This is the desired result. □

Proof of Theorem 3: The layout of the proof closely follows that of Theorem 2. Recall inequality (A.4):

$$\left(\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)\|\theta - \hat{\theta}_n\|^2 \leq 2[Q_n(\theta) - Q_n(\hat{\theta}_n)] \leq \left(\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)\|\theta - \hat{\theta}_n\|^2,$$

where $0 < \underline{\lambda} = \min_{\theta\in\Theta} \lambda_{\min}[G_n(\theta)'W_nG_n(\theta)] \leq \overline{\lambda} = \max_{\theta\in\Theta} \lambda_{\max}[G_n(\theta)'W_nG_n(\theta)] < \infty$, and $C := 2\sqrt{\lambda_{\max}(W_n)}L \geq 0$ are finite. Condition (6') implies $0 < \underline{\lambda} - C\varphi$ since $\underline{\lambda} \geq \underline{\sigma}^2\underline{\lambda}_W$ when (3) holds. Then, for any $\delta \in (0,1)$, we have $(\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}) \geq (1-\delta)[\underline{\lambda} - C\varphi] > 0$, with probability approaching 1 (wpa1). This implies that the norm equivalence holds and is informative, with high probability, in large samples. Now recall inequality (A.8):

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq \left(1 - \gamma\frac{c_{1n}}{2} + \gamma^2 c_2\right)\left[Q_n(\theta_k) - Q_n(\hat{\theta}_n)\right] + \left(\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}\right)Q_n(\hat{\theta}_n),$$

where:

$$\begin{aligned}
c_{1n} &= \min_{\theta\in\Theta} \lambda_{\min}[G_n(\theta)'W_n^{1/2}A_kW_n^{1/2}G_n(\theta)](\overline{\lambda}/2 + C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n})^{-1},\\
c_2 &= L_Q \max_{\theta\in\Theta} \sigma^2_{\max}[P_kG_n(\theta)'W_n^{1/2}],\\
c_{3n} &= 2\max_{\theta\in\Theta} \sigma_{\max}[A_kW_n^{1/2}G_n(\theta)][\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1/2},
\end{aligned}$$

L_Q is the Lipschitz constant of $\partial_\theta Q_n$ and $A_k = W_n^{1/2}G_n(\theta_k)P_kG_n(\theta_k)'W_n^{1/2}$ is an idempotent matrix for GN. Together, (3') and (6') imply the following upper and lower bounds holds wpa1:

$$0 < \underline{c}_1 := \frac{2}{3}\rho^2([\underline{\sigma}/\overline{\sigma}]^2\kappa_W^{-1})^2 \leq c_{1n} \leq 2[\overline{\sigma}/\underline{\sigma}]^2\kappa_W := \overline{c}_1 < \infty,$$

where $\overline{\sigma} \geq \underline{\sigma}$ is such that $\sigma_{\max}[G_n(\theta)] \leq \overline{\sigma}$ for all $\theta \in \Theta$ and $\kappa_W = \overline{\lambda}_W/\underline{\lambda}_W$. The upper bound relies on $\sigma_{\max}(A_k) = 1$ so the numerator is less than $\overline{\sigma}^2\overline{\lambda}_W$, while for the denominator $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \geq 0$ and $\overline{\lambda} \geq \underline{\sigma}^2\underline{\lambda}_W$. For the lower bound, condition (3') implies the numerator is greater than $\inf_{\theta\in\Theta} \sigma_{\min}[G_n(\theta)'W_nG_n(\theta_k)]^2\lambda_{\min}(P_k) \geq \rho^2[\underline{\sigma}^2\underline{\lambda}_W]^2[\overline{\sigma}^2\overline{\lambda}_W]^{-1}$. For the denominator of the lower bound, notice that $0 \leq \varphi < \underline{\lambda}/C$ implies $C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \leq 2C\varphi \leq 2\underline{\lambda} \leq 2\overline{\lambda}$, wpa 1, which – with a bound on $\overline{\lambda}$ – yields the resulting bound $\underline{c}_1$. Also, wpa1:

$$0 \leq c_{3n} \leq \overline{c}_3[\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1/2}.$$

where $\overline{c}_3 = 2\overline{\sigma}\overline{\lambda}_W^{1/2}$ since $\sigma_{\max}(A_k) = 1$ for GN. Combine these bounds to find that, wpa1 and uniformly in θ_k we have:[21]

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq \left(1 - \gamma\frac{\underline{c}_1}{2} + \gamma^2 c_2\right)\left[Q_n(\theta_k) - Q_n(\hat{\theta}_n)\right] + \left(\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}\right)Q_n(\hat{\theta}_n),$$

which does not depend on φ. For $\gamma \in (0,1)$ small enough, pick $\overline{\gamma} \in (0,1)$ such that $(1 - \gamma\frac{\underline{c}_1}{2} + \gamma^2 c_2) = (1-\overline{\gamma})^2$ and

$$C_n = \frac{\gamma^2 c_2 + 2\gamma c_{3n}^2/c_{1n}}{[1-(1-\overline{\gamma})^2][\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]} = O_p(1),$$

as in Theorem 2. Then we have:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq (1-\overline{\gamma})^{2k}\left[Q_n(\theta_0) - Q_n(\hat{\theta}_n)\right] + C_n[\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]Q_n(\hat{\theta}_n),$$

iterate on this inequality and apply the norm equivalence to find that (5') holds wpa1.

