# Convexity Not Required: Estimation of Smooth Moment Condition Models

Jean-Jacques Forneron* Liang Zhong†

July 11, 2025

**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 conditions involving the Jacobian of the 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 moderately non-linear reparameterizations, and 3) moderate misspecification. The conditions preclude non-global optima. Numerical and empirical examples illustrate the condition, non-convexity, and convergence properties of different optimizers.

JEL Classification: C11, C12, C13, C32, C36.
Keywords: Non-linear estimation, over-identification, misspecification, nonlinear systems of equations, injectivity, local and global identification.

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

†Faculty of Business and Economics, The University of Hong Kong, Pokfulam Road, Hong Kong. Email: samzl@hku.hk, Website: https://samzl1.github.io/.

This paper was written while the second author was a doctoral student at Boston University. The authors would like to thank Jessie Li for suggesting to look at misspecified models and Hiro Kaido, David Lakagos, Bernard Salanié and participants at the NY Camp Econometrics Conference for useful comments.

# 1 Introduction

The Generalized and Simulated Method of Moments (GMM, SMM) are commonly used to estimate structural Economic models. To find 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. While sampling properties of estimators are often derived, their practical implementation often receives a less detailed treatment. There is now a vast literature on statistical learning with a convex loss function, using stochastic gradient-descent. However, these results need not directly apply to GMM, as it often involves non-convex minimizations. A number of authors have pointed out the lack of robustness of off-the-shelf methods, and Knittel and Metaxoglou (2014) illustrate this in the context of demand estimation. This is perhaps not surprising since non-convex optimization is subject to a curse of dimensionality (Andrews, 1997, Section 2) and becomes increasingly challenging when the number of parameters is moderate or large.

The main contribution of the paper is to show that *convexity is not required* for some methods to perform well in GMM estimation specifically: some algorithms are globally convergent under a global rank condition involving the Jacobian of the moments and the weighting matrix. This defines a class of non-convex problems that is as hard as convex problems for optimization. Since this is perhaps surprising, the following gives some intuition behind the result. Given sample moments $\overline{g}_n(\theta)$ with Jacobian $G_n(\theta)$, one can minimize the GMM objective function $Q_n(\theta) = 1/2\overline{g}_n(\theta)'W_n\overline{g}_n(\theta)$ iteratively, by minimizing successive quadratic approximations. To this end, convex optimizers rely on a quadratic expansion of $Q_n(\theta)$ using its gradient and Hessian. This quadratic approximation yields a proper minimization problem only if the Hessian is strictly positive definite, i.e. $Q_n$ is convex.

Another approach, is to linearly expand the sample moments using the Jacobian and plug the linearized moments into the GMM objective. Since the approximate moments are linear, this yields a proper minimization problem as long as $G_n$ has full rank. Gauss-Newton (GN) relies on this approach. Gradient-descent can be motivated by either the quadratic of linear approximation. In the just-identified case, it is well known that GN is *locally convergent* when $G_n$ has full rank around the solution. This paper goes further by showing that GN and gradient-descent are *globally convergent* when the product of $G_n$, $W_n$, and an average of $G_n$ has full rank everywhere. Unlike existing results, this applies to just and over-identified moments. The condition can be relaxed for the product to only be non-singular in a specific direction, towards the global minimizer. Under this weaker condition, GN with a Levenberg-Marquardt regularization and gradient descent are globally convergent. Importantly, for

correctly specified models, the conditions imply that there are no local optima, besides the global minimizer; a necessary condition for global convergence of gradient-based optimizers. It is shown that these convergence results are robust to 1) moderate misspecification, and 2) moderately non-linear reparameterizations. However, the results may or may not hold depending on the choice of weighting matrix. In particular, when $W_n$ is ill-conditioned, convergence can be significantly slower.

Several conditions found in the convex and non-convex optimization literatures imply the weaker condition introduced of this paper. These include strong, star, and quasar convexity of the objective function. It also relates to the Polyak-Łojasiewicz condition, an important inequality which has gathered much interest in machine learning to prove convergence of gradient-descent. Strong monotonicity of the moments, a condition for solving just-determined system of non-linear equations, and strong injectivity, introduced here for just and over-identified models, also imply the weaker condition. Hence, the condition introduced in this paper is a common denominator of several existing conditions. In terms of econometrics properties, the conditions are sufficient for the parameters to be both locally and globally identified, when the model is correct or moderately misspecified.

A simple MA(1) estimation from Gourieroux and Monfort (1996) illustrates the results analytically and numerically. The problem is non-convex: the scalar Hessian can be positive, negative, or zero; yet the conditions hold. As predicted, the recommended GN algorithm converges. Newton-Raphson provably diverges, and off-the-shelf optimizers can be unstable. When the model is moderately misspecified, GN remains globally convergent. In line with theory, significant misspecification can produce non-global optima which hinder the global convergence of gradient-descent and Gauss-Newton.

Two empirical applications further illustrate the results. The first application revisits the numerical results of Knittel and Metaxoglou (2014) for estimating random coefficient demand models on Nevo's generated cereal data. The same GN algorithm systematically converges from a wide range of starting values. 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 whereas GN performs well for nearly all starting values.

Numerically, in all three applications, the GMM objective is non-convex at most values. The strong injectivity condition holds at most values, an indication that GN and gradient-

descent are appropriate. The later converges very slowly, however. These findings explain the good performance of GN relative to more commonly used methods. The main takeaway is that non-convexity need not be a deterrent to structural estimation: simple algorithms can converge quickly and globally under alternative conditions.

**Structure of the paper.** Section 2 contains the main assumptions and results. Section 3 reviews existing conditions found in the literature and relates the main assumptions with these conditions. Section 4 suggests a numerical procedure to check whether the main assumption holds or not and a way to set the tuning parameter. Section 5 illustrates the results with one numerical and two empirical applications. Appendices A and B give the proofs to the main results and additional results. The Supplemental Material consists of: Appendices C-G. Appendix C provides additional local convergence results, which complement the main global convergence results in the paper. Appendix D provides of survey of empirical practice in the American Economic Review between 2016 and 2018. Appendix E gives R code to replicate the numerical MA(1) example. Appendix F provides additional simulation and empirical results. Appendix G gives additional details about the methods found in the survey of Appendix D.

**Notation:** In the following $\lambda_{\min}$, $\lambda_{\max}$ return the smallest and largest eigenvalues of a square positive semidefinite matrix. For an arbitrary rectangular matrix $A$ of size $d_g \times d_\theta$ with $d_g \geq d_\theta$, $\sigma_{\min}$, $\sigma_{\max}$ are the smallest and largest singular values of $A$ defined as $\sigma_{\min}(A) = \sqrt{\lambda_{\min}(A'A)}$ and $\sigma_{\max}(A) = \sqrt{\lambda_{\max}(A'A)}$; $A$ has full rank if, and only if, $\sigma_{\min}(A) > 0$.

# 2 GMM Estimation without 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. Their population counterparts are $g(\theta) = \mathbb{E}[g(\theta; x_i)]$ and $G(\theta) = \partial_\theta g(\theta)$. $W_n$ is a weighting matrix which, for simplicity, does not depend on $\theta$ – this excludes continuously-updated estimations. The sample GMM objective function is:

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

and the goal is to find the global minimizer $\hat{\theta}_n$ of $Q_n$ in $\mathbb{R}^{d_\theta}$. The population objective $Q(\theta) = \frac{1}{2} g(\theta)' W g(\theta)$, defined similarly using the limit $W$ of $W_n$, has a global minimizer $\theta^\dagger$. Throughout, it will be assumed that the sample $Q_n$ is continuously differentiable. More

specifically, this paper considers derivative-based optimizers of the form:

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

for $k = 0, 1, \ldots$, some staring value $\theta_0 \in \mathbb{R}^{d_\theta}$ and a matrix $P_{k,n}$, called conditioning matrix, assumed to be symmetric. The tuning parameter $\gamma \in (0, 1]$ is called the learning rate. There are several ways to motivate (1) as a minimization algorithm in the context of GMM estimation. They are conceptually similar but implicitly rely on a different set of assumptions. The first is to consider a quadratic approximation of the GMM objective function $Q_n$:

$$Q_n(\theta) \simeq Q_n(\theta_k) + \partial_\theta Q_n(\theta_k)(\theta - \theta_k) + \frac{1}{2\gamma}(\theta - \theta_k)' \partial^2_{\theta,\theta'} Q_n(\theta_k)(\theta - \theta_k),$$

here $\gamma$ penalizes the quality of the quadratic approximation. For linear models, such as OLS and IV regressions, $Q_n$ is quadratic so that $\gamma = 1$ is feasible. For non-linear models, the approximation is inexact, and $\gamma < 1$ is generally required. Minimizing the right-hand-side with respect to $\theta$ yields a Newton-Raphson (NR) iteration: $\theta_{k+1} = \theta_k - \gamma[\partial^2_{\theta,\theta'} Q_n(\theta_k)]^{-1} \partial_\theta Q_n(\theta_k)$ with $\partial_\theta Q_n(\theta_k) = G_n(\theta_k)' W_n \overline{g}_n(\theta_k)$ and $P_{k,n} = [\partial^2_{\theta,\theta'} Q_n(\theta_k)]^{-1}$. A quasi-Newton (QN) iterations replaces the Hessian matrix $\partial^2_{\theta,\theta'} Q_n(\theta_k)$ with an approximation computed sequentially over $k$. The most popular QN software implementation is called BFGS. Importantly, the quadratic approximation implicitly requires that is $H_n(\theta) = \partial^2_{\theta,\theta'} Q_n(\theta_k)$ strictly positive definite around $\theta_k$ so that (1) yields a minimizer of the quadratic approximation.

Another way to motivate (1) is to consider a linear approximation of the moments and plug it into the GMM objective function:

$$\begin{aligned}
\overline{g}_n(\theta) &\simeq \overline{g}_n(\theta_k) + \frac{1}{\gamma} G_n(\theta_k)(\theta - \theta_k), \\
Q_n(\theta) &\simeq \frac{1}{2}\Big[\overline{g}_n(\theta_k) + \frac{1}{\gamma} G_n(\theta_k)(\theta - \theta_k)\Big]' W_n \Big[\overline{g}_n(\theta_k) + \frac{1}{\gamma} G_n(\theta_k)(\theta - \theta_k)\Big],
\end{aligned}$$

where now $\gamma$ penalizes the quality of the linear approximation. Take the first order condition in the last display to find (1) with $P_{k,n} = (G_n(\theta_k)' W_n G_n(\theta_k))^{-1}$, a Gauss-Newton (GN) iteration. The quadratic approximation requires the Hessian $H_n$ of $Q_n$ to be strictly positive definite at $\theta_k$. A GN iteration minimizes the linear approximation as long as the Jacobian $G_n$ of $\overline{g}_n$ has full rank at $\theta_k$ so that $G_n(\theta_k)' W_n G_n(\theta_k)$ is strictly positive definite. Standard regularity condition imply local convexity around $\hat{\theta}_n$. Still, convexity is more challenging to satisfy away from the solution since $\|\overline{g}_n(\theta_k)\| \gg 0$ can result in a non-definite Hessian $H_n(\theta_k) = G_n(\theta_k)' W_n G_n(\theta_k) + (\overline{g}_n(\theta_k)' W_n \otimes I_d) \partial_\theta \text{vec}[G_n(\theta_k)']$, depending on the last term.

This suggests that quadratic-based methods (NR, BFGS) and linear-based methods (GN) can behave differently when $Q_n$ is globally non-convex. Gradient-Descent (GD) can be motivated by either a linear or a quadratic approximation. The following summarizes the choice of $P_{k,n}$ for each algorithm:

Table 1: Optimizers considered in (1)

| | | |
|---|---|---|
| 1. | Gradient-Descent (GD) | $P_{k,n} = I_d$, |
| 2. | Newton-Raphson (NR) | $P_{k,n} = [\partial^2_{\theta,\theta'} Q_n(\theta_k)]^{-1}$, |
| 3. | quasi-Newton (QN) | $P_{k,n}$ approximates $[\partial^2_{\theta,\theta'} Q_n(\theta_k)]^{-1}$, |
| 4. | Gauss-Newton (GN) | $P_{k,n} = [G_n(\theta_k)' W_n G_n(\theta_k)]^{-1}$. |

## 2.1 Main Assumptions

The following gives the main assumptions on the population moments used to describe the large sample properties of the estimator $\hat{\theta}_n$ and optimization algorithms.

**Assumption 1.** *The observations $x_i$ are iid and:*

*(i)* $Q(\theta) = 1/2\|g(\theta)\|^2_W$ *has a unique minimizer* $\theta^\dagger \in \mathbb{R}^{d_\theta}$,

*(ii)* $g(\theta; x_i)$ *and* $g(\theta) = \mathbb{E}[g(\theta; x_i)]$ *are continuously differentiable on* $\mathbb{R}^{d_\theta}$,

*(iii) for all* $\theta \in \mathbb{R}^{d_\theta}$: $\mathbb{E}[\|G(\theta; x_i)\|^2] < \infty$, $\mathbb{E}[\|g(\theta; x_i)\|^2] < \infty$, $\sigma_{\max}[G(\theta)] < \overline{\sigma} < \infty$; *there exists* $\bar{L}(\cdot) \geq 0$ *such that* $\mathbb{E}[\bar{L}(x_i)] < L < \infty$, $\mathbb{E}[|\bar{L}(x_i)|^2] < \infty$, *and for all* $\theta_1, \theta_2 \in \mathbb{R}^{d_\theta}$: $\|G(\theta_1; x_i) - G(\theta_2; x_i)\| \leq \bar{L}(x_i)\|\theta_1 - \theta_2\|$,

*(iv) there exists* $R_G > 0$ *such that* $\sigma_{\min}[G(\theta)] > \underline{\sigma} > 0$ *for all* $\|\theta - \theta^\dagger\| < R_G$,

*(v) there exists* $\bar{M}(\cdot)$ *such that* $\mathbb{E}[|\bar{M}(x_i)|^2] < \infty$, $\mathbb{E}[\bar{M}(x_i)] < M < \infty$, *and for any* $R > 0$, $\|G(\theta; x_i) - G(\theta_R; x_i)\| \leq \bar{M}(x_i)/(1+R)$, *where* $\theta_R = \frac{R}{\|\theta\|}\theta$ *if* $\|\theta\| > R$, $\theta_R = \theta$ *otherwise,*

*(vi)* $W_n \xrightarrow{p} W$, $0 < \underline{\lambda}_W < \lambda_{\min}(W) \leq \lambda_{\max}(W) < \overline{\lambda}_W < \infty$.

Assumption 1 consists mainly of standard conditions to derive asymptotic properties for $\hat{\theta}_n$. The *iid* assumption can be relaxed to allow for time-series dependence. The parameter space is unbounded to accommodate the unconstrained optimization. The technical condition (v) and the next Assumption imply that $Q_n$ has a strictly quadratic lower bound. This ensures consistency without assuming compactness or uniform consistency of the sample moments. The quantity $\sigma_{\min}[G(\theta)]$ in the local identification condition refers to the smallest singular value of $G(\theta)$. The main Assumption 2 below will rely on the following quantities:

$$\overline{G}(\theta) = \int_0^1 G(\omega\theta + (1-\omega)\theta^\dagger)d\omega, \quad \overline{G}(\theta_1, \theta_2) = \int_0^1 G(\omega\theta_1 + (1-\omega)\theta_2)d\omega.$$

The matrix $\overline{G}(\theta)$ is an average derivative over the path from $\theta$ to the solution $\theta^\dagger$. The matrix plays a role in the mean-value identity: $g(\theta_1) - g(\theta_2) = \overline{G}(\theta_1, \theta_2)(\theta_1 - \theta_2)$ (see Lemma A1).

**Assumption 2.** *There exists* $0 < \rho < \underline{\sigma}\underline{\lambda}_W/2$ *such that, for all* $\theta \in \mathbb{R}^{d_\theta}$*, either:*
*(a)* $\sigma_{\min}[G(\theta)'W\overline{G}(\theta)] > \rho\underline{\sigma}$*, or*
*(b)* $\|G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)\| > \rho\underline{\sigma}\|\theta - \theta^\dagger\|$.

Assumption 2 gives the main conditions used in this paper for global GMM estimation of just and over-identified models.[1] Assumption 2 (a) replaces the convexity condition $0 < \underline{\lambda}_H \leq \lambda_{\min}[H_n(\theta)] \leq \lambda_{\max}[H_n(\theta)] < \overline{\lambda}_H < \infty$ used to derive convergence results for GD, NR and QN.[2], which may not hold for GMM. A sufficient, but restrictive, condition for Assumption 2 (a) is that $g$ is the derivative of a convex function, for instance a Probit log-likelihood function. Further sufficient conditions are listed in Section 3. Assumption 2 (a) implies Assumption 2 (b); the latter is the weaker condition. Assumption 2 (a) implies that $G(\theta)$ has full rank for all $\theta$, Assumption 2 (b) only requires $G(\theta)'W\overline{G}(\theta)$ to be non-singular in the relevant direction $(\theta - \theta^\dagger)$. For over-identified models, both conditions (a) and (b) depend on the choice of weighting matrix $W$. Indeed, unlike square matrices, the product of full rank rectangular matrices does not automatically have full rank,[3] and the weighting matrix changes the way $G$ and $\overline{G}$ are multiplied. It is possible for the product to be singular even when $G$ and $\overline{G}$ have full rank. Importantly, Assumption 2 may or may not hold depending on the choice of weighting matrix $W$. If Assumption 2 is not satisfied using the preferred weighting matrix, the algorithm remains locally convergent. A two-step estimation, with a weighting matrix for which Assumption 2 holds in the first step, would provide a valid estimation strategy in that case. Assumption 2 is invariant to some one-to-one reparameterizations, this is shown in the next section.

Under Assumption 2, the parameters are both locally and globally identified (i.e. Assumption 1 (iv) and (i)). Conversely, Assumption 1 (iv) implies that Assumption 2 (a) holds locally around $\theta^\dagger$. The condition requires that it holds globally rather than locally.[4] Under Assumptions 1 and 2, a sample analog of Assumption 2 holds for the following quantities:

$$\overline{G}_n(\theta) = \int_0^1 G_n(\omega\theta + (1-\omega)\hat{\theta}_n)d\omega, \quad \overline{G}_n(\theta_1, \theta_2) = \int_0^1 G_n(\omega\theta_1 + (1-\omega)\theta_2)d\omega,$$

[1]The factor $\rho$ is assumed to be set, without loss of generality, such that $\sigma_{\min}[\overline{G}(\theta)] > \underline{\sigma}$ under (a) and $\|\overline{G}(\theta)(\theta - \theta^\dagger)\| > \underline{\sigma}\|\theta - \theta^\dagger\|$ under (b) for $\underline{\sigma}$ found in Assumption 1.

[2]See Nesterov (2018, pp33-35), especially equations (1.2.25), (1.2.27) and Theorem 1.2.4 for GD.

[3]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.

[4]See Lemmas A4, A5 and Propositions 1, 5.

with probability approaching 1, this is shown in Lemma A6. When Assumption 2 cannot be verified analytically, a related condition which does not involve the minimizer can be checked numerically on the sample moments and their Jacobian. This is considered in Section 4.1.

**Lemma 1.** *Suppose Assumptions 1 and 2 hold, then* $\hat{\theta}_n \xrightarrow{p} \theta^\dagger$ *and* $Q_n(\hat{\theta}_n) \xrightarrow{p} Q(\theta^\dagger)$.

Lemma 1 shows that, although the parameter space is unbounded and $Q_n$ is non-convex, $\hat{\theta}_n$ is a consistent estimator under Assumptions 1 and 2.

**Assumption 3.** *With probability approaching 1:* $P_{k,n}$ *is symmetric and such that:* $0 < \underline{\lambda}_P \leq \lambda_{\min}(P_{k,n}) \leq \lambda_{\max}(P_{k,n}) \leq \overline{\lambda}_P < \infty$.

Assumption 3 requires $P_{k,n}$ to be finite and strictly positive definite. This is always the case for GD since $P_{k,n} = I_d$, and holds for GN under Assumption 2 (a). If the moments only satisfy Assumption 2 (b), Assumption 3 does not necessarily hold for GN since the Jacobian $G_n(\theta_k)$ can be singular, but it remains valid for GD. When Assumption 3 fails, one approach is to regularize the inverse using the so-called Levenberg-Marquardt (LM) algorithm to GN by setting $P_{k,n} = (G_n(\theta_k)'W_nG_n(\theta_k) + \lambda I_d)^{-1}$ so that $\overline{\lambda}_P < \lambda^{-1} < \infty$ and $P_{k,n}$ is finite. Note that Assumption 3 does hold for GN under strong injectivity conditions introduced in the next Section. Nocedal and Wright (2006, Ch3.4) list several additional approaches to enforce Assumption 3, mainly for convex optimizers.

## 2.2 Global Convergence Results

The following provides the main results: the global convergence properties of gradient-based algorithms. In the following, the initial value $\theta_0$ is taken from $\Theta$, a compact subset of $\mathbb{R}^{d_\theta}$. This is a technical assumption; although the optimization is unconstrained, the sample moments are not uniformly consistent on $\mathbb{R}^{d_\theta}$ which complicates the analysis. The following shows global convergence, uniformly over $\theta_0 \in \Theta$. The main idea is to show that, with probability approaching 1, the optimization path $(\theta_k)_{k\geq 0}$ is restricted to a compact set, determined by $\theta_0$, where the sample moments are uniformly consistent. Without loss of generality, $\Theta$ is assumed convex and large enough that $\theta^\dagger \in \text{interior}(\Theta)$. In addition, local convergence results can be found in Appendix C, those results are new in the case of overidentified and misspecified models as they allow for $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \neq 0$.

**Theorem 1** (Correctly Specified)**.** *Suppose Assumptions 1, 2, 3 hold and* $Q(\theta^\dagger) = 0$*. Then, for* $\gamma$ *small enough, there exists* $\overline{\gamma} \in (0, 1)$*,* $0 < \underline{\lambda} \leq \overline{\lambda} < +\infty$*, and* $C \geq 0$ *such that:*

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1 - \overline{\gamma})^{k+1} \frac{\sqrt{\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}}{\sqrt{\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}} \|\theta_0 - \hat{\theta}_n\|,$$

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

Theorem 1 provides global convergence results that are comparable to the convex case. Because the factor $(1 - \overline{\gamma})$ is less than 1, the distance to the solution $\|\theta_{k+1} - \hat{\theta}_n\|$ decreases exponentially fast with $k$, as in the convex case. Several factors affect convergence. The constants $\underline{\lambda}$, $\overline{\lambda}$ coincide with $C_2 = 1/2\rho^2\underline{\sigma}^2/[\overline{\sigma}^2\overline{\lambda}_W]$, $C_3 = 1/2\overline{\sigma}^2\overline{\lambda}_W$ in Proposition 1 below. The convergence rate $1 - \overline{\gamma}$ depends on $C_1 = 1/2\rho^2\underline{\sigma}^2/[\overline{\sigma}^2\overline{\lambda}_W]$, from the same Proposition.

Through these constants, it appears that identification strength - here measured by $\rho\underline{\sigma}$ - and the choice of weighting matrix $W_n$ affect the convergence properties. In particular, a weighting matrix that is ill-conditioned can lead to slower convergence. This can make optimization challenging. When the sample moments are highly correlated, the optimal weighting matrix can be ill-conditioned. Using equal weighting, a diagonal weighting matrix, or regularizing the optimal weighting matrix with $W_n = (\hat{V}_n + \lambda I_d)^{-1}$, where $\hat{V}_n$ estimates the variance of $\sqrt{n}\overline{g}_n(\theta^\dagger)$, could improve numerical stability.

The size of $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ further affects convergence. The constant $C$ coincides with $C_4 = \overline{\lambda}_W^{1/2} L$ in Proposition 5 below. The constant $L$ measures the non-linearity of the sample moments, $L = 0$ corresponds to linear models. For linear models, $C = 0$ implies that $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ does not affect convergence. Non-linear models have $L > 0$ which makes optimization more sensitive to $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ for overidentified models.

In applications, $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ can be relatively large so that misspecification becomes a concern. Understanding the robustness of Theorem 1 to non-negligible deviations from $Q(\theta^\dagger) = 0$ is then empirically relevant. The following considers models where the quantity:

$$Q_n(\hat{\theta}_n) \xrightarrow{p} Q(\theta^\dagger) := \varphi/2 > 0$$

does not vanish asymptotically which implies that $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ matters for convergence, even in large samples. Since $G_n$ cannot be full rank at $\theta = \hat{\theta}_n$ when the model is both just-

identified and misspecified, the results presented here solely consider over-identified models.[5]

**Theorem 2** (Misspecified)**.** *Suppose Assumptions 1, 2, 3 hold and* $Q(\theta^\dagger) = \varphi/2 > 0$*, such that:*

$$\sqrt{\varphi} < \min\left(\frac{\rho\underline{\sigma}}{\sqrt{\overline{\lambda}_W}L}, \frac{1}{2}\frac{\rho^2\underline{\sigma}^2}{\overline{\lambda}_W^{3/2}\overline{\sigma}^2 L}\right), \tag{2}$$

*then, for* $\gamma$ *small enough, there exists* $\overline{\gamma} \in (0,1)$*,* $0 < \underline{\lambda} \leq \overline{\lambda} < +\infty$ *and* $C > 0$ *such that* $\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \xrightarrow{p} \underline{\lambda} - C\sqrt{\varphi} > 0$*, and:*

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})^{k+1}\frac{\sqrt{\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}}{\sqrt{\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}}\|\theta_0 - \hat{\theta}_n\|,$$

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

Theorem 2 shows that convergence is robust to 'moderate' amounts of misspecification. For linear models, $L = 0$ implies that (2) reads $\varphi < +\infty$, which is not restrictive. In (2), the choice of $W_n$, nonlinearity, and identification strength restrict the amount of misspecification allowed in (2). The restrictions (2) are discussed further with Proposition 5 below. The convergence rate $1-\overline{\gamma}$ also depends on $\varphi$, which slows convergence. In the limit, its expression is given by $(1-\overline{\gamma})^2 = 1-\gamma\underline{\lambda}_P C_1/2$, where $C_1 = (\rho\underline{\sigma} - \overline{\lambda}_W^{1/2}L\sqrt{\varphi})^2/[C_3 + C_4\sqrt{\varphi}]$. The constants $C_3$, $C_4$ appear in Proposition 5 below. The first of the two terms in the upper bound in (2) ensures that $\overline{\gamma} > 0$ is feasible. Having $\varphi \neq 0$ makes convergence slower and estimation more challenging. When $\varphi$ is arbitrarily large, global convergence can fail. This is explained in the next Section, and illustrated with an MA(1) example. Since the magnitude of $\varphi$ depends on the choice of moments $\overline{g}_n$ and weighting matrix $W_n$, a careful selection of these two might mitigate this issue.

# 3 Assumption 2 and its relation to the literature

**Convexity, monotonicity and the Polyak-Łojasiewicz condition.** The following briefly reviews some convexity conditions found in the literature and an important relaxation called the Polyak-Łojasiewicz (PL) condition. The latter has gathered much attention

[5]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.

in the machine learning literature in recent years. Because Assumption 2 is stated on population quantities, the following discussion will focus on $Q$.

For general minimization of an objective $Q$, GD, NR and QN are globally convergent for $\theta^\dagger$ if $Q$ is $\mu$-*strongly convex*, i.e. if for some $\mu > 0$:

$$Q(\theta_2) \geq Q(\theta_1) + \partial_\theta Q(\theta_1)(\theta_2 - \theta_1) + \frac{\mu}{2}\|\theta_1 - \theta_2\|^2,$$

for all $\theta_1, \theta_2 \in \mathbb{R}^{d_\theta}$. When $Q$ is twice continuously differentiable it is strongly convex if its Hessian $H(\theta) = \partial^2_{\theta,\theta'} Q(\theta)$ is strictly positive definite everywhere with $0 < \underline{\lambda}_H < \lambda_{\min}[H(\theta)] \leq \lambda_{\max}[H(\theta)] < \overline{\lambda}_H < \infty$. Under strong convexity, for $\gamma > 0$ sufficiently small and any $\theta_0$:

$$Q(\theta_{k+1}) - Q(\theta^\dagger) \leq (1 - \eta)\left(Q(\theta_k) - Q(\theta^\dagger)\right),$$

for some $\eta \in (0, 1)$ which depends on $\gamma$, the choice of algorithm, i.e. $P_{k,n}$, and the eigenvalues of $H$. Iterating on this inequality indicates that the fit improves rapidly from any starting value $\theta_0$: $Q(\theta_k) - Q(\theta^\dagger) \leq (1-\eta)^k \left(Q(\theta_0) - Q(\theta^\dagger)\right)$. Under strong convexity, $Q$ has a unique global minimizer and no local optima. The literature has considered a number of relaxations of strong convexity under which GD is globally convergent. This includes the so-called *star convexity* condition introduced by Nesterov and Polyak (2006):

$$Q(\theta^\dagger) \geq Q(\theta) + \lambda \partial_\theta Q(\theta)(\theta^\dagger - \theta) + \frac{\mu}{2}\|\theta - \theta^\dagger\|^2$$

for some $\mu \geq 0$ and $\lambda = 1$. Fast convergence results for $\theta$ require $\mu > 0$. This is similar-looking to strong convexity but only involves the pairs $(\theta_1, \theta_2) = (\theta, \theta^\dagger)$. For these functions, the convexity property only holds on line segments toward $\theta^\dagger$. Star convexity implies that $\theta^\dagger$ is the unique global minimizer of $Q$. This condition can be further weakened to *quasar convexity*, which allows for $\lambda > 1$ in the inequality above. Hinder et al. (2020), Figure 1, plot several functions that satisfy these conditions.

Karimi et al. (2016), Guminov et al. (2017) showed that a number of relaxations of strong convexity imply the so-called *Polyak-Łojasiewicz* (PL) inequality, named after Polyak (1963) and Łojasiewicz (1963), which requires that:

$$\|\partial_\theta Q(\theta)\|^2 \geq \mu\left(Q(\theta) - Q(\theta^\dagger)\right), \tag{PL}$$

for all $\theta \in \mathbb{R}^{d_\theta}$ and some $\mu > 0$. When $Q$ satisfies the PL inequality, $\partial_\theta Q(\theta) = 0$ implies $\theta$ is globally optimal, i.e. $Q(\theta) = Q(\theta^\dagger)$. The arg-minimizer may not be unique, however,

unlike strong convexity. If the PL inequality holds and $\partial_\theta Q$ is Lipschitz continuous, it can be shown that for $\gamma > 0$ small enough: $Q(\theta_{k+1}) - Q(\theta^\dagger) \leq (1-\eta)\left(Q(\theta_k) - Q(\theta^\dagger)\right)$ for GD (Karimi et al., 2016, Th1). This does not imply that $\theta_{k+1}$ converges to $\theta^\dagger$, however, unless the arg-minimizer is unique. Because strong convexity implies the PL inequality, Karimi et al. (2016) argue that the latter holds locally over a larger area than strong convexity, predicting better optimization performance. They also note that it is difficult to characterize which functions satisfy the PL inequality. They show that $Q(\theta) = h(A\theta)$, with $h$ strongly convex and $A$ a non-zero matrix, satisfies the PL inequality.

Closely related to the GMM setting, a smaller literature has considered conditions for solving non-linear systems of equations of the form: $g(\theta) = 0$, typically with $g$ and $\theta$ of the same dimension. An important reference is Dennis and Schnabel (1996), who cast the problem as minimizing $Q(\theta) = \|g(\theta)\|^2$, similar to GMM, and derive global convergence results to a local minimum under convexity conditions (Theorems 6.3.3-6.3.4). Deuflhard (2005, Ch3) studies global convergence under alternative conditions. For just and under-determined systems, several authors considered a *strong monotonicity condition*:

$$(g(\theta_1) - g(\theta_2))'(\theta_1 - \theta_2) \geq \mu\|\theta_1 - \theta_2\|^2,$$

with $\mu > 0$, e.g. Solodov and Svaiter (2000), Polyak and Tremba (2020). Note that when $g = \partial_\theta F$, then $g$ is strongly monotone if, and only if, $F$ is strongly convex. Hence, global convergence under strong monotonicity is related to global convergence under strong convexity of $F$. In that case, $g$ is said to be cyclically monotone (Rockafellar, 2015, p238). These results do not consider $g(\theta^\dagger) \neq 0$ which is particularly relevant here.

A companion paper, Forneron (2023), considers correctly specified GMM estimation with non-smooth sample moments that may not satisfy Assumption 2. There are two important differences in that setting: 1) the Jacobian $G_n$ is not defined, and 2) $Q_n$ can have local optima. The methods considered here are not sufficient to find a global optimum, and there is a curse of dimensionality for global convergence. The two papers are complementary.

**Relation between the different conditions.** Narrowing to the GMM setting specifically, the following shows that the PL inequality holds *in the population for correctly specified models* under Assumption 2. A related result is derived under misspecification.

As discussed above, Assumption 2 (a) implies Assumption 2 (b). The latter confers most of the properties required for minimizing $Q$. It can be useful to re-write the condition in terms of $g$: Assumption 2 (b) $\|G(\theta)'W[g(\theta) - g(\theta^\dagger)]\| > \rho\underline{\sigma}\|\theta - \theta^\dagger\|$. For correctly specified

models, $g(\theta^\dagger) = 0$ implies $G(\theta)'Wg(\theta) = \partial_\theta Q(\theta)$. The only critical point is $\theta = \theta^\dagger$. Hence, Assumption 2 excludes local optima and saddle points when the model is correctly specified.[6]

**Proposition 1** (Correct Specification)**.** *Suppose Assumptions 1 (ii), (iii), (vi), 2 (b) hold and $Q(\theta^\dagger) = 0$, then there exists strictly positive constants $C_1, C_2, C_3$ such that for all $\theta \in \mathbb{R}^{d_\theta}$:*

$$(1) \quad \|\partial_\theta Q(\theta)\|^2 \geq C_1 \left(Q(\theta) - Q(\theta^\dagger)\right)$$
$$(2) \quad C_2\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger) \leq C_3\|\theta - \theta^\dagger\|^2.$$

Proposition 1 shows that Assumption 2 (b), together with bounds on $W$ and Lipschitz continuity of $G$ imply the PL inequality (1) for $Q$. In addition, (2) implies global identification and is needed to derive the convergence rate of $\theta_k$. Strong convexity also implies (1) and (2).

**Proposition 2.** *Suppose $W$ is invertible, $Q(\theta^\dagger) = 0$. 1) If $Q$ satisfies the PL inequality with $\mu > 0$ and $C_2\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger)$ for $C_2 > 0$ and all $\theta \in \mathbb{R}^{d_\theta}$, then Assumption 2 (b) holds. 2) If $Q$ is quasar-convex with $\mu > 0$, then Assumption 2 (b) holds.*

Proposition 2 gives a condition under which quasar-convexity and the PL inequality imply Assumption 2 (b). On compact sets, Assumption 1 (i), (iii), (iv) together imply a $C_2 > 0$ exists for correctly specified models. Assumption 2 (b) does not imply quasar-convexity.[7] The following considers strong monotonicity and introduces a *strong injectivity* condition:

$$\|g(\theta_1) - g(\theta_2)\| \geq \mu\|\theta_1 - \theta_2\|. \tag{SI}$$

It can be shown that the strong injectivity property holds on compact convex sets under the Gale-Nikaidô-Fisher-Rothenberg global identification conditions: $\det(G(\theta)) > 0$ and $G(\theta)$ positive quasi-definite, for all $\theta \in \mathbb{R}^{d_\theta}$, where det is the determinant.[8]

**Proposition 3** (Just-Identified)**.** *1) If $Ag$ is strongly monotone for some invertible matrix $A$ and $\mu > 0$, then Assumption 2 (b) holds. 2) If $g$ is strongly injective with $\mu > 0$, then Assumption 2 (b) holds.*

[6] A critical point is a $\theta$ such that $\partial_\theta Q(\theta) = 0$. Assuming $Q$ is twice differentiable, it is a local minimum if $\partial^2_{\theta,\theta'}Q(\theta)$ is positive semidefinite, maximum if $\partial^2_{\theta,\theta'}Q(\theta)$ is negative semidefinite, and a saddle point if $\partial^2_{\theta,\theta'}Q(\theta)$ is indefinite, i.e. has both positive and negative eigenvalues.

[7] Quasar-convexity implies $(\theta-\theta^\dagger)'G(\theta)'W\overline{G}(\theta)(\theta-\theta^\dagger) \geq \frac{\mu}{2\lambda}\|\theta-\theta^\dagger\|^2$ for correctly specified models. This is more restrictive than Assumption 2 (b) when $\theta$ is not scalar.

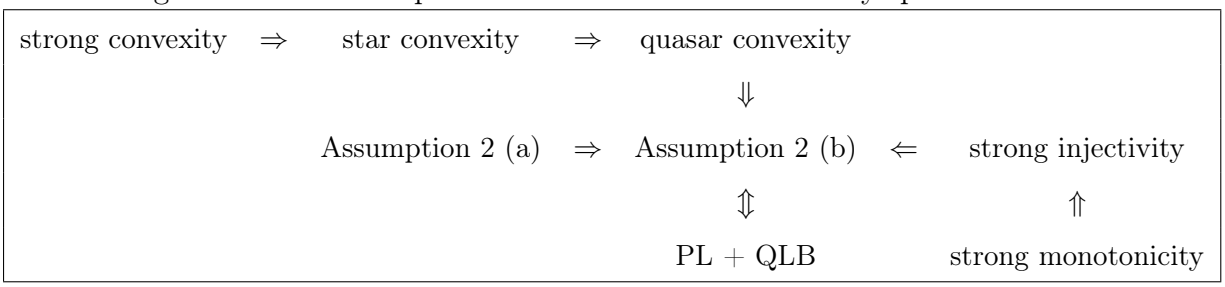
[8] $G$ is positive quasi-definite if, and only if, $G + G'$ is positive definite. See Fisher (1966), Rothenberg (1971); and Komunjer (2012) for a discussion and alternative conditions.

For over-identified models, (SI) is not sufficient. As discussed above, the weighting matrix $W$ plays a role in the convergence properties. The following extends (SI) appropriately:

$$\|G(\theta_1)'W[g(\theta_1) - g(\theta_2)]\| \geq \mu\|\theta_1 - \theta_2\|. \tag{SI'}$$

Relative to (SI), the additional term ensures that $g(\cdot)$ is one-to-one in the row space of $G(\cdot)'W$. Taking $(\theta_1, \theta_2) = (\theta, \theta^\dagger)$ yields Assumption 2 (b). Note that (SI') implies that Assumption 3 holds for GN (Lemma B7). (SI') is more challenging to verify than (SI) as it involves the weighting matrix $W$ and the Jacobian $G$. If (SI) holds for a just-identified subset of moments, then it is possible to regularize $W$ so that (SI') holds (Lemma B8).

Figure 1: Relationship between conditions for correctly specified models




**Legend:** Relations hold when $Q(\theta^\dagger) = 0$. QLB = Quadratic Lower Bound, i.e. $C_2\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger)$ for some $C_2 > 0$. Relation with strong monotonicity is for just-identified models.

Figure 1 summarizes the results of Propositions 1, 2, 3. Since $Q_n(\hat{\theta}_n) = 0$ for just-identified models that are correctly specified, the relationship also applies in the finite samples problems where these conditions are met. When $g$ and $\theta$ are scalar, Assumption 2 implies strict monotonicity, $g$ is either increasing or decreasing, but does not imply convexity of $Q$, however, as the MA example below will illustrate.

It remains to determine if Assumption 2 (b) is minimal for global convergence, or if can be weakened further. The following condition is *necessary* for GD and other gradient-based optimizers of the form (1) to be globally convergent:

$$\partial_\theta Q(\theta) = 0 \Leftrightarrow \theta = \theta^\dagger. \tag{N}$$

The following shows, under regularity conditions, that (N) implies Assumption 2 (b).

**Proposition 4.** *Suppose condition (N) and Assumption 1 (ii)-(iv) and (vi) hold, then Assumption 2 (b) holds on any compact convex set containing $\theta^\dagger$.*

The case of misspecified models is more complicated, as the following shows that the equivalence between the PL inequality and Assumption 2 (b) is not automatic.

**Proposition 5** (Misspecification)**.** *Suppose Assumptions 1 (ii), (iii), (vi), 2 (b) hold, then there exists strictly positive constants $C_2, C_3, C_4$ such that for all $\theta \in \mathbb{R}^{d_\theta}$:*

$$(1) \quad \|\partial_\theta Q(\theta)\| \geq \left(\rho\underline{\sigma} - \sqrt{\varphi}\overline{\lambda}_W^{1/2} L\right) \|\theta - \theta^\dagger\|$$
$$(2) \quad (C_2 - C_4\sqrt{\varphi})\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger) \leq (C_3 + C_4\sqrt{\varphi})\|\theta - \theta^\dagger\|^2,$$

*where $Q(\theta^\dagger) = \varphi > 0$, $C_2, C_3$ are the same as in Proposition 1 and $L$ is the Lipschitz constant of $G$ from in Assumption 1 (iii). If in addition $\rho\underline{\sigma} - \sqrt{\varphi\overline{\lambda}_W} L > 0$, then for all $\theta \in \mathbb{R}^{d_\theta}$:*

$$(1') \quad \|\partial_\theta Q(\theta)\|^2 \geq \frac{(\rho\underline{\sigma} - \sqrt{\varphi\overline{\lambda}_W} L)^2}{C_3 + C_4\sqrt{\varphi}} \left(Q(\theta) - Q(\theta^\dagger)\right).$$

Proposition 5 (1) is only informative when the amount of misspecification is moderate, i.e. $\varphi < \rho^2\underline{\sigma}^2/[\overline{\lambda}_W L^2]$. When this holds, there are no local optima besides $\theta^\dagger$. It also implies the PL inequality (1') holds. To recover convergence for $\theta$, the lower bound in (2) should be informative which further requires $\sqrt{\varphi} < C_2/C_4$.[9] The degree of non-linearity - measured by $L$ - and the choice of weighting matrix - measured by $\overline{\lambda}_W, \underline{\lambda}_W$ and $\varphi$ - constrain the amount of misspecification permitted to get informative bounds. For correctly specified models, $Q_n(\hat{\theta}_n) = o_p(1)$ implies that both (1') and (2) hold asymptotically.