As in Theorem 2, we further need to invoke the local convergence results to show that $\theta_k \to \hat{\theta}_n$ as k increases. For that, we need to show that for some $\delta \in (0,1)$, sufficient small, we have $C_n Q_n(\hat{\theta}_n) \leq (1-\delta)R_n^2$, defined in Proposition 3. Note that condition (3') implies, without loss of generality, that $R_G > \tilde{R}_n = R_n$ in Proposition 3.

If $\varphi = 0$, we have $Q_n(\hat{\theta}_n) = o_p(1)$ and $C_n = O_p(1)$ which together yield $C_n Q_n(\hat{\theta}_n) = o_p(1) \leq (1-\delta)\tilde{R}_n^2$ wpa1, as in Theorem 2. Now suppose $\varphi > 0$, then we have, for any $\delta \in (0,1)$, that $Q_n(\hat{\theta}_n) \leq [1-\delta]^{-1}\varphi^2$ wpa1. We also have: $\{\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\}^{-1} \leq \{(1-\delta)\Delta\}^{-1}$ wpa1, where $\Delta = 1/2[\underline{\sigma}^2\underline{\lambda}_W - 2L\overline{\lambda}_W^{1/2}\varphi]$. Combine these with the bounds for c_{1n}, c_{3n} above to find that wpa1:

$$C_n Q_n(\hat{\theta}_n) \leq \frac{\gamma^2 c_2 + 2\gamma\overline{c}_3^2/[(1-\delta)\Delta\underline{c}_1]}{[\gamma\underline{c}_1 - \gamma^2 c_2](1-\delta)^2\Delta}\varphi^2 \leq \frac{\Delta\gamma^2 c_2 + 2\gamma\overline{c}_3^2/\underline{c}_1}{[\gamma\underline{c}_1 - \gamma^2 c_2](1-\delta)^3\Delta^2}\varphi^2,$$

using $(1-\overline{\gamma})^2 = (1 - \gamma\underline{c}_1/2 + \gamma^2 c_2)$. If inequality (7) holds strictly, then for $\delta \in (0,1)$ small enough we also have:

$$\frac{\Delta\gamma^2 c_2 + 2\gamma\overline{c}_3^2/\underline{c}_1}{[\gamma\underline{c}_1/2 - \gamma^2 c_2]\Delta^2}\varphi^2 \leq (1-\delta)^5\left((1-\varepsilon)\frac{\underline{\sigma}}{L\sqrt{\kappa_W}} - \frac{\varphi}{\underline{\sigma}\sqrt{\underline{\lambda}_W}}\right)^2. \tag{A.9}$$

Next, note that wpa1: $(1-\delta)^2[(1-\varepsilon)\frac{\underline{\sigma}}{L\sqrt{\kappa_W}} - \frac{\varphi}{\underline{\sigma}\sqrt{\underline{\lambda}_W}}]^2 \leq (1-\delta)[(1-\varepsilon)\frac{\underline{\sigma}}{L\sqrt{\kappa_W}} - \frac{\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}{\underline{\sigma}\sqrt{\underline{\lambda}_W}}]^2 =$

[21] The inequality is uniform in θ_k because the bound involves the same event on $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ for all θ_k.

$(1-\delta)\tilde{R}_n^2$ from Proposition 3. Set $\tilde{\gamma}$ such that $\varepsilon = \tilde{\gamma}/\gamma$ (or smaller). Putting these inequalities together implies that wpa1:

$$C_n \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 \leq (1-\delta)\tilde{R}_n^2,$$

for the same small enough $\delta \in (0,1)$. Now take $k \geq k_n$ given in the Theorem, wpa1 $\|\theta_k - \hat{\theta}_n\| \leq \tilde{R}_n$ when $k \geq k_n$ because k_n was chosen such that the leading term is less than $\delta\tilde{R}_n^2$ to be added to $(1-\delta)\tilde{R}_n^2$ above. Since the conditions for Proposition 3 hold, we have for $k = k_n + j$, $j \geq 0$: $\|\theta_k - \hat{\theta}_n\| \leq (1-\tilde{\gamma})^j \tilde{R}_n$, as desired. □

Supplement to

"Convexity Not Required: Estimation of Smooth Moment Condition Models"

Jean-Jacques Forneron* Liang Zhong†

April 19, 2023

This Supplemental Material consists of Appendices B, C, and D to the main text.

*Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA. Email: jjmf@bu.edu, Website: http://jjforneron.com.

†Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA. Email: samzl@bu.edu.

Appendix B R Code for the MA(1) Example

```
library(stats)  # fit an AR(p) model
library(pracma) # compute jacobian

n = 200         # sample size n
theta = -1/2    # MA(1) coefficient

set.seed(123)  # set the seed for random numbers
e = rnorm(n+1)               # draw innovations
y = e[2:(n+1)] - theta*e[1:n]   # generate MA(1) data
p = 12          # number of lags for the AR(p) models

beta <- function(theta) {
    # computes the p-limit of the OLS estimates
    # V = covariance matrix of (y_{t-1},...,y_{t-p})
    V = diag(p+1)*(1+theta^2)   # variances on the diagonal
    diag(V[,-1]) = -theta       # autocovariance
    V = t(V)                    # transpose
    diag(V[,-1]) = -theta       # autocovariance
    return(
        solve( V[2:(p+1),2:(p+1)], V[1,2:(p+1)] )
        # p-limit = inv(V)*( vector of autocovariances )
    )
}

# Fit the AR(p) auxiliary model:
ols_p = c(ar.ols( y, aic = FALSE, order.max = p, demean = FALSE,
   intercept = FALSE )$ar)

moments <- function(theta) {
    # computes the sample moments gn
    return( ols_p - beta(theta) ) # gn = psi_n - psi(theta)
}

objective <- function(theta,disp = FALSE) {
    # compute the sample objective Qn
    if (disp == TRUE) {
        print(round(theta,3)) # print to tack R's optimization paths
    }
    mm = moments(theta)       # compute sample moments gn
```