**Further characterization of Assumption 2 (Just-Identified).** Like star-convexity, Assumption 2 is stated relative to the unknown $\theta^\dagger$. The following Proposition gives several conditions under which Assumption 2 (a) holds and properties implied by these conditions.

**Proposition 6.** *(Sufficient Conditions) Consider the following conditions:*
*(a) $\sigma_{\min}[\overline{G}(\theta_1, \theta_2)] > \underline{\sigma} > 0$, for all $\theta_1, \theta_2 \in \mathbb{R}^{d_\theta}$, (b) for all $\theta \in \mathbb{R}^{d_\theta}$, $G(\theta) = US(\theta)V$ for $U, V$ invertible and $S(\theta)$ symmetric with $0 < \underline{\lambda}_S < \lambda_{\min}[S(\theta)] < \overline{\lambda}_S < \infty$, for all $\theta \in \mathbb{R}^{d_\theta}$, (c) $g(\theta) = \partial_\theta F(\theta)$, for all $\theta \in \mathbb{R}^{d_\theta}$, where $F : \mathbb{R}^{d_\theta} \to \mathbb{R}$ is twice continuously differentiable, strongly convex.*
*The following holds: (1) (c) $\Rightarrow$ (b) $\Rightarrow$ (a) $\Rightarrow$ Assumption 2 (a) holds; (2) (a) implies $g(\cdot)$ is one-to-one; (3) if (a) holds, there exists a reparameterization $h(\cdot) = \psi \circ g \circ \phi(\cdot)$ with $\phi$ one-to-one and $\psi$ affine, such that $1/2h(\theta)'Wh(\theta)$ is strongly convex.*

Condition (a) does not require knowledge of $\theta^\dagger$ and implies that $g(\cdot)$ is one-to-one. The

[9] The derivations give the following bounds $C_2 = 1/2\frac{\rho^2\underline{\sigma}^2}{\overline{\sigma}^2\overline{\lambda}_W}$ and $C_4 = \overline{\lambda}_W^{1/2} L$ so that the condition reads $\sqrt{\varphi} < 1/2\rho^2\underline{\sigma}^2[\overline{\sigma}^2\overline{\lambda}_W^{3/2} L]^{-1}$. It is possible to relax this condition at the cost of more complicated derivations using a combination of global and local convergence arguments.

latter is often assumed for indirect inference.[10] Condition (a) also implies (SI) with $\mu = \underline{\sigma}$. When the Jacobian can be linearly rearranged into a symmetric positive definite matrix $S(\theta) = U^{-1}G(\theta)V^{-1}$, then condition (a) holds. These problems can be thought of as *implicitly convex* in the special case where where $S$ is the second derivative of a convex function. For a given $\theta \in \mathbb{R}^{d_\theta}$, the decomposition (b) always exists: the singular value decomposition gives $G(\theta) = U(\theta)S(\theta)V(\theta)$ where $U(\theta), V(\theta)$ are unitary and $S(\theta)$ is diagonal with positive entries. A lesser known result, due to Frobenius (1910) shows that any square matrix can be written as the product of two real symmetric matrices; here $G(\theta) = S_1(\theta)S_2(\theta)$. The Jordan normal form of $G(\theta)$ can be used to compute this factorization (Bosch, 1986). If $G(\theta)$ is invertible, for all $\theta \in \mathbb{R}^{d_\theta}$, and $U, V$ or one of $S_1, S_2$ do not vary with $\theta$, in the singular value or Frobenius decomposition, then (b) holds. Under condition (c), $g$ is cyclically , and thus strongly, monotone.

**Proposition 7.** *(Reparameterization) Take $h : \mathcal{U} \to \mathbb{R}^{d_\theta}$, one-to-one, continuously differentiable on $\mathcal{U}$, a convex set, with $0 < \underline{\sigma}_h \leq \min_{u\in\mathcal{U}} \sigma_{\min}[\partial_u h(u)] \leq \max_{u\in\mathcal{U}} \sigma_{\max}[\partial_u h(u)] \leq \overline{\sigma}_h < \infty$. Let $u^\dagger = h^{-1}(\theta^\dagger)$, the minimizer of $Q \circ h$.*
*1) Suppose Assumption 2 (a) holds for $g$, let:*

$$L_{1,h} = \sup_{u\in\mathcal{U}} \|\partial_u h(u) - \partial_u h(u^\dagger)\|,\ L_{2,h} = \sup_{u\in\mathcal{U},\omega\in[0,1]} \|h(\omega u + (1-\omega)u^\dagger) - \omega h(u) - (1-\omega)h(u^\dagger)\|.$$

*If $\underline{\sigma} > [L_{1,h}\overline{\sigma} + L_{2,h}L\overline{\sigma}_h]/\underline{\sigma}_h$, where $L$ is the Lipschitz constant of $G$, then Assumption 2 (a) holds for $g \circ h$. In particular, if $h = Au + b$ is affine with $A$ invertible then $L_{1,h} = L_{2,h} = 0$ and Assumption 2 (a) holds for $g \circ h$.*
*2) Suppose Assumption 2 (b) holds for $g$. If $\|h(u) - h(u^\dagger)\| \geq \mu\|u - u^\dagger\|$, for some $\mu > 0$ and all $u \in \mathcal{U}$, then Assumption 2 (b) holds for $g \circ h$.*

Strong convexity is preserved by affine transformations and reparameterization that satisfy particular component-wise monotonicity constraints on the reparameterization (e.g. Boyd and Vandenberghe, 2004, Sec3.2). Proposition 7 shows that Assumption 2 (a) is also preserved by affine transformations and moderately non-linear reparameterizations $h$. Hence, under Assumption 2 (a), optimization should be locally robust to the choice of parameterization. Assumption 2 (b) is preserved if $h$ is strongly injective, a mild requirement. In particular, invertible affine transformations preserve Assumption 2 (b). Similar statements for overidentified models can be found in Propositions B8, B9. Propositions 2 and 7 together

[10] See e.g. Gourieroux et al. (1993), Assumption (A4).

imply that if $Q$ is strongly convex for a particular parameterization, e.g. reduced-form coefficients, then Assumption 2 (b) holds for $Q \circ h$ where $h$ satisfies the conditions above, where $h$ is the mapping from reduced form to structural coefficients.

# 4 Recommendations for Practice

## 4.1 Checking whether Assumption 2 holds

The global convergence results hinge on Assumption 2 (b) as it confers the objective several key properties for optimization. In some cases it may be feasible to verify analytically that one of the conditions in Figure 1 or Proposition 6 hold. For some models, it is possible to construct moments that identify the parameters, typically using injectivity arguments. In that case, (SI) holds which implies Assumption 2 (b) holds, under regularity conditions.

For more complex models, it may only be possible to evaluate numerically over a representative set of points, whether one of these conditions is likely to holds, or not. Since Assumption 2, and its sample counterpart Assumption A2, depend on the unknown minimizer or $Q$, resp. $Q_n$, it is not possible to check the conditions numerically before the performing the estimation. It is possible to check a stronger condition which does not take an estimate $\hat{\theta}_n$ as input, however.

In the main results, the constant $\rho\underline{\sigma}$ can be arbitrarily small. In practice, however, when $\rho\underline{\sigma} \to 0$ the convergence rate $(1 - \overline{\gamma}) \to 1$ is arbitrarily slow. The following approximates an upper bound for $(1 - \overline{\gamma})$, assuming correct specification, and the corresponding number of iterations $\underline{k}$ required to achieve $Q_n(\theta_{\underline{k}}) - Q_n(\hat{\theta}_n) \leq \varepsilon[Q_n(\theta_0) - Q_n(\hat{\theta}_n)]$ for a user-chosen $\varepsilon \in (0, 1)$. In practice, these bounds can be very conservative. The value $\underline{k}$ mainly indicates whether global convergence is practically feasible (e.g. $\underline{k} \leq 10^3$) or not (e.g. $\underline{k} \geq 10^{12}$).

When $\hat{\theta}_n$ is unknown, before the estimation is performed, it is only possible to verify a stronger condition. The following considers a sample analog of (SI'), introduced above:

$$\|G_n(\theta_1)' W_n[\overline{g}_n(\theta_1) - \overline{g}_n(\theta_2)]\| \geq \mu_n \|\theta_1 - \theta_2\|, \tag{SI'}$$

for some $\mu_n > 0$. In the following, the finite grid of pairs $\Theta_K = \{(\theta_1^1, \theta_1^2), \ldots, (\theta_K^1, \theta_K^2)\}$ will be used for that purpose. It construction is discussed in more detail below. Suppose $\theta_k^1 \neq \theta_k^2$

for each $k$, compute:

$$\mu_k = \frac{\|P_{k,n} G_n(\theta_k^1)' W_n [\overline{g}_n(\theta_k^1) - \overline{g}_n(\theta_k^2)]\|}{\|\theta_k^1 - \theta_k^2\|}, \quad C_{3,k} = \frac{\|\theta_k^1 - \theta_k^2\|}{\|\overline{g}_n(\theta_k^1) - \overline{g}_n(\theta_k^2)\|_{W_n}},$$
$$L_{Q,P,k} = \frac{\|P_{k,n} H_n(\theta_k^1)(\theta_k^1 - \theta_k^2)\|}{\|\theta_k^1 - \theta_k^2\|},$$

where $P_{k,n}$ is computed using $\theta_k^1$ and the algorithm of choice.[11] Then compute:

$$(1 - \overline{\gamma})^2 = \max\left(0, 1 - [\hat{\mu}_n \hat{C}_{3,n}]^2 / [4 \hat{L}_{Q,P,n}]\right), \quad \underline{k} \geq \frac{\log(\varepsilon)}{\log(1 - \overline{\gamma})},$$

where $\hat{\mu}_n = \min_k \mu_k$, $\hat{C}_{3,n} = \min_k C_{3,k}$, and $\hat{L}_{Q,P,n} = \max_k L_{Q,P,k}$. The normalization using $P_{k,n}$ ensures that these values are invariant to linear reparameterizations of the parameters and/or moments for GN or NR. As a reference, with the normalization linear models have $\hat{\mu}_n = 1$ under the standard rank condition. To ensure the product $P_{k,n} G_n(\theta_k^1)$ is well behaved, it is recommended to compute a pseudo-inverse of $G_n(\theta_k^1)' W_n G_n(\theta_k^1)$ in the case of GN. This yields $\mu_k = 0$ and $\overline{\gamma} = 0$ when $G_n(\theta_k^1)$ is numerically close to singular in the relevant direction. If the conditions fail or the bounds indicate that convergence is not practically feasible, typically when $\hat{\mu}_n < 10^{-2}$ for GN,[12] gradient-based methods need to be modified to ensure global convergence, using multiple starting values or a hybrid approach with theoretical guarantees, see Forneron (2023) for an explicit algorithm in that setting.

**Constructing $\Theta_K$.** Take a set $\Theta$ large enough that $\hat{\theta}_n \in \Theta$ is likely. The grid $\Theta_K$ should be dense in $\Theta$ so that, as the number of points $K$ increases, any $\theta \in \Theta$ is arbitrarily close to some value in the grid. For $\Theta = [0,1]^{d_\theta}$, the Sobol and Halton sequences have this property, and are readily available in statistical software (R, Matlab, Python, Julia). In general, when $\Theta = [\underline{\theta}_1, \overline{\theta}_1] \times \cdots \times [\underline{\theta}_{d_\theta}, \overline{\theta}_{d_\theta}]$, where $\underline{\theta}_1, \overline{\theta}_1$ denote lower (resp. upper) bounds on each coefficient, a sequence can be constructed from the Sobol or Halton sequence, denoted $(\vartheta_{i,k})$, $i = 1, \ldots, d_\theta$, $k = 1, \ldots, K$, by setting $\theta_{i,k} = \underline{\theta}_i + (\overline{\theta}_i - \underline{\theta}_i)\vartheta_{i,k}$.

[11] Note that $L_{Q,P,k}$ involves a Hessian-vector product, which can be computed using only gradients: $H_n(\theta_k^1)(\theta_k^1 - \theta_k^2) \simeq [\partial_\theta Q_n(\theta_k^1 + \epsilon[\theta_k^2 - \theta_k^1]) - \partial_\theta Q_n(\theta_k^1)]/\epsilon$, for $\epsilon$ small.

[12] Dividing $\mu_n$ by 10 approximately reduces $\overline{\gamma}$ by a factor of 100, by a local expansion argument. Convergence becomes significantly slower when $\hat{\mu}_n$ approaches 0.

### 4.2 Iteration dependent choice of learning rate $\gamma_k$

The results are stated for a fixed learning rate. In practice, adaptive choices of $\gamma_k$ are common, using a line search for instance. If the adaptive algorithm is tuned to satisfy the requirements for global convergence, then it is also globally convergent. To preserve convergence properties, additional tuning parameters are typically involved (Nocedal and Wright, 2006, Ch3.1). A backtracking line search, a simple and popular way to set the learning rate (Nocedal and Wright, 2006, Ch3.1), is used as a benchmark comparison for the fixed learning rate used in the applications.

**Algorithm 1:** Backtracking Line Search for Gauss-Newton

**Tuning Parameters:** Initial $\gamma_{\text{init}}$, $\rho \in (0,1)$, $c \in (0,1)$.
**Inputs** : Previous iterate $\theta_k$, moments $\overline{g}_n(\theta_k)$, Jacobian $G_n(\theta_k)$
**Compute** : Search direction: $p_k = (G_n(\theta_k)'W_nG_n(\theta_k))^{-1}G_n(\theta_k)'W_n\overline{g}_n(\theta_k)$, $J_k = G_n(\theta_k)'W_n\overline{g}_n(\theta_k)$.
**Set** : $\gamma_k = \gamma_{\text{init}}$ and $\theta_{k+1} = \theta_k - \gamma_k p_k$
**while** $Q_n(\theta_{k+1}) > Q_n(\theta_k) - c\gamma_k J_k' p_k$ **do**
| **Set** : $\gamma_k = \rho\gamma_k$ and $\theta_{k+1} = \theta_k - \gamma_k p_k$
**end**
**Output** : New iterate $\theta_{k+1}$, Learning Rate $\gamma_k$.

By construction, $J_k'p_k \geq 0$ so that the final $\gamma_k$ decreases the value of the objective function. The while loop terminates once the so-called *Armijo condition* is met:[13] $Q_n(\theta_{k+1}) \leq Q_n(\theta_k) - c\gamma_k J_k'p_k$. For just-identified models, the termination criterion is feasible if $c$ is sufficiently small.[14] Having $\theta_k = \hat{\theta}_n$ implies $p_k = 0$; the condition holds for any $\gamma_k \in (0,1]$. A common choice is $c = 10^{-4}$, $\gamma_{\text{init}} = 1$, $\rho = 0.8$. These were used in all examples.[15]

# 5 Numerical and Empirical Applications

## 5.1 A pen and pencil example: the MA(1) model

The first example illustrates the main results using a simple MA(1) process:

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

[13] See Nocedal and Wright (2006, p33), Nesterov (2018, pp28-29) for discussions.

[14] A sample analog of Proposition 1 implies that $Q_n(\theta_{k+1}) \leq (1-\overline{\gamma})^2 Q_n(\theta_k)$, for any $\theta_k \in \Theta$, when $\gamma \in (0,1)$ small enough for some $\overline{\gamma} \in (0,\gamma)$. Proposition 1 (1)-(2) further imply for just-identified models that $Q_n(\theta_k) - Q_n(\hat{\theta}_n) \leq c_n J_k'p_k$ for some $c_n > 0$. The Armijo condition is feasible if $c$ is small enough.

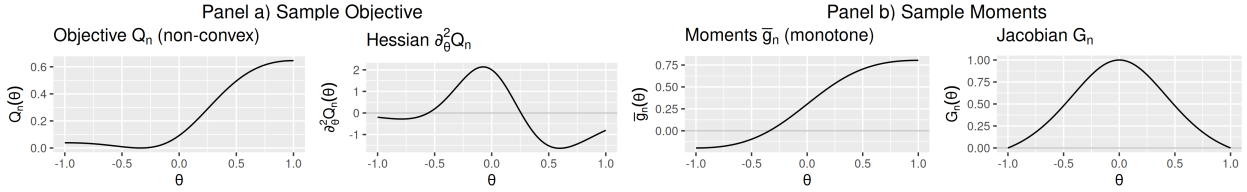
[15] When there are bounds for parameters values, one can set $Q_n(\theta_{k+1}) = +\infty$ if $\theta_{k+1}$ is outside the bounds. Another approach is to project $\theta_{k+1}$ inside the bounds when $\gamma_k$ is too large.

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 2 shows that the Hessian $\partial^2_{\theta,\theta} Q_n(\theta)$ can be positive, negative, or equal to zero depending on the value of $\theta$ – $Q_n$ is non-convex, especially when $\overline{g}_n(\theta)$ is large. Now notice that: $\overline{g}_n(\theta) = \partial_\theta F_n(\theta)$ where $F_n(\theta) = \hat{\beta}_1\theta + \frac{1}{2}\log(1+\theta^2)$, which not a GMM objective but is nevertheless convex on $[-1,1]$, strongly convex on any $[-1+\varepsilon, 1-\varepsilon]$, $\varepsilon \in (0,1)$. Hence, $\overline{g}_n$ is cyclically monotone and statisfies Assumption 2 (a). Note that implicitly, GN minimizes the convex $F_n$ – whereas NR explicitly minimizes the non-convex $Q_n$. This is specific to the just-identified case ($p = 1$), since an $F_n$ cannot be defined in the over-identified case ($p > 1$).

Figure 2: 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 $\theta_0$ is to $\theta^\dagger$. Although $Q_n$ is locally convex around $\hat{\theta}_n$, which is useful for local optimization, the corresponding neighborhood can be fairly small from a practical standpoint.