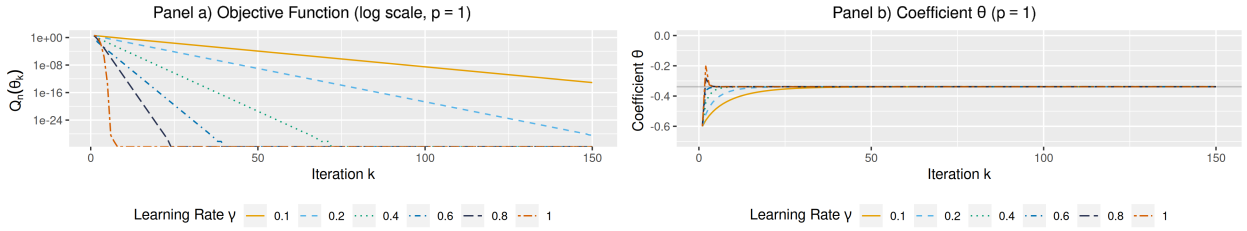
```
    return( t(mm)%*%mm )       # compute Qn = gn'*gn (W = Id)
}

dQ <- function(theta,disp=FALSE) {
    # compute the derivative of Qn
    # gradient of Qn = -2*d psi(theta)/ d theta' * gn(theta)
    return(-2*t(jacobian(beta,theta))%*%moments(theta))
}

# L-BFGS-B: with bound constraints
o1 = optim(0.95,objective,gr=dQ,method="L-BFGS-B",lower=c(-1),upper=
   c(1),disp=TRUE)
# BFGS: without bound constraints
o2 = optim(0.95,objective,gr=dQ,method="BFGS",disp=TRUE)

# **********************************
#           Gauss-Newton
# **********************************
gamma = 0.1 # learning rate
coefsGN = rep(0,150) # 150 iterations in total
coefsGN[1] = 0.95    # starting value: theta = 0.95

for (b in 2:150) { # main loop for Gauss-Newton
    Gn  = -jacobian(beta,coefsGN[b-1])  # 1. compute Jacobian
    mom =  moments(coefsGN[b-1])        # 2. compute moments
    coefsGN[b] = coefsGN[b-1] - gamma*solve(t(Gn)%*%Gn,t(Gn)%*%mom)
       # 3. update
}   # repeat for each b

# Put the results into a table:
results = matrix(NA,2,3)
colnames(results) = c('L-BFGS-B','BFGS','GN')
results[1,] = c(o1$par,o2$par,coefsGN[150])
results[2,] = sapply(results[1,],objective)
rownames(results) = c('theta','Qn(theta)')

print(results,digits=3)
# Output should look like this:
#           L-BFGS-B   BFGS     GN
# theta         -1.0 -6.979 -0.626
# Qn(theta)      1.7  0.397  0.101
```

Appendix C Additional Simulation, Empirical Results

C.1 Estimating an MA(1) model

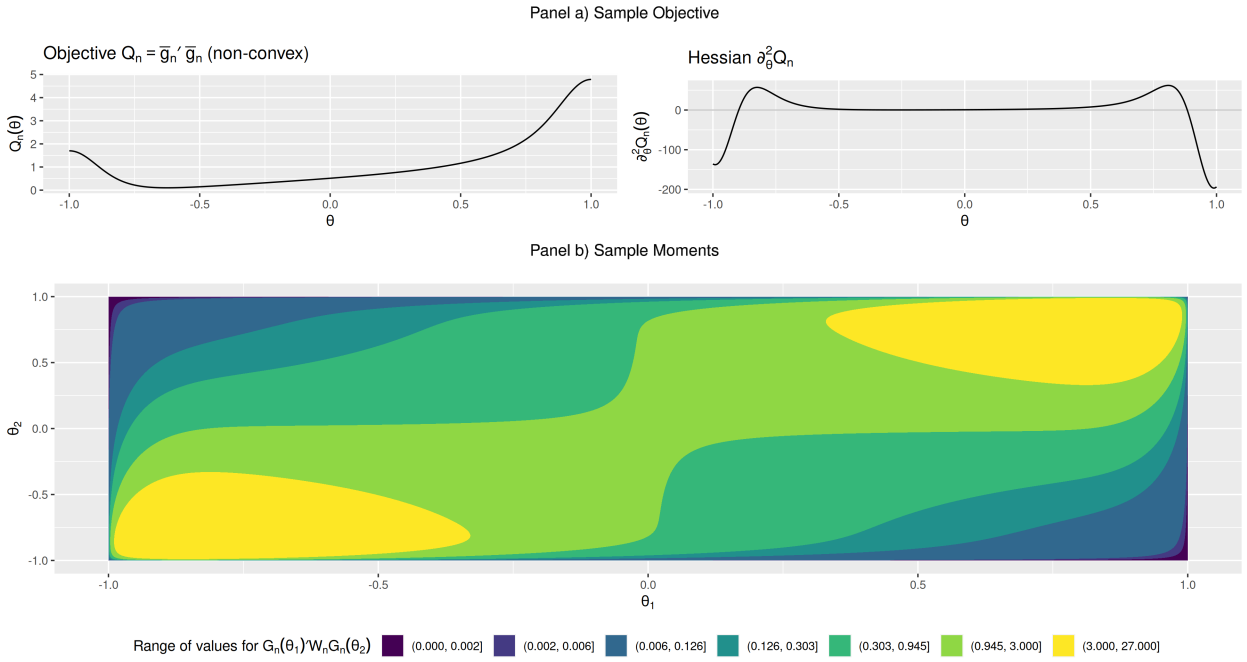
The following reports GN results with $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$, $p = 1$ and $p = 12$, equal and optimal weighting ($p = 12$).