GN converges steadily from the same $\theta_0$. 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). While here this is not too problematic, the objective function is well defined outside the bounds, this is more concerning in applications where the model cannot be solved outside the bounds – this is illustrated in Section 5.2. A natural solution is to introduce bounds using L-BFGS-B. The search, however, remains

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

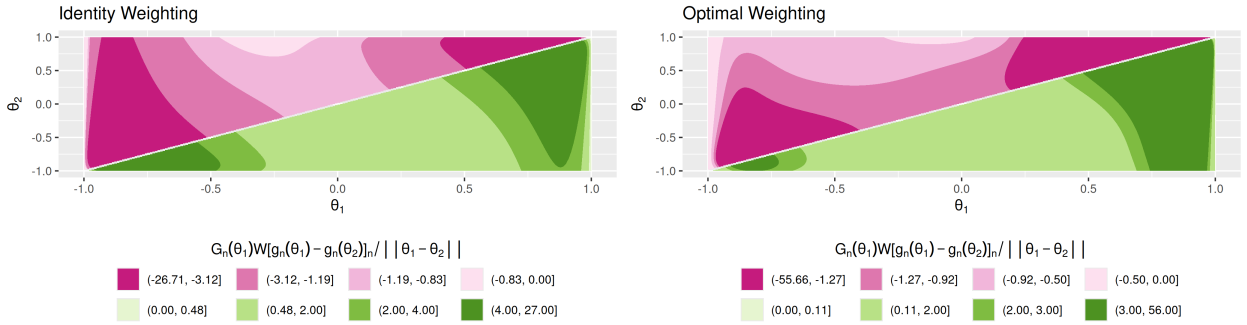
| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | ... | 99 | $Q_n(\theta_{99})$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $p = 1$ | | | | | | | | | | |
| NR | -0.600 | -0.689 | -0.722 | -0.749 | -0.772 | -0.793 | -0.811 | -0.828 | ... | -0.993 | 0.038 |
| GN | -0.600 | -0.560 | -0.529 | -0.504 | -0.484 | -0.466 | -0.451 | -0.438 | ... | -0.338 | $7 \cdot 10^{-8}$ |
| GN-BACK | -0.600 | -0.202 | -0.326 | -0.338 | -0.338 | -0.338 | -0.338 | -0.338 | ... | -0.338 | $7 \cdot 10^{-8}$ |
| BFGS | -0.600 | -0.505 | 4.425 | -0.307 | -0.359 | -0.338 | -0.337 | -0.337 | ... | -0.337 | $7 \cdot 10^{-8}$ |
| L-BFGS-B | -0.600 | -0.505 | 1.000 | -0.455 | -0.375 | -0.318 | -0.341 | -0.339 | ... | -0.338 | $7 \cdot 10^{-8}$ |
| BFGS$^\star$ | -0.600 | -0.462 | -0.286 | -0.345 | -0.340 | -0.338 | -0.338 | -0.338 | ... | -0.338 | $7 \cdot 10^{-8}$ |
| L-BFGS-B$^\star$ | -0.600 | -0.462 | -0.286 | -0.345 | -0.339 | -0.338 | -0.338 | -0.338 | ... | -0.338 | $7 \cdot 10^{-8}$ |
| | $p = 12$ | | | | | | | | | | |
| NR | 0.950 | 0.956 | 0.961 | 0.965 | 0.969 | 0.972 | 0.975 | 0.978 | ... | 1.000 | 4.786 |
| GN | 0.950 | 0.890 | 0.860 | 0.834 | 0.810 | 0.787 | 0.763 | 0.740 | ... | -0.623 | 0.101 |
| GN-BACK | 0.950 | 0.350 | -0.089 | -0.478 | -0.591 | -0.616 | -0.616 | -0.623 | ... | -0.626 | 0.101 |
| BFGS | 0.950 | -8.290 | -8.279 | -8.267 | -8.256 | -8.244 | -8.233 | -8.221 | ... | -6.979 | 0.397 |
| L-BFGS-B | 0.950 | -1.000 | -1.000 | -1.000 | -1.000 | -1.000 | -1.000 | -1.000 | ... | -1.000 | 1.7 |

**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 $\theta$. $W_n = I_d$. The solutions are $\hat{\theta}_n = -0.339$ ($p = 1$) and $\hat{\theta}_n = -0.626$ ($p = 12$). NR = Newton-Raphson, GN = Gauss-Newton, GN-BACK = Gauss-Newton with backtracking line search (Algorithm 1). 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$^\star$ and L-BFGS-B$^\star$ apply the same optimizers to $F_n$ instead of $Q_n$.

somewhat erratic as seen in the Table. Compare these to BFGS$^\star$ and L-BFGS-B$^\star$ which minimize $F_n$, instead of $Q_n$, using the same *optim*. Like GN, they steadily converge to $\hat{\theta}_n$.

For $p = 12$, the model is over-identified and the conditions are more challenging to check analytically. Figure 3 indicates that the strong injectivity condition (SI') appear to hold, the value is bounded away from zero, except at the boundary. The choice of weighing matrix affects the constant $\mu$ in (SI'), as it appears to be smaller with optimal weighting.

Figure 3: MA(1): illustration of the strong injectivity condition




**Legend:** the color groups are given by quantiles on the positive and negative values, so that the colors represent the same fraction of values on the left and right panels.

Table 2 shows that NR, BFGS and L-BFGS-B all fail to converge from $\theta_0 = 0.95$, a starting

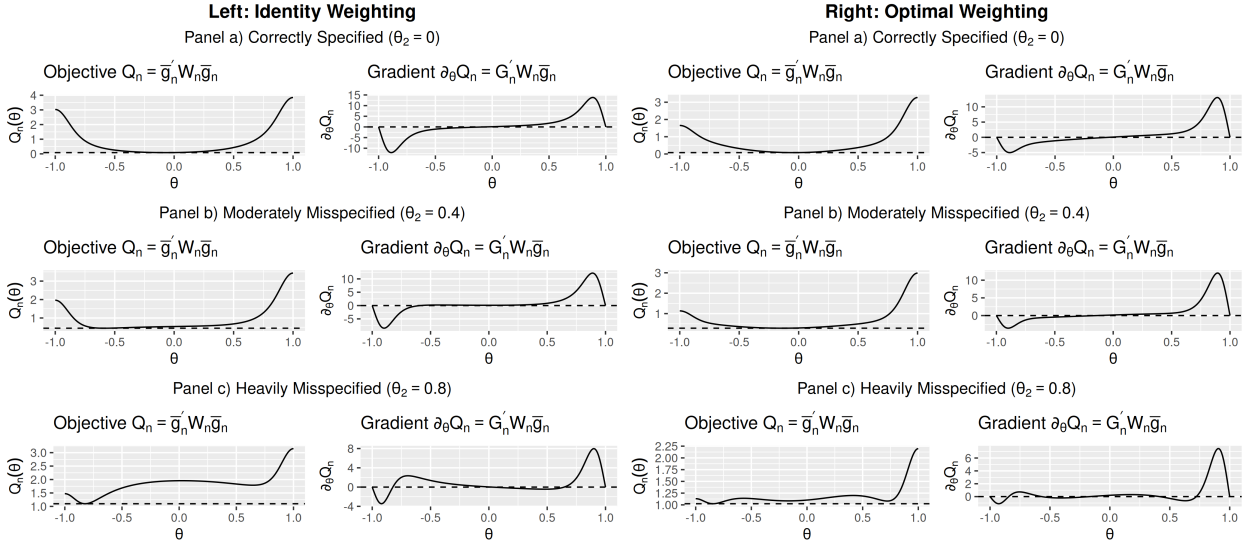
value with negative curvature, with identity weighting.[16] Compare with GN, which steadily converges to $\hat{\theta}_n$. Starting closer to the solution, BFGS and L-BFGS-B also fail to converge using $\theta_0 = 0.6$; GN remains accurate (not reported). R codes can be found in Appendix E.

**The impact of misspecification on estimation.** As discussed in Section 3, when the degree of misspecification $\varphi$ becomes large, local optima may appear and making gradient-based optimizers non-globally convergent. To illustrate this issue, consider the true DGP:

$$y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, \quad e_t \overset{iid}{\sim} \mathcal{N}(0,1), \quad \theta^\dagger \in (-1,1),$$

where $\theta_2 \in \{0, 0.4, 0.8\}$ determines the degree of misspecification. The following sets $\theta_1 = -0.1$ to ensure invertibility as $\theta_2$ varies. The moments are the same as above with $p = 12$.

Figure 4: MA(1): effect of misspecification on $Q_n$




**Legend:** simulated sample of size $n = 200$, $\theta_1 = -0.1$, $p = 12$.

Figure 4 compares $Q_n$ and its gradient $\partial_\theta Q_n$, with identity and optimal weighting, at different degrees of misspecification. On intervals $[-1+\varepsilon, 1-\varepsilon]$, $Q_n$ has no local optima for $\theta_2 \in \{0, 0.4\}$, in line with Proposition 5. For the larger $\theta_2 = 0.8$, there are local optima: equal weighting has two (one maximum and minimum), optimal weighting has four (two maxima and minima). Also, the objective becomes flatter as $\theta_2$ increases, making optimization more

[16] 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).

Table 3: MA(1): estimates under misspecification

| | Correctly Specified | | | | Moderately Misspecified | | | | Heavily Misspecified | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $W_n$ | Identity | | Optimal | | Identity | | Optimal | | Identity | | Optimal | |
| | $\hat{\theta}_n$ | $Q_n$ | $\hat{\theta}_n$ | $Q_n$ | $\hat{\theta}_n$ | $Q_n$ | $\hat{\theta}_n$ | $Q_n$ | $\hat{\theta}_n$ | $Q_n$ | $\hat{\theta}_n$ | $Q_n$ |
| TRUE | -0.070 | 0.084 | -0.040 | 0.078 | -0.590 | 0.447 | -0.140 | 0.285 | -0.82 | 1.10 | -0.84 | 1.02 |
| GN | -0.070 | 0.084 | -0.043 | 0.078 | -0.588 | 0.447 | -0.139 | 0.285 | 0.645 | 1.789 | 0.722 | 1.081 |
| BFGS | -14.4 | 0.084 | -22.8 | 0.078 | -6.95 | 0.52 | -7.35 | 0.285 | -1.21 | 1.10 | -5.92 | 1.09 |
| L-BFGS-B | -1.00 | 3.03 | -1.00 | 1.65 | -1.00 | 1.97 | -1.00 | 1.13 | -1.00 | 1.47 | -1.00 | 1.13 |

**Legend:** simulated sample of size $n = 200$, $\theta_1 = -0.1$, $p = 12$, $W_n = I_d$. TRUE is the actual sample estimator. Starting value $\theta_0 = 0.9$. $\hat{\theta}_n$: estimates returned by optimizer, $Q_n$: minimized objective.

challenging. Table 3 shows how this translates into estimation properties. As predicted, GN is robust to moderate misspecification but only converges to a local minimum under heavier misspecification. Other methods (L-BFGS-B, BFGS) systematically fail to converge.

## 5.2 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 'fake' cereal data generated by Nevo (2001).[17] 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 parameters of interest are the random coefficients $\theta$,[18] which enter $\delta_{jt}$, recovered from market shares $s_{jt}$ using the fixed point algorithm of Berry et al. (1995). The 25 linear coefficients $\beta$ are nuisance parameters concentrated out by two-stage least squares for each $\theta$. The replication sets the maximum number of iterations for the contraction mapping to 20000 and the tolerance level for convergence to $10^{-12}$. This is important for the optimization to be well-behaved; see e.g. Brunner et al. (2017), Conlon and Gortmaker (2020). The range of starting values used here is much wider than in these papers,[19] which explains why optimizers are more prone to crashing here than in their replications. Initial values are constructed as follows: the Sobol sequence generates values in $[0, 1]^8$, the coefficients for standard deviations are adjusted to lie in $[0, 10]$, those for income in $[-10, 10]$. Values for which the contraction mapping produces an error are discarded until 50 valid starting values are available.

[17] It available in the R package BLPestimatoR (Brunner et al., 2017). The data consists 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.

[18] 8 parameters are the unobserved standard deviation and the income coefficient on the constant term, price, sugar, and mushy.

[19] Conlon and Gortmaker (2020, p25) draw "starting values from a uniform distribution with support 50% above and below the true parameter value."

Table 4: Demand for Cereal: performance comparison

| | | STDEV | | | | INCOME | | | | objs | crash | time |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 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 | 00:03:51 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | |
| GN-B | avg | 0.28 | 2.03 | -0.01 | -0.08 | 3.58 | 0.47 | -0.17 | 0.69 | 33.84 | 0 | 00:00:20 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | |
| GD-B | avg | 0.29 | 2.18 | -0.01 | -0.08 | 3.59 | 0.43 | -0.17 | 0.71 | 35.17 | 0 | 09:59:04 |
| | std | 0.01 | 0.23 | 0.00 | 0.01 | 0.70 | 4.58 | 0.01 | 0.06 | 1.14 | | |
| BFGS | avg | 0.53 | 1.90 | -0.29 | -1.72 | 5.03 | 0.97 | -0.22 | 0.21 | 4555.97 | 23 | 00:01:46 |
| | std | 1.26 | 0.69 | 1.45 | 8.54 | 7.55 | 2.63 | 0.24 | 2.51 | $2.35 \cdot 10^4$ | | |
| NM | avg | 1.10 | 5.28 | -0.09 | 0.78 | 4.99 | 3.68 | -0.28 | 3.29 | 543.20 | 3 | 00:01:19 |
| | std | 1.44 | 7.74 | 0.11 | 1.86 | 4.43 | 8.99 | 0.26 | 3.47 | 700.29 | | |
| SA | avg | 7.66 | 9.52 | -0.94 | 10.45 | -0.27 | 2.01 | 3.73 | 3.07 | $8.27 \cdot 10^4$ | 2 | 01:45:59 |
| | std | 3.25 | 3.73 | 0.58 | 4.09 | 5.78 | 6.66 | 3.94 | 6.35 | $8.60 \cdot 10^4$ | | |
| SA+NM | avg | 1.02 | 8.90 | -0.13 | 1.00 | 4.75 | 7.64 | -0.29 | 4.65 | 613.70 | 2 | 01:47:33 |
| | std | 1.26 | 8.95 | 0.15 | 1.63 | 4.19 | 11.08 | 0.26 | 5.68 | 558.19 | | |

**Legend:** Comparison for 50 starting values where $[0, 10] \times \cdots \times [0, 10]$ for standard deviations and $[-10, 10] \times \cdots \times [-10, 10]$ for income coefficients. 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. crash: optimization terminated by an error. time: average run time for optimizers in hours:minutes:seconds. GN uses $\gamma = 0.1$, $k = 150$ iterations. GN-B and GD-B use a backtracking line search, terminates once $Q_n(\theta_k) - Q_n(\theta_{k+1}) \leq 10^{-8}$. Additional results can be found in Appendix F.1.

Table 4 and Figure 5 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) and Gradient-Descent (GD) for 50 different starting values.[20] As reported in Knittel and Metaxoglou (2014), optimization can crash often.[21] 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.[22] Only GN systematically produces accurate estimates; BFGS crashes 46% of the time and has one highly inaccurate estimate. Derivative-free optimizers (NM, SA, SA+NM) can produce inaccurate estimates. GD can be very slow to converge. Using a backtracking line search, GN converges in 11 iterations on average, compared to 8816 for GD – which has a higher maximum number of iterations set at 10000, compared to 150 for GN. Increasing the maximum number of iterations for GD would improve the estimates at the expense of

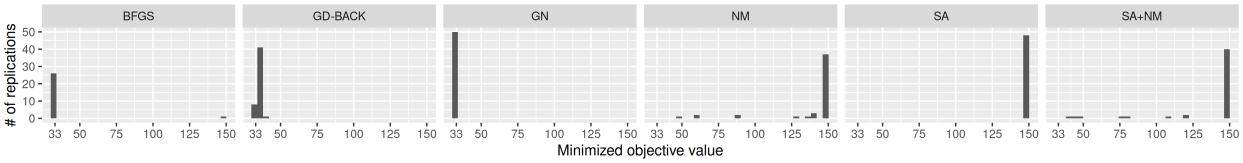
[20]The solution of the contraction mapping is not well defined for all values in $\Theta$, so we use the first 50 values produced by the Sobol sequence such that $\delta_{jt}$ is finite for all $j, t$.

[21]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.

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

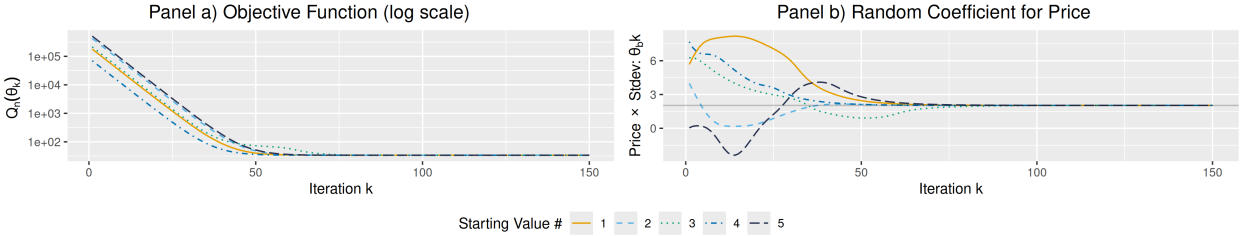
further computation time.

Figure 5: 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 6: Demand for Cereal: Gauss-Newton iterations for 5 starting values




**Legend:** 150 GN iterations for 5 starting values in $[0, 10] \times \cdots \times [0, 10]$ for standard deviations and $[-10, 10] \times \cdots \times [-10, 10]$ for income coefficients. Panel b) horizontal grey line = full sample estimate.

Figure 6, illustrates the convergence of GN for the first 5 starting values. In line with the predictions of Theorem 1, 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 two-stage least-squares for linearized BLP.

## 5.3 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.

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 (SIMS in the Table, Figures)[23]

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

algorithm with a reparameterization which bounds the coefficients.[24] Although this type of reparameterization is commonly used, the Jacobian is singular at the boundary; this matters for both local and global convergence, according to the results. As in the demand estimation, initial values are constructed using the Sobol sequence and adjusted to match the bounds used in the original study, reported in the last two rows of Table 5.

Table 5: Impulse Response Matching: performance comparison

| | | $\eta$ | $\nu$ | $\rho_s$ | $\sigma_s$ | objs | crash | time | $\eta$ | $\nu$ | $\rho_s$ | $\sigma_s$ | objs | crash | time |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TRUE | est | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | - | - | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | - | - |
| | | WITHOUT REPARAMETERIZATION | | | | | | | WITH REPARAMETERIZATION | | | | | | |
| GN | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 1 | 00:00:56 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 9 | 00:00:55 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | |
| GN-B | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 | 00:00:04 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 1 | 00:00:06 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | |
| GD-B | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 | 04:25:16 | 0.31 | 0.29 | 0.39 | 0.17 | 4.65 | 27 | 08:58:30 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | |
| BFGS | avg | -0.04 | -0.11 | -0.38 | 4.87 | $2 \cdot 10^4$ | 0 | 00:00:12 | 0.44 | 0.27 | 0.29 | 0.15 | 65.1 | 0 | 00:00:08 |
| | std | 0.25 | 0.93 | 0.45 | 3.79 | $2 \cdot 10^4$ | | | 0.32 | 0.16 | 0.50 | 0.07 | 101 | | |
| SIMS | avg | 0.23 | -0.23 | 0.31 | 0.18 | 42.2 | 0 | 00:00:40 | 0.61 | 0.25 | 0.09 | 0.14 | 118 | 0 | 00:00:38 |
| | std | 0.42 | 2.00 | 0.38 | 0.12 | 105 | | | 0.36 | 0.26 | 0.73 | 0.07 | 123 | | |
| NM | avg | 0.43 | -4.98 | 0.38 | 0.17 | 16.96 | 0 | 00:00:17 | 0.56 | 0.25 | 0.41 | 0.15 | 21.6 | 0 | 00:00:16 |
| | std | 0.44 | 37.3 | 0.22 | 0.05 | 39.9 | | | 0.34 | 0.16 | 0.29 | 0.05 | 31.5 | | |
| SA | avg | 1.55 | -1.45 | 0.50 | 0.09 | 74.8 | 0 | 00:04:45 | 0.66 | 0.19 | 0.63 | 0.05 | 194 | 0 | 00:02:32 |
| | std | 2.14 | 2.71 | 0.25 | 0.09 | 92.0 | | | 0.45 | 0.28 | 0.66 | 0.07 | 87.2 | | |
| SA | avg | 0.96 | -79.0 | 0.44 | 0.10 | 63.2 | 0 | 00:04:52 | 0.66 | 0.24 | 0.59 | 0.06 | 168 | 0 | 00:02:49 |
| +NM | std | 2.03 | 122 | 0.15 | 0.09 | 78.7 | | | 0.43 | 0.27 | 0.66 | 0.07 | 98.5 | | |
| lower b. | | 0.05 | 0.01 | -0.95 | 0.01 | - | - | - | 0.05 | 0.01 | -0.95 | 0.01 | - | - | - |
| upper b. | | 0.99 | 0.90 | 0.95 | 12 | - | - | - | 0.99 | 0.90 | 0.95 | 12 | - | - | - |

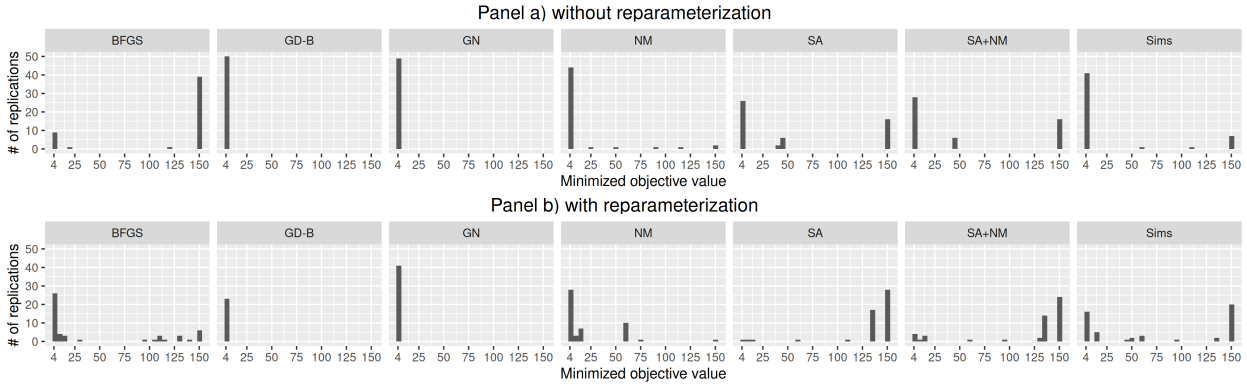
**Legend:** Comparison for 50 starting values. TRUE: full sample estimate (est). Objs: avg and std of minimized objective value. crash: optimization terminated because objective returned error. time: average run time for optimizers in hours:minutes:seconds. Lower/upper bound used for the reparameterization. GN run with $\gamma = 0.1$ for $k = 150$ iterations for all starting values. Standard errors were not computed in the original study. GN-B and GD-B use a backtracking line search, terminates once $Q_n(\theta_k) - Q_n(\theta_{k+1}) \leq 10^{-8}$. Additional results for GN, using a range of values $\gamma \in (0, 1]$ can be found in Appendix F.2.