Figure C7: GN iterations: different learning rates



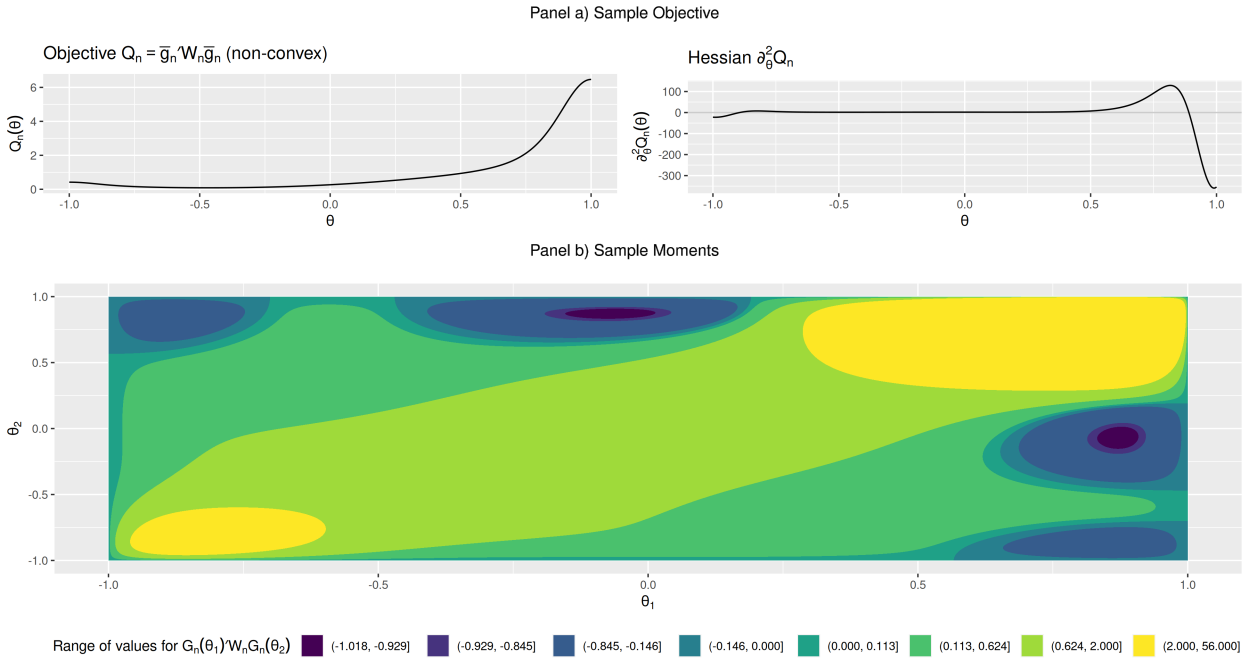
Legend: simulated sample of size $n = 200$, $\theta^\dagger = -1/2$, $\overline{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$. Grey horizontal line: $\hat{\theta}_n = -0.339$. Just Identified ($p = 1$).

Figure C8: Non-convexity and the rank condition ($p = 12$, equal weighting $W_n = I_d$)



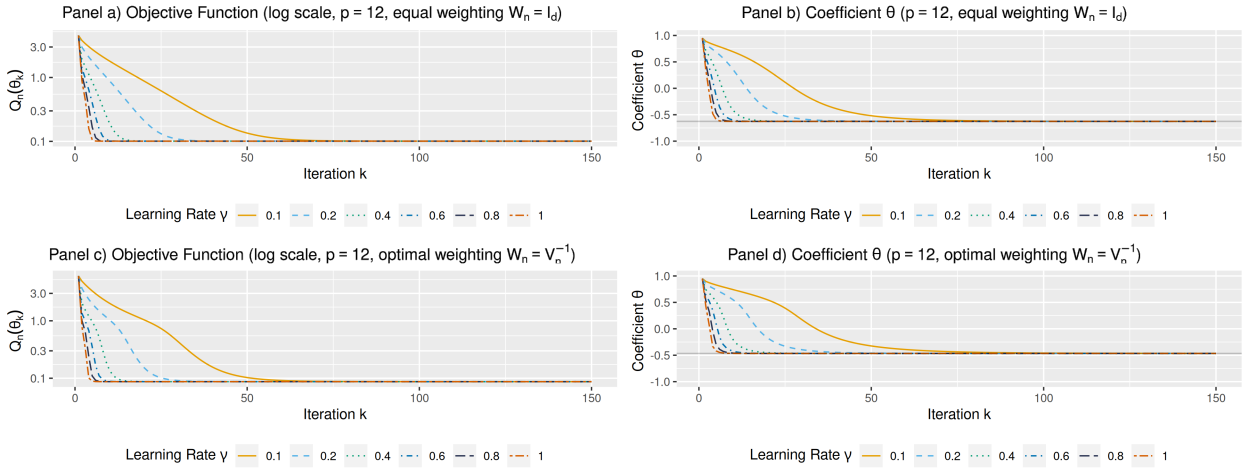
Legend: simulated sample of size $n = 200$, $\theta^\dagger = -1/2$, $\overline{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$, $W_n = I_d$. The GMM objective (panel a) is non-convex but the sample moments (panel b) satisfy the rank condition: $G_n(\theta_1)'G_n(\theta_2)$ is full rank (non-zero) for all $(\theta_1, \theta_2) \in (-1, 1) \times (-1, 1)$.

Figure C9: Non-convexity and the rank condition ($p = 12$, optimal weighting $W_n = V_n^{-1}$)



Legend: simulated sample of size $n = 200$, $\theta^\dagger = -1/2$, $\bar{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$, $W_n = V_n^{-1}$ with $V_n = n\hat{\text{var}}(\hat{\beta}_n)$. The GMM objective (panel a) is non-convex but the sample moments (panel b) does not satisfy the rank condition: $G_n(\theta_1)' V_n^{-1} G_n(\theta_2)$ is not full rank (non-zero) for all $(\theta_1, \theta_2) \in (-1, 1) \times (-1, 1)$.

Figure C10: GN iterations: equal and optimal weighting, different learning rates



Legend: simulated sample of size $n = 200$, $\theta^\dagger = -1/2$, $\bar{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$, $W_n = I_d$ (equal weighting, top panel), or $W_n = V_n^{-1}$ (optimal weighting, bottom panel), with $V_n = n\hat{\text{var}}(\hat{\beta}_n)$. Grey horizontal line: $\hat{\theta}_n = -0.626$ (equal weighting), $\hat{\theta}_n = -0.466$ (optimal weighting).

C.2 Demand for Cereal

Table C5: Demand for Cereal: GN with different learning rates

		STDEV				INCOME				objs	# of crashes
		const.	price	sugar	mushy	const.	price	sugar	mushy		
TRUE	est	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	-
	se	0.11	0.76	0.01	0.15	0.56	3.06	0.02	0.26	-	
$\gamma = 0.1$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Legend: Comparison for 50 starting values in $[-10, 10] \times \cdots \times [-10, 10]$. Avg, Std: sample average and standard deviation of optimizer outputs. TRUE: full sample estimate (est) and standard errors (se). Objs: avg and std of minimized objective value. # of crashes: optimization terminated because the objective function returned an error. GN run with $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ for $k = 150$ iterations for all starting values.

C.3 Impulse Response Matching

The following tables report results for GN using a range of tuning parameters γ. Since the rank condition (3') does not hold towards the lower bound for η, ν, GN alone can crash and/or fail to converge. Following Forneron (2023), we can introduce a global step:

$$\begin{aligned} &\theta_{k+1} = \theta_k - \gamma P_k G_n(\theta_k)' W_n \overline{g}_n(\theta_k) \\ &\text{if } \|\overline{g}_n(\theta^{k+1})\|_{W_n} < \|\overline{g}_n(\theta_{k+1})\|_{W_n}, \text{ set } \theta_{k+1} = \theta^{k+1} \end{aligned} \tag{1}$$

where the sequence $(\theta^k)_{k \geq 0}$ is predetermined and dense in Θ. The results rely on the Sobol sequence, independently randomized for each of the 50 starting values.[1] Results are reported with and without the global step. Also, the former implements error-handling (try-catch).

[1] We take $(s_k)_{k \geq 0}$ in $[0, 1]^p$, $p \geq 1$ is the number of parameters, draw one vector $(u_1, \ldots, u_p) \sim \mathcal{U}_{[0,1]^p}$, for each starting value, and compute $\tilde{s}_k = (s_k + u)$ modulo 1, then map $\tilde{s}_k$ to the bounds for $\theta = (\theta_1, \ldots, \theta_p)$. The randomization is used to create independent variation in the global step between starting values to emphasize that convergence does not rely on a specific value in the sequence $(\theta^k)_{k \geq 0}$; this is called a random shift (see Lemieux, 2009, Ch6.2.1).

Table C6: Without reparameterization : GN with different learning rates

		η	ν	ρ_s	σ_s	objs	# of crashes
TRUE	est	0.30	0.29	0.39	0.17	4.65	-
GN WITHOUT GLOBAL STEP							
$\gamma = 0.1$	avg	0.30	0.29	0.39	0.17	4.65	2
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.30	0.29	0.39	0.17	4.65	2
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.30	0.29	0.39	0.17	4.65	4
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.30	0.29	0.39	0.17	4.65	8
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.30	0.29	0.39	0.17	4.65	12
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.30	0.29	0.39	0.17	4.65	28
	std	0.00	0.00	0.00	0.00	0.00	
GN WITH GLOBAL STEP							
$\gamma = 0.1$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
lower bound		0.05	0.01	-0.95	0.01	-	-
upper bound		0.99	0.90	0.95	12	-	-

Legend: Comparison for 50 starting values. TRUE: full sample estimate (est). GN WITH GLOBAL STEP: Gauss-Netwon augmented with a global sequence. Both are run for $k = 150$ iterations in total, for all starting values. Objs: avg and std of minimized objective value. # of crashes: optimization terminated because objective returned error. Lower/upper bound used for the reparameterization.