In the original paper, the authors initialize 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 5. The model is estimated using CSMINWEL and the same set of optimizers used in the previous replication. Table 5 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 constraints in these cases. GN correctly estimates the parameters for all starting values but

[24]The replication uses the mapping $\theta_j = \underline{\theta}_j + \frac{\overline{\theta}_j - \underline{\theta}_j}{1+\exp(-\vartheta_j)}$, where each $\vartheta_j$ 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.

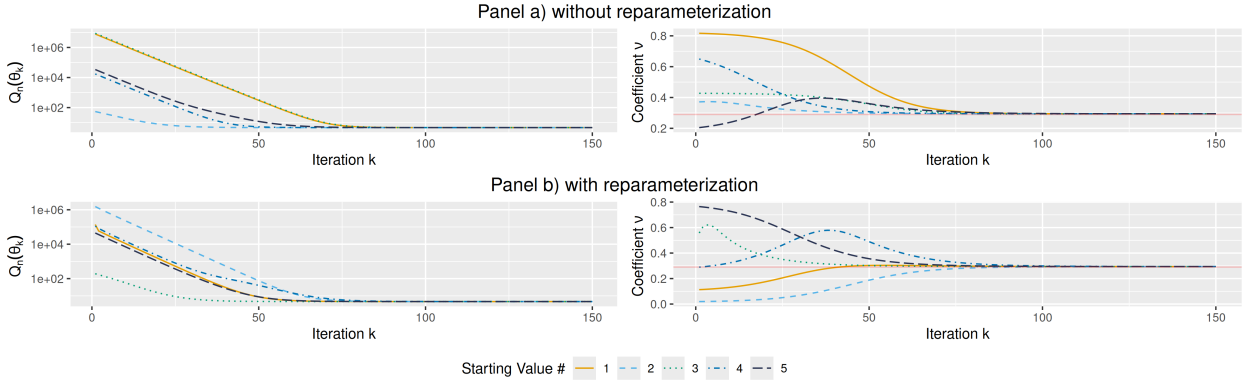
crashes twice for starting values for which both $\eta$ and $\nu$ are close to their lower bounds where the Jacobian is nearly singular. With the reparameterization, GN crashed more often, nines times in total, but is otherwise accurate. With backtracking line search, crashes are fewer for GN, and converges in 14 iterations, on average, with or without reparameterization, compared to 832 for GD without reparameterization and 2912 with reparameterization (both with cap of 10000). The crashes might also occur at values strictly within the parameter bounds for which Dynare cannot solve the model and returns an error. There is no obvious way to modify GN or GD to avoid this problem.

Figure 7: 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 8: Impulse Response Matching: Gauss-Newton iterations for 5 starting values




**Legend:** 150 GN iterations for 5 non-crashing starting values. Left: value of the objective function at each iteration; Right: coefficient $\eta$ at each iteration; horizontal light red line = full sample estimate.

The other two gradient-based optimizers, BFGS and SIMS(CSMINWEL), never crash because of better error handling in Matlab. They produce valid estimates less often than GN.

Figure 7 illustrates that CSMINWEL is sensitive to reparameterization. Likewise, derivative-free methods can be inaccurate, as illustrated in Table 5 and Figure 7; some crashes occur despite Matlab's error handling. Finally, Figure 8 shows 5 optimization paths for which GN does not crash with and without the reparameterization. Appendix F.2 gives additional results for larger values of $\gamma \in (0, 1]$ and error handling.

## 5.4 Convexity, Strong Injectivity, and Assumption 2 (b)

Table 6 illustrates the strong injectivity conditions, Assumption 2 (b), and convexity for the MA(1) model and the two empirical applications. A grid of 100 Sobol points was used to construct values within parameter bounds, respectively SB and LB, as described in Section 4.1; the first grid value is enforced to take only values from the bounds. The objective $Q_n$ is locally convex at $\theta$ if the Hessian is positive definite, i.e. $H_n(\theta) > 0$. The Table reports an estimate for $\mu_n$, $\rho\underline{\sigma}$ and the proportion of grid values where $Q_n$ is locally convex. To evaluate $\rho\underline{\sigma}$, the same step from Section 4.1 were used setting $\theta_k^2 = \hat{\theta}_n$ for all $k$. The $P_{k,n}$ for GN was used so that the $\mu_n$, $\rho\underline{\sigma}$ reported here are invariant to linear reparameterizations of both the parameters and the moments.

Table 6: Empirical and Illustrative Examples: Conditions, Convexity

| | | Strong Injectivity (SI') | | | Assumption 2 (b) | | | Convexity |
|---|---|---|---|---|---|---|---|---|
| | | $\hat{\mu}_n$ | $\overline{\gamma}$ | $\underline{k}$ | $\widehat{\rho\underline{\sigma}}_n$ | $\overline{\gamma}$ | $\underline{k}$ | $H_n > 0$ (%) |
| MA(1), $p = 1$ | SB | 0.5 | $3.1 \cdot 10^{-3}$ | $2.2 \cdot 10^{3}$ | 0.9 | $3.1 \cdot 10^{-3}$ | $2.1 \cdot 10^{3}$ | 46 |
| | LB | 0.0 | 0.0 | $\infty$ | 0.0 | 0.0 | $\infty$ | 40 |
| MA(1), $p = 12$ | SB | 0.15 | $4 \cdot 10^{-3}$ | $1.6 \cdot 10^{3}$ | 0.17 | $6 \cdot 10^{-3}$ | $1.2 \cdot 10^{3}$ | 98 |
| $W_n = I_d$ | LB | 0.0 | 0.0 | $\infty$ | 0.0 | 0.0 | $\infty$ | 90 |
| MA(1), $p = 12$ | SB | 0.12 | $3.5 \cdot 10^{-3}$ | $1.9 \cdot 10^{3}$ | 0.13 | $3.4 \cdot 10^{-3}$ | $2.0 \cdot 10^{3}$ | 98 |
| $W_n = \hat{V}_n^{-1}$ | LB | 0.0 | 0.0 | $\infty$ | 0.0 | 0.0 | $\infty$ | 90 |
| BLP | SB | 0.38 | 1.0 | 1 | 0.75 | 1.0 | 1 | 95 |
| | LB | 0.62 | $1.6 \cdot 10^{-5}$ | $4.3 \cdot 10^{5}$ | 0.68 | $2.3 \cdot 10^{-7}$ | $3.0 \cdot 10^{7}$ | 1 |
| DSGE | SB | 0.16 | $4.0 \cdot 10^{-10}$ | $1.7 \cdot 10^{10}$ | 0.37 | $1.3 \cdot 10^{-9}$ | $5.2 \cdot 10^{9}$ | 3 |
| | LB | 0.13 | $1.3 \cdot 10^{-11}$ | $5.5 \cdot 10^{11}$ | 0.33 | $2.7 \cdot 10^{-10}$ | $2.5 \cdot 10^{10}$ | 5 |
| DSGE | SB | 0.13 | $3.5 \cdot 10^{-12}$ | $2.0 \cdot 10^{12}$ | 0.23 | $3.5 \cdot 10^{-13}$ | $1.98 \cdot 10^{13}$ | 0 |
| (RE) | LB | $1.2 \cdot 10^{-14}$ | 0 | $\infty$ | 0.24 | $5.0 \cdot 10^{-13}$ | $1.4 \cdot 10^{13}$ | 0 |

**Legend:** Results for 100 sobol grid points, adjusted to match the bounds (Smaller Bounds SB, or Larger Bounds LB), for which the moments are well defined. DSGE, DSGE (RE) with/without reparameterization. **Bounds:** MA(1): SB $\Theta = [-0.9, 0.9]$ (rank conditions hold); LB $\Theta = [-1.0, 1.0]$ (rank conditions fail). BLP: SB $\Theta$ = values 50% above/below the true value (Conlon and Gortmaker, 2020, p25), LB $\Theta = [-10, 10] \times \cdots \times [-10, 10]$. DSGE: SB same as original paper plus/minus 0.1 for lower/upper bounds; LB same as original paper $\Theta = [0.05, 0.99] \times [0.01, 0.90] \times [-0.95, 0.95] \times [0.01, 12]$. **Convexity:** percentage (%) of points for which $H_n$ is strictly positive definite. **Sample sizes:** MA(1) $n = 200$, BLP $n = 2256$, DSGE $n = 63$.

For the MA(1) model, strong injectivity and Assumption 2 (b) fail at the boundary where $\theta = \pm 1$. This is visible in the results for LB. For SB, both conditions hold as illustrated in the

Table. Optimal weighting has some effect on the conditions and the predicted convergence properties. Convexity fails more often with a single moment condition ($p = 1$).

For BLP, the conditions appear to hold and predict fast convergence for SB, used in Conlon and Gortmaker (2020), where $Q_n$ is almost everywhere locally convex. With wider bounds (LB), convexity almost always fails, but the estimates for $\mu_n$, $\rho\underline{\sigma}$ are very close to SB. This confirm the good optimization properties for GN reported above.

For the DSGE model, $\mu_n$, $\rho\underline{\sigma}$ are of the same order of magnitude as the other applications without reparameterization. With reparameterization, the conditions can fail at the boundary which is visible in the Table under LB. With and without reparameterization, $Q_n$ is rarely locally convex, which confirms the challenges BFGS and CSMINWELL can have.

In both empirical applications, the estimates for $\overline{\gamma}$ and $\underline{k}$ tend to be very small and large, respectively, despite $\mu_n$, $\rho\underline{\sigma}$ being away from zero. This reflect the large amount of non-linearity, measured by $C_{3,K}$ and $L_{Q,P,K}$. As discussed in Section 4.1, these estimates can be fairly conservative which is clearly the case here. Also, because the moments are evaluated numerically, using a fixed-point algorithm for BLP, and the derivatives are computed by finite differences, the second-order derivatives can be fairly inaccurate. This issue is explained in Appendix F.3. Innacurate second-order derivatives can make optimizers like BFGS and CSMINWELL numerically unstable, but will also affect the value of $L_{Q,P}$, which tend to be very large in the empirical applications resulting in a very conservative bound for $\overline{\gamma}$.[25]

# 6 Conclusion

Non-convexity of the GMM objective function is considered to be an important challenge for structural estimation. This paper considers alternative conditions under which there are globally convergent algorithms. The results are robust to non-convexity, moderately non-linear one-to-one reparameterizations, and moderate misspecification. Though off-the-shelf methods might fail to converge due to the non-convexity of the optimization problem, the paper has shown that this does not necessarily imply that it will be difficult in practice. Econometric theory emphasizes the role of the weighting matrix $W_n$ on the statistical efficiency of the estimator $\hat{\theta}_n$. Here, Assumption 2 may or may not hold, depending on the choice of weighting matrix $W_n$. Its condition number $\kappa_W$ also affects local convergence which highlights an important role for the weighting matrix: it may facilitate or hinder the estimation itself.

[25]The estimate is $L_{Q,P,K} = 2 \cdot 10^4$ for BLP with large bounds, and $L_{Q,P,K} = 9$ with small bounds.

# 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.

Bosch, A. (1986): "The factorization of a square matrix into two symmetric matrices," *The American Mathematical Monthly*, 93, 462–464.

Boyd, S. and L. Vandenberghe (2004): *Convex optimization*, Cambridge university press.

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

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.

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.

Fisher, F. (1966): *The Identification Problem in Econometrics*, Economics handbook series, McGraw-Hill.

Forneron, J.-J. (2023): "Noisy, Non-Smooth, Non-Convex Estimation of Moment Condi-

tion Models," *arXiv preprint arXiv:2301.07196*.

Frobenius, G. (1910): "Über die mit einer Matrix vertauschbaren Matrizen," in *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften: Jahrgang 1910; Erster Halbband Januar bis Juni*, Verlag der Königlichen Akademie der Wissenschaften, 3–15.

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

Gourieroux, C., A. Monfort, and E. Renault (1993): "Indirect inference," *Journal of applied econometrics*, 8, S85–S118.

Guminov, S., A. Gasnikov, and I. Kuruzov (2017): "Accelerated Methods for $\alpha$-Weakly-Quasi-Convex Problems," *arXiv preprint arXiv:1710.00797*.

Hinder, O., A. Sidford, and N. Sohoni (2020): "Near-optimal methods for minimizing star-convex functions and beyond," in *Conference on learning theory*, PMLR, 1894–1938.

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

Karimi, H., J. Nutini, and M. Schmidt (2016): "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *Joint European conference on machine learning and knowledge discovery in databases*, Springer, 795–811.

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

Komunjer, I. (2012): "Global identification in nonlinear models with moment restrictions," *Econometric Theory*, 28, 719–729.

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.

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.

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

Nesterov, Y. and B. T. Polyak (2006): "Cubic regularization of Newton method and its global performance," *Mathematical programming*, 108, 177–205.

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.

Polyak, B. and A. Tremba (2020): "New versions of Newton method: step-size choice, convergence domain and under-determined equations," *Optimization Methods and Software*, 35, 1272–1303.

Polyak, B. T. (1963): "Gradient methods for minimizing functionals," *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3, 643–653.

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

Rockafellar, R. T. (2015): *Convex Analysis*, Princeton Landmarks in Mathematics and Physics, Princeton, NJ: Princeton University Press,.

Rothenberg, T. J. (1971): "Identification in parametric models," *Econometrica: Journal of the Econometric Society*, 577–591.

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

Solodov, M. V. and B. F. Svaiter (2000): "A truly globally convergent Newton-type method for the monotone nonlinear complementarity problem," *SIAM Journal on Optimization*, 10, 605–625.

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

Łojasiewicz, S. (1963): "A topological property of real analytic subsets," *Coll. du CNRS, Les équations aux dérivées partielles*, 117, 2.

# Appendix A Proofs for the Main Results

The proofs will make repeated use of the following mean value identity.

**Lemma A1** (Mean Value Identity)**.** *For any $g(\cdot)$ continuous differentiable on $\mathbb{R}^{d_\theta}$ with Jacobian $G(\cdot)$, let $\overline{G}(\theta_1,\theta_2) = \int_0^1 G(\omega\theta_1 + (1-\omega)\theta_2)d\omega$. For any $\theta_1, \theta_2 \in \mathbb{R}^{d_\theta}$:*

$$g(\theta_1) - g(\theta_2) = \overline{G}(\theta_1,\theta_2)(\theta_1 - \theta_2).$$

**Proof of Lemma A1:** Let $h : [0,1] \to \mathbb{R}^{d_g}$ be defined as $h(\omega) = g(\omega\theta_1 + (1-\omega)\theta_2)$, so that $g(\theta_1) - g(\theta_2) = h(1) - h(0) = \int_0^1 \partial_\omega h(\omega)d\omega$. By composition and the chain rule: $\partial_\omega h(\omega) = \partial_\theta g(\omega\theta_1 + (1-\omega)\theta_2)(\theta_1 - \theta_2) = G(\omega\theta_1 + (1-\omega)\theta_2)(\theta_1 - \theta_2)$. Plug this into the integral to find: $g(\theta_1) - g(\theta_2) = \overline{G}(\theta_1,\theta_2)(\theta_1 - \theta_2)$, as desired. □

## A.1 Implications of Assumptions 1, 2

In the following we will use the notation: $\overline{g}_n(\theta) = 1/n\sum_{i=1}^n g(\theta; x_i)$, $g(\theta) = \mathbb{E}[\overline{g}_n(\theta)]$, $G(\theta; x_i) = \partial_\theta g(\theta; x_i)$, $G_n(\theta) = 1/n\sum_{i=1}^n G(\theta; x_i)$, $G(\theta) = \mathbb{E}[G_n(\theta)]$, $Q_n(\theta) = 1/2\overline{g}_n(\theta)'W_n\overline{g}_n(\theta)$, and $Q(\theta) = 1/2g(\theta)'Wg(\theta)$. $W_n$ 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 $\theta^\dagger$. In the following, $\tilde{\Theta}$ generically denotes a compact convex subset of $\mathbb{R}^{d_\theta}$ such that $\theta^\dagger \in \text{interior}(\tilde{\Theta})$.

**Assumption A1.** *With probability approaching 1: i. $Q_n$ has a global minimizer on $\tilde{\Theta}$, $\hat{\theta}_n \in$ interior$(\tilde{\Theta})$, ii. $\overline{g}_n$ is twice continuously differentiable on $\tilde{\Theta}$, iii. $G_n$ is Lipschitz continuous with constant $L \geq 0$ on $\tilde{\Theta}$, 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. $W_n$ is such that $0 < \underline{\lambda}_W \leq \lambda_{\min}(W_n) \leq \lambda_{\max}(W_n) \leq \overline{\lambda}_W < \infty$.*

**Remarks.** 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 1i, iii which are used to derive uniform laws of large numbers for $\overline{g}_n$ and $G_n$.

**Lemma A2.** *Assumption 1 implies Assumption A1.*

**Lemma A3.** *Suppose Assumption 1 holds, then $\sup_{\theta\in\mathbb{R}^{d_\theta}} \|G_n(\theta) - G(\theta)\| = o_p(1)$. This implies that $\sup_{\theta_1,\theta_2\in\mathbb{R}^{d_\theta}} \|\overline{G}_n(\theta_1,\theta_2) - \overline{G}(\theta_1,\theta_2)\| = o_p(1)$ and $\sigma_{\max}[\overline{G}_n(\theta_1,\theta_2)] \leq \overline{\sigma}$, wpa1, uniformly in $\theta_1, \theta_2 \in \mathbb{R}^{d_\theta}$.*

**Lemma A4.** *Suppose Assumption 1 holds. Then, for some $r > 0$, Assumption 2 (a) holds for all $\theta \in \mathcal{B}_r(\theta^\dagger)$ with the same choice of $\rho$, $\underline{\sigma}$.*

**Lemma A5.** *Suppose Assumption 1 (iii), (v), (vi) and 2 (b) hold. Then, Assumption 1 (iv) holds for some strictly positive* $\tilde{\underline{\sigma}}, \tilde{R}$.

The following results are stated in terms of $\overline{G}_n(\theta) = \int_0^1 \{G_n(\omega\theta + (1-\omega)\hat{\theta}_n)\} d\omega$.

**Assumption A2.** *With probability approaching 1, for all* $\theta \in \mathbb{R}^{d_\theta}$*: (a)* $\sigma_{\min}[G_n(\theta)' W_n \overline{G}_n(\theta)] \geq \rho\underline{\sigma}$*, (b)* $\|G_n(\theta)' W_n \overline{G}_n(\theta)(\theta - \theta^\dagger)\| \geq \rho\underline{\sigma}\|\theta - \theta^\dagger\|$.

**Lemma A6.** *Suppose Assumptions 1 holds. 1) If Assumption 2 (a) holds, Assumption A2 (a) holds. 2) If Assumption 2 (b), Assumption A2 (b) holds.*

**Proof of Lemma 1.** Lemma A3 implies that $\overline{G}_n(\theta_1, \theta_2)$ is uniformly consistent in $\theta_1, \theta_2 \in \mathbb{R}^{d_\theta}$. With this in mind, Lemma A1 implies:

$$\overline{g}_n(\theta) - \overline{g}_n(\theta^\dagger) = \overline{G}_n(\theta, \theta^\dagger)(\theta - \theta^\dagger) = [\overline{G}(\theta) + o_p(1)](\theta - \theta^\dagger),$$

uniformly in $\theta \in \mathbb{R}^{d_\theta}$. Now Assumption 2 (b) implies:

$$\|\overline{g}_n(\theta) - \overline{g}_n(\theta^\dagger)\| \geq (\rho\underline{\sigma}/[\overline{\lambda}_W \overline{\sigma}] - o_p(1))\|\theta - \theta^\dagger\|.$$

Using the triangular inequality: $\|\overline{g}_n(\theta)\|_{W_n} \geq \underline{\lambda}_W^{1/2}(\rho\underline{\sigma}/[\overline{\lambda}_W\overline{\sigma}] - o_p(1))\|\theta - \theta^\dagger\| - \|\overline{g}_n(\theta^\dagger)\|_{W_n}$, uniformly in $\theta \in \mathbb{R}^{d_\theta}$. For any $\|\theta - \theta^\dagger\| \geq [2\|g(\theta^\dagger)\|_W + 1]\overline{\lambda}_W\overline{\sigma}/[\underline{\lambda}_W^{1/2}\rho\underline{\sigma}]$, this implies:

$$\|\overline{g}_n(\theta)\|_{W_n} \geq \|g(\theta^\dagger)\|_W + 1 - o_p(1).$$

Now, given that $\|\overline{g}_n(\theta^\dagger)\|_{W_n} \leq \|g(\theta^\dagger)\|_W + 1$ wpa1, this implies that $\|\hat{\theta}_n - \theta^\dagger\| \leq [2\|g(\theta^\dagger)\|_W + 1]\overline{\lambda}_W\overline{\sigma}/[\underline{\lambda}_W^{1/2}\rho\underline{\sigma}]$, wpa1. Then, uniform convergence on compact sets (Lemma A2), and the identification conditions imply that $\hat{\theta}_n \xrightarrow{p} \theta^\dagger$, using standard arguments (e.g. Newey and McFadden, 1994, Th2.1). Again, $Q_n$ is uniformly consistent on compact sets, so $Q_n(\hat{\theta}_n) \xrightarrow{p} Q(\theta^\dagger)$. This concludes the proof. □

**Proof of Lemma A2.** In the following, all the strict inequalities are replaced by weak inequalities with some slackness $\delta > 0$, e.g. $\sigma_{\min}(G(\theta)) \geq (1+\delta)\underline{\sigma} > 0$ instead of $\sigma_{\min}(G(\theta)) > \underline{\sigma} > 0$, and $\lambda_{\max}(W) \leq (1-\delta)\overline{\lambda}_W < \infty$ instead of $\lambda_{\max}(W) < \overline{\lambda}_W < \infty$. Assumption A1ii, iv follow from 1ii, 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 A1v.