Table C7: With reparameterization : GN with different learning rates

		η	ν	ρ_s	σ_s	objs	# of crashes
TRUE	est	0.30	0.29	0.39	0.17	4.65	-
GN WITHOUT GLOBAL STEP							
$\gamma = 0.1$	avg	0.30	0.29	0.39	0.17	4.65	5
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.30	0.29	0.39	0.17	4.65	10
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.30	0.29	0.39	0.17	4.65	20
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.30	0.29	0.39	0.17	4.65	22
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.30	0.29	0.39	0.17	4.65	25
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.30	0.29	0.39	0.17	4.65	29
	std	0.00	0.00	0.00	0.00	0.00	
GN WITH GLOBAL STEP							
$\gamma = 0.1$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
lower bound		0.05	0.01	-0.95	0.01	-	-
upper bound		0.99	0.90	0.95	12	-	-

Legend: Comparison for 50 starting values. TRUE: full sample estimate (est). GN WITH GLOBAL STEP: Gauss-Netwon augmented with a global sequence. Both are run for $k = 150$ iterations in total, for all starting values. Objs: avg and std of minimized objective value. # of crashes: optimization terminated because objective returned error. Lower/upper bound used for the reparameterization.

Appendix D Additional Material for Section 2.2

The Nelder-Mead algorithm. The following description of the algorithm is based on Nash (1990, Ch14) which R implements in the optimizer *optim*. The first step is to build a simplex for the p-dimensional parameters, i.e. $p+1$ distinct points $\theta_1, \ldots, \theta_{p+1}$ ordered s.t. $Q_n(\theta_1) \leq \cdots \leq Q_n(\theta_{p+1})$. The simplex is then transformed at each iteration using four operations called *reflection*, *expansion*, *reduction*, and *contraction*. The algorithm also repeatedly computes the centroid θ_c of the best p points, to do so: take the best p guesses $\theta_1, \ldots, \theta_p$ and compute their average: $\theta_c = 1/p \sum_{\ell=1}^{p} \theta_\ell$. Once this is done, go to step **R** below.

Nelder-Mead Algorithm:

Inputs: Initial simplex $\theta_1, \ldots, \theta_{p+1}$, parameters $\alpha, \gamma, \beta, \beta'$. NM suggest to use $\alpha = 1, \gamma = 2, \beta = \beta' = 1/2$.

Re-order the points so that $Q_n(\theta_1) \leq \cdots \leq Q_n(\theta_{p+1})$, compute the centroid $\theta_c = 1/p \sum_{\ell=1}^{p} \theta_\ell$ (average of the best p points)

Start at **R** and run until convergence:

R: The *reflection* step computes $\theta_r = \theta_c + \alpha(\theta_c - \theta_{p+1}) = 2\theta_c - \theta_{p+1}$ for $\alpha = 1$. There are now several possibilities:

- If $Q_n(\theta_r) < Q_n(\theta_1)$ got to step **E**.
- If $Q_n(\theta_1) \leq Q_n(\theta_r) \leq Q_n(\theta_p)$, replace θ_{p+1} with θ_r, re-order the points, compute the new θ_c, and do **R** again.
- By elimination: $Q_n(\theta_r) > Q_n(\theta_p)$. If $Q_n(\theta_r) < Q_n(\theta_{p+1})$, replace θ_{p+1} with θ_r. Either way, go to step **R'**.

E: The *expansion* step computes $\theta_e = \theta_r + (\gamma - 1)(\theta_r - \theta_c) = 2\theta_r - \theta_c$ for $\gamma = 2$. If $Q_n(\theta_e) < Q_n(\theta_r)$, then θ_e replaces θ_{p+1}. Otherwise, θ_r replaces θ_{p+1}. Once θ_{p+1} is replaced, re-order the points, compute the new θ_c, and go to **R**.

R': The *reduction* step computes $\theta_s = \theta_c + \beta(\theta_{p+1} - \theta_c) = (\theta_c + \theta_{p+1})/2$ for $\beta = 1/2$. If $Q_n(\theta_s) < Q_n(\theta_{p+1})$, θ_s replaces θ_{p+1}, then re-order the points, compute the new θ_c, and go to **R**. Otherwise, go to **C**.

C: The *contraction* step updates $\theta_2, \ldots, \theta_{p+1}$ using $\theta_\ell = \theta_1 + \beta'(\theta_\ell - \theta_1) = (\theta_\ell + \theta_1)/2$ for $\beta' = 1/2$. Re-order the points, compute the new θ_c, and go to **R**.

Clearly, the choice of initial simplex can affect the convergence of the algorithm. Typically, one provides a starting value θ_1 and then the software picks the remaining p points of

the simplex without user input. NM proposed their algorithm with statistical estimation in mind, so they considered using the standard deviation $\sqrt{\sum_{\ell=1}^{n+1}(Q_n(\theta_\ell) - \bar{Q}_n)^2/n} < \text{tol}$ as a convergence criterion, setting $\text{tol} = 10^{-8}$ and $\bar{Q}_n$ the average of $Q_n(\theta_\ell)$ in their application. Here convergence occurs when the simplex collapses around a single point.

The Grid-Search algorithm. The procedure is very simple, pick a grid of k points $\theta_1, \ldots, \theta_k$, and compute:

$$\tilde{\theta}_k = \text{argmin}_{\ell=1,\ldots,k} Q_n(\theta_\ell).$$

The optimization error $\|\tilde{\theta}_k - \hat{\theta}_n\|$ depends on both k and the choice of grid. The following gives an overview of the approximation error and feasible error rates.

For simplicity, suppose that the parameter space is the unit ball in $\mathbb{R}^p$: $\Theta = \mathcal{B}_2^p$, and Q_n is continuous. Under these assumptions, there is an $L \geq 0$ such that $|Q_n(\theta_1) - Q_n(\theta_2)| \leq L\|\theta_1 - \theta_2\|$. $L > 0$, unless Q_n is constant. This implies: $|Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n)| \leq L(\inf_{1\leq\ell\leq k} \|\theta_\ell - \hat{\theta}_n\|)$. Suppose we want to ensure $|Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n)| \leq \varepsilon$, then we need $\inf_{1\leq\ell\leq k} \|\theta_\ell - \hat{\theta}_n\| \leq \varepsilon/L$. Packing arguments (e.g. Vershynin, 2018, Proposition 4.2.12) give a lower bound for k over all grids, and all possible $\hat{\theta}_n$: $k \geq \text{vol}(\mathcal{B}_2^p)/\text{vol}([\varepsilon/L]\mathcal{B}_2^p) = [\varepsilon/L]^{-p}$, where vol is the volume.

For the choice of grid, Niederreiter (1983, Theorem 3) shows that low-discrepancy sequences, e.g. the Sobol or Halton points sets, can achieve this rate, up to a logarithmic term.[2] This is indeed a common choice for multi-start and grid search optimization.

In practice, $Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n)$ is typically not the quantity of interest for empirical estimations, rather we are interested in $\|\tilde{\theta}_k - \hat{\theta}_n\|$. Suppose, in addition, that $\hat{\theta}_n \in \text{int}(\Theta)$, and Q_n is twice continuously differentiable with positive definite Hessian $H_n(\hat{\theta}_n)$, a local identification condition. Then there exists $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$ and $\varepsilon_1 > 0$ s.t. $\|\theta - \hat{\theta}_n\| \leq \varepsilon_1$ implies:

$$\underline{\lambda}\|\theta - \hat{\theta}_n\|^2 \leq Q_n(\theta) - Q_n(\hat{\theta}_n) \leq \overline{\lambda}\|\theta - \hat{\theta}_n\|^2, \tag{D.10}$$

i.e. Q_n is locally strictly convex.[3] If $\hat{\theta}_n$ is the unique minimizer of Q_n, there is a $0 < \varepsilon_2 \leq \varepsilon_1$ such that $\inf_{\|\theta - \hat{\theta}_n\| \geq \varepsilon_1} Q_n(\theta) > Q_n(\hat{\theta}_n) + \overline{\lambda}\varepsilon_2^2$, using a global identification condition. Now, by local identification: $\|\theta - \hat{\theta}_n\| \leq \varepsilon_2 \Rightarrow Q_n(\theta) \leq Q_n(\hat{\theta}_n) + \overline{\lambda}\varepsilon_2^2 < \inf_{\|\theta - \hat{\theta}_n\| \geq \varepsilon_1} Q_n(\theta)$. As soon as $k \geq k_0$ where $\inf_{1\leq\ell\leq k_0} \|\theta_\ell - \hat{\theta}_n\| \leq \varepsilon_2$, we have $\|\tilde{\theta}_k - \hat{\theta}_n\| \leq \varepsilon_1$. Then, for any $k \geq k_0$: $\underline{\lambda}\|\tilde{\theta}_k - \hat{\theta}_n\|^2 \leq Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n) \leq \overline{\lambda}(\inf_{1\leq\ell\leq k} \|\theta_\ell - \hat{\theta}_n\|^2)$ and $\|\tilde{\theta}_k - \hat{\theta}_n\| \leq [\overline{\lambda}/\underline{\lambda}]^{1/2}(\inf_{1\leq\ell\leq k} \|\theta_\ell - \hat{\theta}_n\|)$.

[2] In comparison, using uniform random draws in a grid search would require $O([\varepsilon/L]^{-2p})$ iterations to achieve the same level of accuracy with high-probability. Fang and Wang (1993, Ch3.1) give a review of these results.

[3] The three $\varepsilon_1, \underline{\lambda}, \overline{\lambda}$ only depend on $H_n(\cdot)$.

This reveals the interplay between the identification conditions and the optimization error. The best value $\tilde{\theta}_k$ is only guaranteed to be near $\hat{\theta}_n$ when $k \geq \varepsilon_2^{-p}$ iterations (using packing arguments for the unit ball), where ε_2 depends on the global identification condition. Local convergence depends on the ratio $\overline{\lambda}/\underline{\lambda} \geq 1$ which is infinite when $H_n(\hat{\theta}_n)$ is singular. The main drawback of a grid search is its slow convergence. To illustrate, Colacito et al. (2018, pp3443-3445) estimate $p = 5$ parameters using a grid search with $k = 1551$ points. For simplicity, suppose $\overline{\lambda}/\underline{\lambda} = 1$, $k_0 < k$, and $\Theta = \mathcal{B}_2^p$, the unit ball, then the worst-case optimization error is $\sup_{\hat{\theta}_n \in \Theta}(\inf_{1 \leq \ell \leq k} \|\theta_\ell - \hat{\theta}_n\|) \geq k^{-1/p} \simeq 0.23$. This is ten times larger than all but one of the standard errors reported in the paper.

Simulated Annealing. Implementations can vary across software, the following will focus on the implementation used in R's *optim* function.

Simulated Annealing Algorithm:

Inputs: Starting value $\theta_1 \in \Theta$, temperature schedule $\infty > T_2 \geq T_3 \geq \cdots > 0$, a sequence $\infty > \eta_2 \geq \eta_3 \geq \cdots > 0$, and maximum number of iterations k. Common choice: $T_\ell = T_1/\log(\ell)$ for $\ell \geq 2$ and η_ℓ proportional to T_ℓ.