Assumption 1iii and compactness imply uniform convergence of the sample Jacobian $\sup_{\theta\in\tilde{\Theta}} \|G_n(\theta) - G(\theta)\| = o_p(1)$, 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 $\theta \in \tilde{\Theta}$. Notice that $\|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\| = \|[\overline{G}_n(\theta_1, \theta_2) - \overline{G}(\theta_1, \theta_2)](\theta_1 - \theta_2)\| \leq [\sup_{\theta\in\tilde{\Theta}} \|G_n(\theta) - G(\theta)\|]\|\theta_1 - \theta_2\|$, where the sup is a $o_p(1)$ by uniform convergence of $G_n$, and $\overline{G}(\theta_1, \theta_2) = \int_0^1 G(\omega\theta_1 + (1-\omega)\theta_2)d\omega$. Using a finite cover and arguments similar to Jennrich (1969), this implies uniform convergence: $\sup_{\theta\in\tilde{\Theta}} \|\overline{g}_n(\theta) - g(\theta)\| = o_p(1)$.

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 1i. 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. $\mathcal{B}_{R_G}(\hat{\theta}_n) \subseteq \mathcal{B}_{(1+\delta)R_G}(\theta^\dagger) \subseteq \tilde{\Theta}$. This implies $\hat{\theta}_n \in \text{interior}(\tilde{\Theta})$, wpa1. Then, for the same $\theta$, $\sigma_{\min}[G(\theta)] \geq (1+\delta)\underline{\sigma}$, wpa1. Apply Weyl's inequality for singular values to find that, uniformly in $\theta$: $\sigma_{\min}[G_n(\theta)] \geq \sigma_{\min}[G_n(\theta)] - \sigma_{\max}[G(\theta) - G_n(\theta)] \geq (1+\delta)\underline{\sigma} - o_p(1) \geq \underline{\sigma} > 0$, wpa 1. Take any two $\theta_1, \theta_2$ in $\tilde{\Theta}$, $\|G_n(\theta_1) - G_n(\theta_2)\| \leq 1/n \sum_{i=1}^n \bar{L}(x_i)\|\theta_1 - \theta_2\| \leq [(1-\delta)L + o_p(1)]\|\theta_1 - \theta_2\| \leq L\|\theta_1 - \theta_2\|$, wpa1, using a law of large numbers for $\bar{L}(x_i)$. This yields all the conditions in Assumption A1iii. □

**Proof of Lemma A3.** Pick $\delta > 0$, set $(1+R) \geq \frac{3M}{\delta}$ so that $\|G(\theta) - G(\theta_R)\| \leq \delta/3$ for any $\theta \in \mathbb{R}^{d_\theta}$. Since $\Theta_R = \{\theta \in \mathbb{R}^{d_\theta}, \|\theta\| \leq R\}$ is compact, $\sup_{\theta\in\Theta_R} \|G_n(\theta) - G(\theta)\| = o_p(1)$, using Lemma A2. Likewise,

$$\|G_n(\theta) - G_n(\theta_R)\| \leq \left[\frac{1}{n}\sum_{i=1}^n \bar{M}(x_i)\right]/(1+R) \leq [M + o_p(1)]/(1+R) \leq \delta/3 + o_p(1).$$

Then, combine these results to find:

$$\begin{aligned}\|G_n(\theta) - G(\theta)\| &\leq \|G_n(\theta) - G_n(\theta_R)\| + \|G_n(\theta_R) - G(\theta_R)\| + \|G(\theta) - G(\theta_R)\| \\ &\leq 2/3\delta + o_p(1),\end{aligned}$$

uniformly in $\theta \in \mathbb{R}^{d_\theta}$. This implies uniform consistency: $\lim_{n\to\infty} \mathbb{P}(\sup_{\theta\in\mathbb{R}^{d_\theta}} \|G_n(\theta) - G(\theta)\| > \delta) = 0$. Also,$\|\overline{G}_n(\theta_1, \theta_2) - \overline{G}(\theta_1, \theta_2)\| \leq \sup_{\theta\in\mathbb{R}^{d_\theta}} \|G_n(\theta) - G(\theta)\| = o_p(1)$ and $\sigma_{\max}[\overline{G}_n(\theta_1, \theta_2)] \leq \sup_{\theta\in\mathbb{R}^{d_\theta}} \sigma_{\max}[G_n(\theta)] \leq \overline{\sigma}$, wpa1, which is the desired result. □

**Proof of Lemma A4:** Under Assumption 1, $\sigma_{\min}[G(\theta)] \geq (1+\delta)\underline{\sigma}$ for all $\theta \in \mathcal{B}_{R_G}(\theta^\dagger)$ and some $\delta > 0$. Also, $G$ is Lipschitz continuous with constant $L$ since $\|G(\theta_1) - G(\theta_2)\| \leq$

$\mathbb{E}[\|G(\theta_1; x_i) - G(\theta_2; x_i)\|] \leq L\|\theta_1 - \theta_2\|$. As a result, $\|\overline{G}(\theta) - G(\theta^\dagger)\| \leq L\|\theta - \theta^\dagger\|$. Then,

$$\|G(\theta)'W\overline{G}(\theta) - G(\theta^\dagger)'WG(\theta^\dagger)\| \leq 2\overline{\sigma}\overline{\lambda}_W L\|\theta - \theta^\dagger\|.$$

Apply Weyl's inequality to find:

$$\sigma_{\min}[G(\theta)'W\overline{G}(\theta)] \geq \big\{(1+\delta)[\underline{\lambda}_W\underline{\sigma}] - 2\frac{\overline{\sigma}\overline{\lambda}_W L}{\underline{\sigma}}\|\theta - \theta^\dagger\|\big\}\underline{\sigma}.$$

Pick $\|\theta - \theta^\dagger\| \leq r$ with $r$ such that $\delta > 2\overline{\sigma}L\overline{\lambda}_W/[\underline{\lambda}_W\underline{\sigma}^2]r$ to find: $\sigma_{\min}[G(\theta)'W\overline{G}(\theta)] > [\underline{\lambda}_W\underline{\sigma}]\underline{\sigma}$, given that $0 < \rho \leq \underline{\lambda}_W\underline{\sigma}$ in Assumption 2 (a), this yields the result. □

**Proof of Lemma A5:** Take $\theta = \theta^\dagger + \varepsilon v$, with $v$ unitary. Assumption 2 (b) and the regularity conditions (Assumption 1 (iii), (v), (vi)) imply: $\|G(\theta^\dagger)'WG(\theta^\dagger)v\| > \rho\underline{\sigma} - 2\overline{\sigma}\overline{\lambda}_W L\varepsilon$, take $\varepsilon \to 0$, to find $\lambda_{\min}[G(\theta^\dagger)'WG(\theta^\dagger)] > \rho\underline{\sigma}$ so that $\sigma_{\min}[G(\theta^\dagger)] > \rho\underline{\sigma}/[\overline{\sigma}\overline{\lambda}_W]$. Pick $\tilde{\underline{\sigma}} = 1/2\rho\underline{\sigma}/[\overline{\sigma}\overline{\lambda}_W]$. The Lipschitz continuity of $G$ implies $\sigma_{\min}[G(\theta)] \geq \sigma_{\min}[G(\theta^\dagger)] - L\|\theta - \theta^\dagger\| > 1/2\rho\underline{\sigma}/[\overline{\sigma}\overline{\lambda}_W]$ for $\|\theta - \theta^\dagger\| \leq \tilde{R} < 1/2\rho\underline{\sigma}/[L\overline{\sigma}\overline{\lambda}_W]$. □

**Proof of Lemma A6.** Lemmas 1 and A3 apply so that $G_n$ is uniformly convergent and Lipschitz continuous, $\hat{\theta}_n$ is consistent. 1) This implies that:

$$\begin{aligned}\|\overline{G}_n(\theta) - \overline{G}(\theta)\| &= \|\int_0^1 \{G_n(\omega\theta + (1-\omega)\hat{\theta}_n) - G(\omega\theta + (1-\omega)\theta^\dagger)\}d\omega\| \\ &\leq L\|\hat{\theta}_n - \theta^\dagger\| + \sup_{\theta\in\Theta}\|G_n(\theta) - G(\theta)\| = o_p(1).\end{aligned}$$

Then apply Weyl's inequality to find that, uniformly in $\theta$ and wpa1: $\sigma_{\min}[G_n(\theta)] \geq \sigma_{\min}[G(\theta)] - o_p(1)$, $\sigma_{\min}[\overline{G}_n(\theta)] \geq \sigma_{\min}[\overline{G}(\theta)] - o_p(1)$, and $\sigma_{\min}[G_n(\theta)'W_n\overline{G}_n(\theta)] \geq \sigma_{\min}[G(\theta)'W\overline{G}(\theta)] - o_p(1)$, which yields the result.

2) Lemma A4 implies Assumption 2 (a) holds locally, i.e. for $\|\theta - \theta^\dagger\| \leq r$, with $r > 0$. With the derivations above, this implies that Assumption A2 (a) holds locally as well, i.e. for $\|\theta - \hat{\theta}_n\| \leq r/2$, wpa1. Recall that Assumption A2 (a) implies Assumption A2 (b).

Take $\|\theta - \hat{\theta}_n\| \geq r/2$. By uniform consistency and boundedness of $G_n$ and $\overline{G}_n$, we have: $G_n(\theta)'W_n\overline{G}_n(\theta) = G(\theta)'W\overline{G}(\theta) + o_p(1)$, uniformly in $\theta$ using $\sigma_{\max}[G_n(\theta)] \leq \overline{\sigma}$ wpa1. Since

$\hat{\theta}_n$ is consistent, we have uniformly in $\|\theta - \hat{\theta}_n\| \geq r/2$:

$$\begin{aligned}
\|G_n(\theta)'W_n\overline{G}_n(\theta)(\theta - \hat{\theta}_n)\| &\geq \|G(\theta)'W\overline{G}(\theta)(\theta - \hat{\theta}_n)\| - o_p(1)\|\theta - \hat{\theta}_n\| \\
&\geq \|G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)\| - o_p(1)\|\theta - \hat{\theta}_n\| - \overline{\sigma}^2\overline{\lambda}_W o_p(1) \\
&\geq (1+\delta)\rho\underline{\sigma}\|\theta - \theta^\dagger\| - o_p(1)\|\theta - \hat{\theta}_n\| - \overline{\sigma}^2\overline{\lambda}_W o_p(1) \\
&\geq [(1+\delta)\rho\underline{\sigma} - o_p(1)]\|\theta - \hat{\theta}_n\| - [\overline{\sigma}^2\overline{\lambda}_W + (1+\delta)\rho\underline{\sigma}]o_p(1) \\
&\geq \left[(1+\delta)\rho\underline{\sigma} - o_p(1) - o_p(1)2\frac{\overline{\sigma}^2\overline{\lambda}_W + (1+\delta)\rho\underline{\sigma}}{r}\right]\|\theta - \hat{\theta}_n\|,
\end{aligned}$$

using $\|\theta - \hat{\theta}_n\|/(r/2) \geq 1$ for the last inequality. The leading term is greater or equal than $\rho\underline{\sigma}$ wpa1 which yields the result. □

## A.2 Proofs for Section 2.2

**Proof of Theorem 1:** Take $\theta_0 \in \Theta$, let $\Theta_n = \left\{\tilde{\theta} \in \mathbb{R}^{d_\theta}, Q_n(\tilde{\theta}) \leq Q_n(\theta_0)\right\}$. From the proof of Lemma 1, we have:

$$\sqrt{2Q_n(\tilde{\theta})} \geq (\rho\underline{\sigma}/[\overline{\lambda}_W\overline{\sigma}] - o_p(1))\|\tilde{\theta} - \theta^\dagger\| - \|g(\theta^\dagger)\|_W - o_p(1),$$

uniformly in $\tilde{\theta} \in \mathbb{R}^{d_\theta}$. Now take $Q_0 = \sup_{\theta\in\Theta} Q(\theta)$ and let:

$$\Theta_0 = \left\{\tilde{\theta} \in \mathbb{R}^{d_\theta}, \|\tilde{\theta} - \theta^\dagger\| \leq 2\frac{\sqrt{2Q_0} + \sqrt{2Q(\theta^\dagger)} + 1}{\rho\underline{\sigma}/[\overline{\lambda}_W]\overline{\sigma}}\right\},$$

a compact subset of $\mathbb{R}^{d_\theta}$. We have $\Theta \subseteq \Theta_0$ and $\Theta_n \subseteq \Theta_0$, wpa1. Now, let: $R_\Theta = 4\overline{\lambda}_P\overline{\sigma}^2\overline{\lambda}_W\text{diam}(\Theta_0)$, which bounds $\|\theta_{k+1} - \theta_k\|$ wpa1, uniformly in $\theta_k \in \Theta_0$, for any choice of $\gamma \in [0,1]$. Uniformly in $\theta_k \in \Theta_0$, $\theta_{k+1}$ computed in (1) satisfies $\theta_{k+1} \in \Theta_R = \cup_{\theta\in\Theta_0} B_{R_\Theta}(\theta)$ wpa1. The sample moments and Jacobian are uniformly consistent on $\Theta_R \supseteq \Theta_0$. Then, by recursion over $k \geq 0$, the following establishes that uniformly in $\theta_k \in \Theta_0$, $Q_n(\theta_{k+1}) \leq Q_n(\theta_k)$. Hence, $\theta_k \in \Theta_0$ for all $k \geq 0$ wpa1, uniformly in $\theta_0 \in \Theta$. So the derivations below can proceed under Assumption A1, with $\tilde{\Theta} = \Theta_R$, since the path is compact-valued wpa1.

**Case 1) Just-identifed:** Since Assumptions A1 and A2 hold (using Lemmas A2, A6), Proposition 1 (1)-(2) holds, with probability approaching 1, for the sample moments with the same choice of strictly positive constants $C_1, C_2, C_3$. Denote by $L_Q$ the Lipschitz constant of

$\partial_\theta Q_n$. The mean value value theorem implies that for some $\tilde{\theta}_k$ between $\theta_k$ and $\theta_{k+1}$:

$$\begin{aligned}
Q_n(\theta_{k+1}) &= Q_n(\theta_k) - \gamma \partial_\theta Q_n(\theta_k) P_{k,n} \partial_\theta Q_n(\theta_k) - \gamma \{\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)\} P_{k,n} \partial_\theta Q_n(\theta_k) \\
&\leq Q_n(\theta_k) - \gamma \underline{\lambda}_P \|\partial_\theta Q_n(\theta_k)\|^2 + \gamma^2 L_Q \overline{\lambda}_P^2 \|\partial_\theta Q_n(\theta_k)\|^2 \\
&\leq Q_n(\theta_k) + \gamma \{-\underline{\lambda}_P + \gamma L_Q \overline{\lambda}_P^2\} \|\partial_\theta Q_n(\theta_k)\|^2 \\
&\leq Q_n(\theta_k) - \gamma \underline{\lambda}_P / 2 \|\partial_\theta Q_n(\theta_k)\|^2 \\
&\leq Q_n(\theta_k) - \gamma \underline{\lambda}_P C_1 / 2 (Q_n(\theta_k) - Q_n(\hat{\theta}_n)),
\end{aligned}$$

if $0 < \gamma \leq \underline{\lambda}_P / [2 L_Q \overline{\lambda}_P^2]$. Substract $Q_n(\hat{\theta}_n)$ on both sides to find:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) = \{1 - \gamma C_1 \underline{\lambda}_P / 2\} (Q_n(\theta_k) - Q_n(\hat{\theta}_n)).$$

Set $(1 - \overline{\gamma})^2 = 1 - \gamma \underline{\lambda}_P C_1 / 2$ and iterate over $k = 0, \ldots$ to find:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1 - \overline{\gamma})^{k+1} \sqrt{C_3 / C_2} \|\theta_0 - \hat{\theta}_n\|,$$

which is the desired result.

**Case 2) Over-identifed:** Since Assumptions A1 and A2 hold (using Lemmas A2, A6), and $Q_n(\hat{\theta}_n) = o_p(1)$, Proposition 5 (1')-(2) holds, with probability approaching 1, for the sample moments with the same choice of strictly positive constants $C_2, C_3, C_4$. Let

$$C_{1n} = \frac{(\rho \underline{\sigma} - \overline{\lambda}_W^{1/2} L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n})^2}{C_3 + C_4 \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}} = C_1 + o_p(1),$$

for the same $C_1$ found in Proposition 1 (1). Denote by $L_Q$ the Lipschitz constant of $\partial_\theta Q_n$. The mean value theorem implies that for some $\tilde{\theta}_k$ between $\theta_k$ and $\theta_{k+1}$:

$$\begin{aligned}
Q_n(\theta_{k+1}) &= Q_n(\theta_k) - \gamma \partial_\theta Q_n(\theta_k) P_{k,n} \partial_\theta Q_n(\theta_k) - \gamma \{\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)\} P_{k,n} \partial_\theta Q_n(\theta_k) \\
&\leq Q_n(\theta_k) + \gamma \{-\underline{\lambda}_P + \gamma L_Q \overline{\lambda}_P^2\} \|\partial_\theta Q_n(\theta_k)\|^2 \\
&\leq Q_n(\theta_k) - \gamma \underline{\lambda}_P C_{1n} / 2 (Q_n(\theta_k) - Q_n(\hat{\theta}_n)),
\end{aligned}$$

if $0 < \gamma \leq \underline{\lambda}_P / [2 L_Q \overline{\lambda}_P^2]$. Substract $Q_n(\hat{\theta}_n)$ on both sides to find:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) = \{1 - \underline{\lambda}_P \gamma C_{1n} / 2\} (Q_n(\theta_k) - Q_n(\hat{\theta}_n)).$$

Set $(1-\overline{\gamma})^2 = 1 - \gamma\underline{\lambda}_P C_{1n}/2$ and iterate over $k = 0, \ldots$ to find:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})^{k+1} \frac{\sqrt{C_3 + C_4\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}}{\sqrt{C_2 - C_4\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}} \|\theta_0 - \hat{\theta}_n\|,$$

which is the desired result. □

**Proof of Theorem 2:** The proof is similar to Theorem 1, the condition on $\varphi$ ensures that inequalities (1')-(2) in Proposition 5 hold with strictly positive constants, with probability approaching 1, for the sample moments. □

# Appendix B Proofs and additional results for Section 3

## B.1 Properties related to Strong Injectivity

**Lemma B7** (From (SI') to (SI), Assumption 3)**.** *Suppose Assumption 1 (iii), (vi) hold, then: 1) (SI') implies (SI), and 2) (SI') implies Assumption 3 holds for* GN*.*

**Lemma B8** (From (SI) to (SI'))**.** *Suppose Assumption 1 (iii), (vi) hold and $g = (g_1', g_2')'$ where $g_1$ is just-identified and satisfies (SI) for some $\mu_1 > 0$. Let $\tilde{W}(\lambda) = \lambda W + (1-\lambda)\,blockdiag(W_1, 0)$, where $W_1$ is the upper block of $W$ corresponding to $g_1$. If $W_1$ is invertible, then there exists $\lambda^\star \in (0,1]$ such that (SI') holds using $\tilde{W}(\lambda)$ for any $0 \leq \lambda \leq \lambda^\star$.*

## B.2 Additional Results for Over-Identified Models

**Proposition B8.** *(Sufficient Conditions: Over-Identified) Consider the following three conditions: (a) $\sigma_{\min}[G(\theta)'W\overline{G}(\theta_1,\theta_2)] > \underline{\sigma} > 0$, for all $\theta, \theta_1, \theta_2 \in \mathbb{R}^{d_\theta}$, (b) for all $\theta \in \mathbb{R}^{d_\theta}$, $G(\theta) = US(\theta)V$ for $U, V$ full rank, $S(\theta)$ symmetric with $0 < \underline{\lambda}_S < \lambda_{\min}[S(\theta)] < \lambda_{\max}[S(\theta)] < \overline{\lambda}_S < \infty$, and $U'WU$ invertible.*
*The following holds: (1) (b) $\Rightarrow$ (a) $\Rightarrow$ Assumption 2 (a), (2) (a) implies $G(\theta_1)'Wg(\cdot)$ is one-to-one, for any $\theta_1 \in \mathbb{R}^{d_\theta}$.*

**Proposition B9.** *(Reparameterization: Over-Identified) Take $h$ as in Proposition 7. 1) If Assumption 2 (a) holds for $g$ and $\underline{\sigma} > \overline{\lambda}_W[C_1\overline{\sigma}_h\overline{\sigma}^2 + C_2 L\overline{\sigma}_h^2\overline{\sigma}]/\underline{\sigma}_h^2$, then Assumption 2 (a) holds for $g \circ h$. In particular, if $h = Au + b$ is affine with $A$ invertible then $C_1 = C_2 = 0$ and Assumption 2 (a) holds for $g \circ h$. 2) Suppose Assumption 2 (b) holds for $g$. If*

$\|h(u) - h(u^\dagger)\| \geq \mu\|u - u^\dagger\|$, *for some* $\mu > 0$ *and all* $u \in \mathcal{U}$, *then Assumption 2 (b) holds for* $g \circ h$.