For $\ell \in \{2, \ldots, k\}$, repeat:

1. Draw $\theta^\star \sim \mathcal{N}(\theta_{\ell-1}, \eta_\ell I_d)$, and $u_\ell \sim \mathcal{U}_{[0,1]}$
2. Set $\theta_\ell = \theta^\star$ if $u_\ell \leq \exp(-[Q_n(\theta^\star) - Q_n(\theta_{\ell-1})]/T_\ell)$, otherwise set $\theta_\ell = \theta_{\ell-1}$

Output: Return $\tilde{\theta}_k = \text{argmin}_{1 \leq \ell \leq k} Q_n(\theta_\ell)$

The implementation described above relies on the random-walk Metropolis update. Notice that if $Q_n(\theta^\star) \leq Q_n(\theta_{\ell-1})$, the exponential term in step 2 is greater than 1 and $\theta^\star$ is always accepted as the next θ_ℓ, regardless of u_ℓ. Bélisle (1992) gave sufficient condition for $\tilde{\theta}_k \xrightarrow{a.s.} \hat{\theta}_n$ when $k \to \infty$ and Q_n is continuous. In practice, the performance of the Algorithm can be measured by its convergence rate. To get some intuition, we give some simplified derivations below which highlight the role of T_k and several quantities which appeared in our discussion of the grid search.

First, notice that for each k, steps 1-2 implement the Metropolis algorithm also used for Bayesian inference using random-walk Metropolis-Hastings. The invariant distribution of these two steps is:

$$f_k(\theta) = \frac{\exp(-[Q_n(\theta) - Q_n(\hat{\theta}_n)]/T_k)}{\int_\Theta \exp(-[Q_n(\theta) - Q_n(\hat{\theta}_n)]/T_k) d\theta},$$

this is called the Gibbs-Boltzmann distribution. When $T_\infty = +\infty$, f_∞ puts all the probability mass on the unique minimum $\hat{\theta}_n$. To build intuition, suppose that $k \geq 1$: $\theta_k \sim f_k$. Because SA is a stochastic algorithm, the approximation error $\|\theta_k - \hat{\theta}_n\|$ is random, but can be quantified using $\mathbb{P}(\|\theta_k - \hat{\theta}_n\| \geq \varepsilon)$. In the following we will assume the temperature schedule to be $T_k = T_1/\log(k)$, as implemented in R.

The following relies on the same setting, notation and assumptions as the grid search above. First, we can bound the probability that θ_k is outside the ε_1-local neighborhood of $\hat{\theta}_n$ where Q_n is approximately quadratic: $\mathbb{P}(\|\theta_k - \hat{\theta}_n\| \geq \varepsilon_1)$. Using the global identification condition:

$$\exp(-[Q_n(\theta) - Q_n(\hat{\theta}_n)]/T_k) \leq \exp(-\overline{\lambda}\varepsilon_2^2/T_k) = k^{-\overline{\lambda}\varepsilon_2^2/T_1}, \text{ if } \|\theta - \hat{\theta}_n\| \geq \varepsilon_1,$$

where ε_1, ε_2 were defined in the grid search section above. This gives an upper bound for the numerator in $f_k(\theta_k)$. A lower bound is also required for the denominator. Using (D.10) and the change of variable $\theta = \hat{\theta}_n + \sqrt{T_k}h$, we have:

$$\exp(-\overline{\lambda}\|h\|^2) \leq \exp(-[Q_n(\hat{\theta}_n + \sqrt{T_k}h) - Q_n(\hat{\theta}_n)]/T_k) \leq \exp(-\underline{\lambda}\|h\|^2), \text{ if } \|\sqrt{T_k}h\| \leq \varepsilon_1.$$

Suppose $T_k \leq \varepsilon_1^2$, the two inequalities give us the bound:

$$\mathbb{P}(\|\theta_k - \hat{\theta}_n\| \geq \varepsilon_1) \leq \frac{k^{-\overline{\lambda}\varepsilon_2^2/T_1}\text{vol}(\Theta)}{|T_k|^{p/2}\int_{\|h\|\leq 1}\exp(-\overline{\lambda}\|h\|^2)dh} = C[\log(k)]^{d/2}k^{-\overline{\lambda}\varepsilon_2^2/T_1}.$$

This upper bound declines more slowly than for the grid search when $\overline{\lambda}\varepsilon_2^2/T_1 < 1/p$, which can be the case if T_1 large and/or ε_2 is small. For the lower bound, pick any $\varepsilon \in (0, \varepsilon_1/\sqrt{T_k})$:

$$\mathbb{P}(\|\theta_k - \hat{\theta}_n\| \leq \sqrt{T_k}\varepsilon) \geq \frac{\int_{\|h\|\leq\varepsilon}\exp(-\overline{\lambda}\|h\|^2)dh}{\int_{\|h\|\in\mathbb{R}}\exp(-\underline{\lambda}\|h\|^2)dh + |T_k|^{-p/2}\text{vol}(\Theta)k^{-\overline{\lambda}\varepsilon_2^2/T_1}},$$

which has a strictly positive limit. This implies that $\sqrt{\log(k)}\|\theta_k - \hat{\theta}_n\| \geq O_p(1)$, since $T_k = T_1/\log(k)$. This $\sqrt{\log(k)}$ rate is slower than the grid search. To get faster convergence, some authors have suggested using $T_k = T_1/k$ and, by default, Matlab sets $T_k = T_1 \cdot 0.95^k$. However, theoretical guarantees to have $\theta_k \xrightarrow{p} \hat{\theta}_n$, as $k \to \infty$ are only available when $T_k = T_1/\log(k)$.[4]

[4]See Spall (2005, Ch8.4-8.6) for additional details and references.