## B.3 Proofs for Section 3 and the additional results

**Proof of Proposition 1:** We first prove (2). For any $\theta \in \mathbb{R}^{d_\theta}$, $g(\theta) = g(\theta) - g(\theta^\dagger) = \overline{G}(\theta)(\theta - \theta^\dagger)$, for correctly specified models. This implies that $Q(\theta) = 1/2(\theta - \theta^\dagger)'\overline{G}(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)$. Assumption 1 (iii) implies $\sigma_{\max}[\overline{G}(\theta)] \leq \max_{\theta \in \mathbb{R}^{d_\theta}} \sigma_{\max}[G(\theta)] \leq \overline{\sigma} < +\infty$. Assumption 2 (b) implies $\overline{\sigma}\overline{\lambda}_W^{1/2}\|W^{1/2}\overline{G}(\theta)(\theta - \theta^\dagger)\| \geq \rho\underline{\sigma}\|\theta - \theta^\dagger\|$ and $\|W^{1/2}\overline{G}(\theta)(\theta - \theta^\dagger)\| = \sqrt{2[Q(\theta) - Q(\theta^\dagger)]}$. Putting these together yields:

$$1/2\frac{\rho^2\underline{\sigma}^2}{\overline{\sigma}^2\overline{\lambda}_W}\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger) \leq 1/2\overline{\sigma}^2\overline{\lambda}_W\|\theta - \theta^\dagger\|^2.$$

Now, we prove (1). We have $\partial_\theta Q(\theta) = G(\theta)'Wg(\theta) = G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)$. Assumption 2 (b) implies:

$$\|\partial_\theta Q(\theta)\|^2 \geq \rho^2\underline{\sigma}^2\|\theta - \theta^\dagger\|^2 \geq \frac{\rho^2\underline{\sigma}^2}{1/2\overline{\sigma}^2\overline{\lambda}_W}[Q(\theta) - Q(\theta^\dagger)],$$

using (2). This is the desired result. □

**Proof of Proposition 2:** For correctly specified models, $\partial_\theta Q(\theta) = G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)$.
1) If the PL inequality holds, the quadratic lower bound implies $\|G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)\|^2 \geq \mu C_2\|\theta - \theta^\dagger\|^2$, i.e. Assumption 2 (b) holds.
2) By definition, $Q$ is quasar-convex if, and only if, there are $\lambda \geq 1$ and $\mu \geq 0$ such that:

$$\partial_\theta Q(\theta)(\theta - \theta^\dagger) \geq \frac{1}{\lambda}\{Q(\theta) - Q(\theta^\dagger)\} + \frac{\mu}{2\lambda}\|\theta - \theta^\dagger\|^2,$$

where $\partial_\theta Q(\theta)(\theta - \theta^\dagger) = (\theta - \theta^\dagger)'G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)$. Since $Q(\theta) - Q(\theta^\dagger) \geq 0$ we have:

$$(\theta - \theta^\dagger)'G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger) \geq \frac{\mu}{2\lambda}\|\theta - \theta^\dagger\|^2.$$

Now apply the Cauchy-Schwarz inequality to find:

$$\|\theta - \theta^\dagger\|\|G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)\| \geq (\theta - \theta^\dagger)'G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger) \geq \frac{\mu}{2\lambda}\|\theta - \theta^\dagger\|^2,$$

which implies Assumption 2 (b). □

**Proof of Proposition 3:** 1) Strong monotonicity of $Ag$ implies $(\theta_1-\theta_2)'A\overline{G}(\theta_1,\theta_2)(\theta_1-\theta_2) \geq \mu\|\theta_1-\theta_2\|^2$ since $g(\theta_1)-g(\theta_2)=\overline{G}(\theta_1,\theta_2)(\theta_1-\theta_2)$. For any unit vector $v$, take $\theta_2=\theta_1+\varepsilon v$ and let $\varepsilon\to 0$ to find $v'AG(\theta_1)v=\frac{1}{2}v'[AG(\theta_1)+G(\theta_1)'A']v\geq\mu$ so that $G(\theta_1)$ has full rank and $AG(\theta_1)+G(\theta_1)'A'$ is positive definite. We have $\sigma_{\min}[G(\theta)]\geq\mu\sigma_{\min}(A)^{-1}:=\underline{\sigma}>0$, as a normalization. Pick $\theta_2=\theta^\dagger$, use the Cauchy-Schwarz inequality to find $\|A'(\theta-\theta^\dagger)\|\|\overline{G}(\theta,\theta^\dagger)(\theta-\theta^\dagger)\|\geq(\theta-\theta^\dagger)'A\overline{G}(\theta,\theta^\dagger)(\theta-\theta^\dagger)\geq\mu\|\theta-\theta^\dagger\|^2$.

Because $G(\theta)'W$ is invertible, we can write $\|G(\theta)'W\overline{G}(\theta,\theta^\dagger)(\theta-\theta^\dagger)\|\geq\underline{\sigma}\underline{\lambda}_W\|\overline{G}(\theta,\theta^\dagger)(\theta-\theta^\dagger)\|\geq\underline{\sigma}\mu\underline{\lambda}_W\sigma_{\max}(A)^{-1}\|\theta-\theta^\dagger\|$; Assumption 2 (b) holds for any appropriate choice of $0<\rho\leq\mu\underline{\lambda}_W\sigma_{\max}(A)^{-1}$.
2) Strong injectivity of $g$ implies $\|\overline{G}(\theta_1,\theta_2)(\theta_1-\theta_2)\|\geq\mu\|\theta_1-\theta_2\|$, for any pair $\theta_1,\theta_2$. Using the same arguments as above: $G(\theta)$ has full rank for all $\theta$ and $\|G(\theta)'W\overline{G}(\theta,\theta^\dagger)(\theta-\theta^\dagger)\|\geq\underline{\sigma}\underline{\lambda}_W\mu\|\theta-\theta^\dagger\|$; Assumption 2 (b) holds for any appropriate choice of $0<\rho\leq\mu\underline{\lambda}_W$. □

**Proof of Proposition 4.** Assumption 1 (ii)-(vi) implies Assumption 2 (a) holds locally (Lemma A4). Hence, for $\|\theta-\theta^\dagger\|\leq r$, we have $\|G(\theta)'W\overline{G}(\theta-\theta^\dagger)\|\geq\rho\underline{\sigma}\|\theta-\theta^\dagger\|$. Condition (N) implies that for $R\geq\|\theta-\theta^\dagger\|\geq r$ we have:

$$\inf_{\theta,R\geq\|\theta-\theta^\dagger\|\geq r}\|\partial_\theta Q(\theta)\|\geq\delta(r,R)\geq\frac{\delta(r,R)}{R}\|\theta-\theta^\dagger\|,$$

by continuity, compactness and the Weierstrass Theorem. We can pick $\rho<\frac{\delta(r,R)}{R\underline{\sigma}}$. □

**Proof of Proposition 5:** For any $\theta\in\mathbb{R}^{d_\theta}$, we have:

$$\begin{aligned}
Q(\theta)-Q(\theta^\dagger)&=\frac{1}{2}\left(g(\theta)'Wg(\theta)-g(\theta^\dagger)'Wg(\theta^\dagger)\right)\\
&=\frac{1}{2}\left(g(\theta)+g(\theta^\dagger)\right)'W\left(g(\theta)-g(\theta^\dagger)\right)\\
&=\frac{1}{2}\left(g(\theta)+g(\theta^\dagger)\right)'W\overline{G}(\theta)(\theta-\theta^\dagger)\\
&=\frac{1}{2}(\theta-\theta^\dagger)'\overline{G}(\theta)'W\overline{G}(\theta)(\theta-\theta^\dagger)-g(\theta^\dagger)'W\overline{G}(\theta)(\theta-\theta^\dagger),
\end{aligned}$$

the first term in the last display matches the one in the proof of Proposition 1. Note that $g(\theta^\dagger)'WG(\theta^\dagger)=0$ and $\|G(\theta^\dagger)-\overline{G}(\theta)\|\leq L\|\theta-\theta^\dagger\|$, together these allow to bound the second term:

$$\|g(\theta^\dagger)'W\overline{G}(\theta)(\theta-\theta^\dagger)\|=\|g(\theta^\dagger)'W[\overline{G}(\theta)-G(\theta^\dagger)](\theta-\theta^\dagger)\|\leq\overline{\lambda}_W^{1/2}L\sqrt{\varphi}\|\theta-\theta^\dagger\|^2.$$

Let $C_2 = 1/2\frac{\underline{\rho}^2\underline{\sigma}^2}{\overline{\sigma}^2\overline{\lambda}_W}$ and $C_3 = 1/2\overline{\sigma}^2\overline{\lambda}_W$, as in the proof of Proposition 1. Take $C_4 = \overline{\lambda}_W^{1/2}L$, this yields (2):

$$(C_2 - C_4\sqrt{\varphi})\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger) \leq (C_3 + C_4\sqrt{\varphi})\|\theta - \theta^\dagger\|^2.$$

For (1), we have $\partial_\theta Q(\theta) = G(\theta)'Wg(\theta)$ and $G(\theta^\dagger)'Wg(\theta^\dagger) = 0$, so that:

$$\partial_\theta Q(\theta) = G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger) + \{G(\theta) - G(\theta^\dagger)\}'Wg(\theta^\dagger).$$

Apply the reverse triangular inequality to find:

$$\begin{aligned}\|\partial_\theta Q(\theta)\| &\geq \rho\underline{\sigma}\|\theta - \theta^\dagger\| - \sqrt{\varphi\overline{\lambda}_W}L\|\theta - \theta^\dagger\| \\ &= \left(\rho\underline{\sigma} - \sqrt{\varphi\overline{\lambda}_W}L\right)\|\theta - \theta^\dagger\|,\end{aligned}$$

where $L$ is the Lipschitz constant of $G$. Finally, (1') can be derived from (1) and (2) assuming $(\rho\underline{\sigma} - \sqrt{\varphi\overline{\lambda}_W}L) > 0$. □

**Proof of Proposition 6:** We first prove (1). (a) ⇒ Assumption 2 (a) is immediate. Under (c), $G(\theta) = \partial^2_{\theta,\theta'}F(\theta)$ is symmetric and strictly positive definite so (b) holds. Suppose (b) holds, then $\overline{G}(\theta_1, \theta_2) = U\{\int_0^1 S(\omega\theta_1 + (1-\omega)\theta_2)d\omega\}V$ where $\int_0^1 S(\omega\theta_1 + (1-\omega)\theta_2)d\omega$ is symmetric. Concavity of the smallest positive eigenvalue on the set of positive definite matrices, and Jensen's inequality imply: $\lambda_{\min}[\int_0^1 S(\omega\theta_1+(1-\omega)\theta_2)d\omega] \geq \int_0^1 \lambda_{\min}[S(\omega\theta_1+(1-\omega)\theta_2)]d\omega \geq \min_{\theta\in\Theta}\lambda_{\min}[S(\theta)] > 0$, by positive definiteness and continuity of $S(\cdot)$. Finally,

$$\sigma_{\min}[\overline{G}(\theta_1, \theta_2)] \geq \sigma_{\min}(U)\sigma_{\min}(V)\min_{\theta\in\Theta}\lambda_{\min}[S(\theta)] > \underline{\lambda}_S\underline{\sigma}_U\underline{\sigma}_V > 0,$$

taking $\underline{\sigma}_U\underline{\sigma}_V$ to be smallest singular values of $U, V$. Hence (a) holds.

For (2), note that $g(\theta_1) - g(\theta_2) = \overline{G}(\theta_1, \theta_2)(\theta_1 - \theta_2)$, using Lemma A1. With condition (a), we have $g(\theta_1) - g(\theta_2) = 0 \Leftrightarrow \theta_1 = \theta_2$, i.e. $g(\cdot)$ is one-to-one.

For (3), $g(\cdot)$ is one-to-one, take $\phi(\cdot) = g^{-1}(\cdot)$, one-to-one, and $\psi = I_d - \theta^\dagger$, we get that $h(\theta) = \theta - \theta^\dagger$ is linear, the associated GMM loss is strictly quadratic; i.e. strongly convex. □

**Proof of Proposition 7:** 1) Under Assumption 2 (a), $G$ has full rank for all $\theta \in \mathbb{R}^{d_\theta}$. Take $u \in \mathcal{U}$, let $\theta = h(u)$, the chain rule implies that $\partial_u g \circ h(u) = \partial_\theta g \circ h(u)\partial_u h(u)$ has full

rank for all $u \in \mathcal{U}$. Then, we have:

$$\begin{aligned}
&\int_0^1 G \circ h(\omega u + (1-\omega)u)\partial_u h(\omega u + (1-\omega)u^\dagger)d\omega \\
&= \int_0^1 G(\omega\theta + (1-\omega)\theta^\dagger)d\omega \partial_u h(u^\dagger) \\
&+ \int_0^1 G(\omega\theta + (1-\omega)\theta^\dagger)[\partial_u h(\omega u + (1-\omega)u^\dagger) - \partial_u h(u^\dagger)]d\omega \\
&+ \int_0^1 [G \circ h(\omega u + (1-\omega)u) - G(\omega\theta + (1-\omega)\theta^\dagger)]\partial_u h(\omega u + (1-\omega)u^\dagger)d\omega,
\end{aligned}$$

using Weyl's inequality and a minoration of the singular value for a matrix product, we get:

$$\sigma_{\min}[\int_0^1 \partial_u h(\omega u + (1-\omega)u^\dagger)G \circ h(\omega u + (1-\omega)u)d\omega] \geq \underline{\sigma}_h \underline{\sigma} - C_1\overline{\sigma} - C_2 L\overline{\sigma}_h,$$

which is strictly positive under the stated condition. After the change of variable, the Assumption 2 (a) holds if:

$$\partial_u h(u)'G(g(u))'W\Big\{\int_0^1 \partial_u h(\omega u + (1-\omega)u^\dagger)G \circ h(\omega u + (1-\omega)u)d\omega\Big\},$$

has singular values bounded below by a strictly positive term, which is the case for $C_1, C_2$ bounded as in the Proposition statement. In particular, when $h$ is affine, $C_1 = C_2 = 0$ and $0 < \underline{\sigma}_h = \sigma_{\min}[A] \leq \sigma_{\max}[A] \leq \overline{\sigma}_h < \infty$, so that the condition is automatically satisfied.
2) Take $u$, let $\theta = h(u)$, since Assumption 2 (b) holds, we have:

$$\|G \circ h(u)'W[g \circ h(u) - g \circ h(u^\dagger)]\| \geq \rho\underline{\sigma}\|h(u) - h(u^\dagger)\| \geq \rho\underline{\sigma}\mu\|u - u^\dagger\|.$$

Using the bounds on the Jacobian of $\partial_u h$, we get the desired result:

$$\|\partial_u h(u)'G \circ h(u)'W[g \circ h(u) - g \circ h(u^\dagger)]\| \geq \underline{\sigma}_h\rho\underline{\sigma}\mu\|u - u^\dagger\|.$$

□

**Proof of Lemma B7** 1) (SI') implies $\|g(\theta_1) - g(\theta_2))\| \geq \mu/[\overline{\sigma}\overline{\lambda}_W]\|\theta_1 - \theta_2\|$, where $\mu/[\overline{\sigma}\overline{\lambda}_W] > 0$. 2) Take $\|v\| = 1$, $\theta_2 = \theta_1 + \varepsilon v$ and let $\varepsilon \to 0$ in (SI') to find: $\|G(\theta_1)'WG(\theta_1)v\| \geq \mu\|v\|$. Apply the min-max theorem for singular values (Bhatia, 2013, p75) to find that $\sigma_{\min}[G(\theta)'WG(\theta)] \geq \mu$ for all $\theta \in \mathbb{R}^{d_\theta}$. Since $G(\theta)'WG(\theta)$ is positive semidefinite, singular and eigenvalues coincide, which implies that $\underline{\lambda}_P = \mu > 0$. Then Assumption 1 (iii), (vi)

implies $\overline{\lambda}_P \leq \overline{\sigma}^2 \overline{\lambda}_W$. Uniform consistency then yield the desired result (Lemma A3). □

**Proof of Lemma B8:** For just-identified models, (SI) implies (SI') as long as the weighting matrix is finite and invertible. Indeed, (SI) implies $\sigma_{\min}[G_1(\theta)] \geq \mu_1 > 0$ so that $\|G_1(\theta_1)'W[g_1(\theta_1) - g_1(\theta_2)]\| \geq \mu_1^2 \lambda_{\min}(W_1)\|\theta_1 - \theta_2\|$ so that (SI') holds using $W_1$ as weighting matrix. Some calculations imply that, by construction of $\tilde{W}$:

$$\|G(\theta_1)'\tilde{W}(\lambda)[g(\theta_1) - g(\theta_2)]\| \geq \mu_1^2 \lambda_{\min}(W_1)\|\theta_1 - \theta_2\| - \lambda\Big[\|G_2(\theta_1)'W_{21}[g_1(\theta_1) - g_1(\theta_2)]\| + \|G_1(\theta_1)'W_{12}[g_2(\theta_1) - g_2(\theta_2)]\| + \|G_2(\theta_1)'W_{22}[g_2(\theta_1) - g_2(\theta_2)]\|\Big],$$

where $W_2$ is the lower block of $W$ corresponding to $g_2$, $W_{12}$ and $W_{21}$ are the top right and bottom left corners of $W$, respectively. Let $L_1, L_2$ be the Lipschitz constants of $g_1, g_2$ respectively we can conservatively bound the last terms with:

$$\lambda\|G\|_\infty\|W\|_\infty[L_1 + 2L_2]\|\theta_1 - \theta_2\| < \mu_1^2\lambda_{\min}(W_1)\|\theta_1 - \theta_2\|,$$

for any $0 \leq \lambda \leq \lambda^\star < \mu_1^2\lambda_{\min}(W_1)/(\|G\|_\infty\|W\|_\infty[L_1 + 2L_2])$, where $\|\cdot\|_\infty$ denotes the $\ell_\infty$ norm. □

**Proof of Proposition B8:** First, we prove (1). (a) ⇒ Assumption 2 (a) is immediate. Suppose (b) holds, take any $\theta, \theta_1, \theta_2 \in \mathbb{R}^{d_\theta}$, then $G(\theta)'W\overline{G}(\theta_1, \theta_2) = V'S(\theta)U'WU\int_0^1\{S(\omega\theta_1 + (1-\omega)\theta_2)\}d\omega V$. By assumption, $V'S(\theta)$ and $U'WU$ have full rank. As in the proof of Proposition 6, $\int_0^1\{S(\omega\theta_1 + (1-\omega)\theta_2)\}d\omega$ has full rank for any $\theta_1, \theta_2$, and $V$ is invertible. Hence, $S(\theta)U'WU\int_0^1\{S(\omega\theta_1 + (1-\omega)\theta_2)\}d\omega V$ is invertible, $U$ has full rank so that $G(\theta)'W\overline{G}(\theta_1, \theta_2)$ has full rank for all $\theta, \theta_1, \theta_2$.

For part (2), take any $\theta_1, \theta_2, \theta_3$. Suppose $G(\theta_1)'Wg(\theta_2) = G(\theta_1)'Wg(\theta_3)$, apply Lemma A1 to find $G(\theta_1)'W\overline{G}(\theta_2, \theta_3)(\theta_2 - \theta_3) = 0 \Rightarrow \theta_2 = \theta_3$ under condition (a). □

**Proof of Proposition B9:** 1) We'll proceed similarly to the proof of Proposition 7:

$$\int_0^1 \partial_u' h(\omega u + (1-\omega)u^\dagger) G \circ h(\omega u + (1-\omega)u^\dagger)' d\omega W G \circ h(u) \partial_u h(u)$$
$$= \partial_u' h(u^\dagger) \int_0^1 G(\omega\theta + (1-\omega)\theta^\dagger)' d\omega W G(\theta) \partial_u h(u)$$
$$+ \int_0^1 [\partial_u h(\omega u + (1-\omega)u^\dagger) - \partial_u h(u^\dagger)]' G(\omega\theta + (1-\omega)\theta^\dagger)' d\omega W G(\theta) \partial_u h(u)$$
$$+ \int_0^1 \partial_u' h(\omega u + (1-\omega)u^\dagger) [G \circ h(\omega u + (1-\omega)u^\dagger) - G(\omega\theta + (1-\omega)\theta^\dagger)]' d\omega W G(\theta) \partial_u h(u).$$

As before, we get: $\sigma_{\min}[\int_0^1 \partial_u' h(\omega u + (1-\omega)u^\dagger) G \circ h(\omega u + (1-\omega)u^\dagger)' d\omega W G \circ h(u) \partial_u h(u)] \geq \underline{\sigma}\underline{\sigma}_h^2 - C_1 \overline{\sigma}_h \overline{\sigma}^2 \overline{\lambda}_W - C_2 L \overline{\sigma}_h^2 \overline{\sigma} \overline{\lambda}_W$ which is positive under the stated condition. As before, for $h$ affine we have $C_1 = C_2 = 0$ so that the condition holds for $A$ finite and invertible. 2) The proof is the same as in the just-identified case. □

Supplement to

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

Jean-Jacques Forneron* Liang Zhong†

July 11, 2025

This Supplemental Material consists of Appendices C, D, E, F, and G 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, Website: https://samzl1.github.io/.

# Appendix C Local Convergence Results

The following considers local convergence under correct specification, where $g(\theta^\dagger) = 0$, and misspecification, where $g(\theta^\dagger) \neq 0$. These results highlight how several quantities affect the estimation. Here, Assumptions 1, 3 are sufficient to study local convergence. Throughout it is assumed that $\hat{\theta}_n$ and $Q_n(\hat{\theta}_n)$ are consistent for $\theta^\dagger$ and $Q(\theta^\dagger)$.

**Proposition C10** (Correctly Specified)**.** *If Assumptions 1, 3 hold, then for $\gamma \in (0,1)$ small enough, with probability approaching 1, there exist $0 < R_n \leq R_G$ and $\tilde{\gamma} \in (0,1)$ such that :*

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

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

This result is comparable to those found for non-linear systems of equations (e.g. Dennis and Schnabel, 1996; Nocedal and Wright, 2006, Ch11), with some notable differences. First, if the model is over-identified, $\overline{g}_n(\theta) = 0$ does not have a solution, and standard results do not apply. Second, the area of local convergence $R_n$ is tied to a) the choice of tuning parameter $\gamma$, b) the size of the moments at the solution $\overline{g}_n(\hat{\theta}_n)$, c) the choice of weighting matrix. For GN, the area of local convergence $R_n = \min(R_G, \tilde{R}_n)$ 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 in finite samples. For correctly specified models $\overline{g}_n(\hat{\theta}_n) = o_p(1)$ implies $\tilde{R}_n \xrightarrow{p} \tilde{R} = (1-\tilde{\gamma}/\gamma)\underline{\sigma}/(\sqrt{\kappa_W}L) > 0$. Note that for GN, Proposition C10 holds for any choice of $\gamma \in (0,1)$. This is typically not the case for other choices of $P_{k,n}$: GD requires $0 < \gamma < [\overline{\lambda}_W\overline{\sigma}^2]^{-1}$ to be sufficiently small. GD and GN iterations use the same inputs $G_n$ and $\overline{g}_n$, but the latter converges more quickly.

The expression for $\tilde{R}_n$ illustrates that the choice of weighting matrix $W_n$ matters. Equal weighting, $W = I_d$, has $\kappa_W = 1$ whereas an ill-conditioned matrix has $\kappa_W \gg 1$.

In applications, $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ can be relatively large so that misspecification becomes a concern. Understanding the robustness of Proposition C10 to non-negligible deviations from $Q(\theta^\dagger) = 0$ is then empirically relevant. The following considers models where the quantity:

$$Q_n(\hat{\theta}_n) \xrightarrow{p} Q(\theta^\dagger) := \varphi/2 > 0$$

does not vanish asymptotically which implies that $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ matters for local convergence, even in large samples. Since $G_n$ cannot be full rank at $\theta = \hat{\theta}_n$ when the model is both just-identified and misspecified, the results presented here solely consider over-identified models.[1]

**Proposition C11** (Misspecified)**.** *Suppose Assumptions 1, 3 hold, and $\varphi$ is such that:*

$$\sqrt{\varphi} < \frac{\underline{\sigma}^2 \sqrt{\underline{\lambda}_W}}{L \kappa_W \kappa_P}, \tag{C.2}$$

*where $\kappa_P = \overline{\lambda}_P / \underline{\lambda}_P$. For $\gamma \in (0,1)$ small enough, there exists $\tilde{\gamma} \in (0, \gamma)$, such that, with probability approaching 1, for any $\|\theta_0 - \hat{\theta}_n\| \leq R_n$, and all $k \geq 0$:*

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

*with the same $R_n$ found in Proposition C10; such that $plim_{n \to \infty} R_n = R > 0$ when (C.2) holds.*

The result shows that under 'moderate' amounts of misspecification, the area of local convergence is asymptotically non-empty. For GN, the condition simplifies to: $\sqrt{\varphi} < \frac{\underline{\sigma}^2 \sqrt{\underline{\lambda}_W}}{L \sqrt{\kappa_W}}$. Several terms restrict the amount of misspecification in (C.2): $\underline{\sigma}$, $L$, and the pair $\underline{\lambda}_W, \kappa_W$. The first measures local identification strength, the second non-linearity, and the latter comes from the weighting matrix. For linear models, $L = 0$, the conditions reads $\sqrt{\varphi} < +\infty$; misspecification only matters in nonlinear problems with $L > 0$. Note that the area of local convergence is asymptotically smaller than in Proposition C10.

**Proof of Proposition C10 (Gauss-Newton).** Take $\theta_k \in \mathbb{R}^{d_\theta}$, the update (1) can be re-written as:

$$\begin{aligned} \theta_{k+1} - \hat{\theta}_n = & \Big( I_d - \gamma P_{k,n} G_n(\theta_k)' W_n G_n(\theta_k) \Big) (\theta_k - \hat{\theta}_n) \\ & - \gamma P_{k,n} G_n(\theta_k)' W_n [\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)]. \end{aligned} \tag{C.3}$$

[1]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.

For GN, $P_{k,n}G_n(\theta_k)'W_nG_n(\theta_k) = I_d$ so that we have:

$$\begin{aligned}\theta_{k+1} - \hat{\theta}_n =& (1-\gamma)(\theta_k - \hat{\theta}_n) \\ & - \gamma P_{k,n}G_n(\theta_k)'W_n[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)] \\ & - \gamma P_{k,n}[G_n(\theta_k) - G_n(\hat{\theta}_n)]'W_n\overline{g}_n(\hat{\theta}_n),\end{aligned} \tag{C.3'}$$

using the first-order condition $G_n(\hat{\theta}_n)'W_n\overline{g}_n(\hat{\theta}_n) = 0$. From Assumption A1, 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 $P_{k,n}$ is well defined and bounded. Since $G_n$ is Lipschitz continuous with constant $L \geq 0$:

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

We also have:

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

Combine these two inequalities into (C.3') to find:

$$\begin{aligned}&\|\theta_{k+1} - \hat{\theta}_n\| \\ &\leq \left(1 - \gamma + \gamma\left[\underline{\sigma}^{-1}\sqrt{\overline{\lambda}_W/\underline{\lambda}_W}L\|\theta_k - \hat{\theta}_n\| + \underline{\sigma}^{-2}(\sqrt{\overline{\lambda}_W}/\underline{\lambda}_W)L\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right]\right)\|\theta_k - \hat{\theta}_n\|.\end{aligned} \tag{C.3"}$$

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 (C.1). □

**Proof of Proposition C10 (General Case).** Take $\theta_k \in \mathbb{R}^{d_\theta}$, the update (1) can be re-written as:

$$\begin{aligned}\theta_{k+1} - \hat{\theta}_n =& \Big(I_d - \gamma P_{k,n} G_n(\theta_k)' W_n G_n(\theta_k)\Big)(\theta_k - \hat{\theta}_n) \\ & - \gamma P_{k,n} G_n(\theta_k)' W_n [\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)].\end{aligned} \tag{C.3}$$

Taking norms on both sides this identity yields:

$$\begin{aligned}\|\theta_{b+1} - \hat{\theta}_n\| \leq & \sigma_{\max}\Big[I_d - \gamma P_{k,n} G_n(\theta_k)' W_n G_n(\theta_k)\Big] \|\theta_b - \hat{\theta}_n\| \\ & + \gamma \|P_{k,n} G_n(\theta_k)' W_n [\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)]\|,\end{aligned} \tag{C.3'}$$

where $\sigma_{\max}$ returns the largest singular value. We will now bound each of these two terms. First, note that $\sigma_{\max}[I_d - \gamma P_{k,n} G_n(\theta_k)' W_n G_n(\theta_k)] = \sigma_{\max}[I_d - \gamma P_{k,n}^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_{k,n}^{1/2}] = \max_{j=1,\ldots,d} |\lambda_j[I_d - \gamma P_{k,n}^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_{k,n}^{1/2}]|$, where $\lambda_j$ 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:

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

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:

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

so that we are only concerned with the upper bound. From Assumption A1, $\|\theta - \hat{\theta}_n\| \leq R_G \Rightarrow \sigma_{\min}[G_n(\theta)] \geq \underline{\sigma}$. Combine with the bound for $\gamma$ to find:

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

for any choice of $\gamma \in (0, [\overline{\lambda}_P \overline{\lambda}_W \overline{\sigma}^2]^{-1})$. For the second term in (C.3), using the identity $G_n(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) = 0$ and Lemma A1:

$$\begin{aligned}P_{k,n} G_n(\theta_k)' W_n [\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)] =& P_{k,n} G_n(\theta_k)' W_n [\overline{G}_n(\theta_k) - G_n(\theta_k)](\theta_k - \hat{\theta}_n) \\ & + P_{k,n} [G_n(\theta_k) - G_n(\hat{\theta}_n)]' W_n \overline{g}_n(\hat{\theta}_n),\end{aligned}$$

where $\overline{G}_n(\theta_k) = \int_0^1 \{G_n(\omega\theta_k + (1-\omega)\hat{\theta}_n)\}d\omega$. Since $G_n$ is Lipschitz continuous with constant $L \geq 0$:

$$\begin{aligned}\|(C.3')\| &\leq (1 - \gamma\underline{\lambda}_P\underline{\lambda}_W\underline{\sigma}^2)\|\theta_b - \hat{\theta}_n\| + \gamma\overline{\lambda}_P\overline{\lambda}_W\overline{\sigma}L\|\theta_b - \hat{\theta}_n\|^2 + \gamma\overline{\lambda}_P\overline{\lambda}_W^{1/2}L\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\|\theta_b - \hat{\theta}_n\| \\ &= \left(1 - \gamma\underline{\lambda}_P\underline{\lambda}_W\underline{\sigma}^2 + \gamma\left[\overline{\lambda}_P\overline{\lambda}_W\overline{\sigma}L\|\theta_b - \hat{\theta}_n\| + \overline{\lambda}_P\overline{\lambda}_W^{1/2}L\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right]\right)\|\theta_b - \hat{\theta}_n\|.\end{aligned}$$

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:

$$\|\theta_k - \hat{\theta}_n\| \leq \frac{\gamma\underline{\lambda}_P\underline{\lambda}_W\underline{\sigma}^2 - \tilde{\gamma}}{\gamma c_1} - \frac{c_2}{c_1}\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} := \tilde{R}_n. \tag{C.4}$$

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 Proposition C11 (Gauss-Newton):** The proof is similar to the proof of Proposition C10 with the difference that $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \xrightarrow{p} \sqrt{\varphi/2} > 0$. The radius is convergence is asymptotically non-zero for $0 < \tilde{\gamma} < \gamma < 1$ small enough if: $\sqrt{\varphi} < \frac{\sigma^2\sqrt{\underline{\lambda}_W}}{L\sqrt{\kappa_W}}$. □

**Proof of Proposition C11 (General Case):** The proof is similar to the proof of Proposition C10 with the difference that $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \xrightarrow{p} \sqrt{\varphi/2} > 0$. The radius is convergence is asymptotically non-zero for $0 < \tilde{\gamma} < \gamma < 1$ small enough if: $\sqrt{\varphi} < \frac{\sigma^2\sqrt{\underline{\lambda}_W}}{L\kappa_W\kappa_P}$, where $\kappa_P = \overline{\lambda}_P/\underline{\lambda}_P$. □

# Appendix D Commonly used methods and their properties

## D.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.[2] Table D7 was constructed by reading through the main text, supplemental material, and all available replication codes of the selected papers.

Table D7: American Economic Review 2016-2018: GMM and related empirical estimations

| Method | # Papers | # Parameters (p) | Data available |
|---|---|---|---|
| Nelder-Mead - one starting value | 7 | 2,6 (×2),11,13 (×2),147 | 3 |
| Simulated Annealing + Nelder-Mead | 2 | 4,13 | 1 |
| Nelder-Mead - multiple starting values | 2 | ?,6 | 1‡ |
| Pattern Search | 2 | 6,147 | 1† |
| Genetic Algorithm | 2 | 9,14 | 1 |
| Simulated Annealing | 2 | 4,13 | 2† |
| MCMC | 1 | 15 | 1 |
| Grid Search | 1 | 5 | 1 |
| No description | 3 | - | - |
| Stata/Mata default | 4 | 3,6 (×2),38 | 3⋆ |

**Legend:** # Parameters correspond to the size of the largest specification. Data avail. 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 reported to include data requires to download the PSID dataset separately. †: two papers in total also rely separately on Nelder-Mead, so they are also reported under Nelder-Mead. ‡: one paper provides data without codes.

**Survey results:** Table D7 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 7 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 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.

[2]The search function in JSTOR allows to search for keywords within the title, abstract, main text, and supplemental material of a paper. Further screening ensures 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."

As mentioned in the introduction, although convex optimizers such as (stochastic) gradient-descent and quasi-Newton methods are commonly used to solve large scale convex minimization problems, they are virtually absent from the survey. Overall, 11 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 provides a brief overview of the properties of the main Algorithms found used in Table D7.

### D.2 A brief summary of the Algorithms' properties

The following briefly discussed the properties of four algorithms from Table D7: Nelder-Mead, Grid Search, Multi-Start, and Simulated Annealing. Further discussion, descriptions, and references can be found in Appendix G.

Nelder-Mead (NM) is the most popular method in the survey, it can be used even if $Q_n$ is discontinuous. Its convergence properties, which measure its ability to find valid estimates, are somewhat limited however. For some smooth convex problems, it can be shown to converge to values that are neither locally nor globally optimal. The grid-search converges to the solution under weak conditions, unlike NM. It is very slow, however, and often not practical when estimating three or more coefficients. Simulated annealing (SA) is not deterministic. Still it converges, in probability, under weak conditions to the solution. Albeit, the convergence is predicted to be slower than grid search. A common approach to improve the convergence of a given algorithm is to combine it with multiple starting values. The required number of starting values depends on $Q_n$ and the choice of algorithm. Andrews (1997) provides an asymptotically valid stopping rule for correctly specified GMM models.

When $Q_n$ is strongly convex, several gradient-based methods discussed below are rapidly, globally convergent and do not suffer from a curse of dimensionality. This implies that it is possible to estimate a large number of parameters in a reasonable amount of time. Similar convergence properties are derived in this paper, under rank conditions instead of convexity.

## Appendix E 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 F Additional Empirical Results

## F.1 Demand for Cereal

Table F8: Demand for Cereal: GN with different learning rates

| | | STDEV | | | | INCOME | | | | crash | objs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 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 where $[0, 10] \times \cdots \times [0, 10]$ for standard deviations and $[-10, 10] \times \cdots \times [-10, 10]$ for income coefficients. 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. crash: 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.

## F.2 Impulse Response Matching

The following tables report results for GN using a range of tuning parameters $\gamma$. Since Assumption 2 does not hold towards the lower bound for $\eta, \nu$, GN alone can crash and/or fail to converge. Following Forneron (2023), we can introduce a global step:

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

where the sequence $(\theta^k)_{k\geq 0}$ is predetermined and dense in $\Theta$. The results rely on the Sobol sequence, independently randomized for each of the 50 starting values.[3] Results are reported with and without the global step. Also, the former implements error-handling (try-catch).

[3]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 F9: GN with different learning rates

| | | WITHOUT REPARAMETERIZATION | | | | | | WITH REPARAMETERIZATION | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\eta$ | $\nu$ | $\rho_s$ | $\sigma_s$ | objs | crash | $\eta$ | $\nu$ | $\rho_s$ | $\sigma_s$ | objs | crash |
| TRUE | est | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | - | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | - |
| $\gamma =$ | | GN WITHOUT GLOBAL STEP | | | | | | | | | | | |
| 0.1 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 1 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 9 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 0.2 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 1 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 16 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 0.4 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 1 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 20 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 0.6 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 1 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 21 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 0.8 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 1 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 27 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 1.0 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 10 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 29 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| | | GN WITH GLOBAL STEP | | | | | | | | | | | |
| 0.1 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 0.2 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 0.4 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 0.6 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 0.8 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| 1.0 | avg | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 | 0.30 | 0.29 | 0.39 | 0.17 | 4.65 | 0 |
| | std | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| lower bound | | 0.05 | 0.01 | -0.95 | 0.01 | - | - | 0.05 | 0.01 | -0.95 | 0.01 | - | - |
| upper bound | | 0.99 | 0.90 | 0.95 | 12 | - | - | 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 estimation and reparameterization.

## F.3 Sensitivity of Numerical Derivatives

In some of the applications, the moments are computed using numerical routines, using e.g. fixed point iterations, which evaluate the moments up to some tolerance level $\eta$. This can affect the optimization as the precision of first and second order numerical derivatives can be sensitive to this approximation. The following gives an brief overview for a scalar moment and parameter. Suppose we can only compute $g_\eta(\theta)$ such that $|g_\eta(\theta) - g(\theta)| \leq \eta$, for all $\theta$. In order to implement a derivative-based optimizer, the derivative $\partial_\theta g(\theta)$ is approximated by finite differences: $G_\epsilon(\theta) = \frac{1}{\epsilon}(g(\theta+\epsilon) - g(\theta))$ with some tuning parameter $\epsilon$, the default in

R is $\epsilon = 6 \cdot 10^{-6}$. The approximation error for this derivative is at most: $|G(\theta) - G_\epsilon(\theta)| \leq \epsilon L$ where $L$ is the Lipschitz constant of $G$. Since $g$ itself is not available, a further approximation is needed: $G_{\eta,\epsilon}(\theta) = \frac{1}{2\epsilon}(g_\eta(\theta + \epsilon) - g_\eta(\theta - \epsilon))$. This has a larger approximation error: $|G_{\eta,\epsilon}(\theta) - G(\theta)| \leq \epsilon L + \frac{\eta}{\epsilon}$.

In the BLP application, the fixed-point tolerance level is set to $\eta = 10^{-12}$, this yields an approximation error of order $10^{-6}$ for the Jacobian $G_n$, when the inner loop did not terminate because of the limit on the number of iterations (set at 2000). BFGS further approximates second derivatives using finite differences. The second-order derivative can be computed as: $\partial_\theta G_\epsilon(\theta) = \frac{1}{\epsilon^2}[g(\theta + \epsilon) + g(\theta - \epsilon) - 2g(\theta)] = \frac{1}{\epsilon}[G_\epsilon(\theta + \epsilon) - G_\epsilon(\theta - \epsilon)]$ which has an approximation error: $|\partial_\theta G_\epsilon(\theta) - \partial_\theta G(\theta)| \leq L_2\epsilon$, where $L_2$ is the Lipschitz constant of $\partial_\theta G$. Again, since $g$ is not available we need a further approximation error: $\partial_\theta G_{\eta,\epsilon}(\theta) = \frac{1}{\epsilon^2}[g_\eta(\theta + \epsilon) + g_\eta(\theta - \epsilon) - 2g_\eta(\theta)]$ which has an error of size $|\partial_\theta G_{\eta,\epsilon}(\theta) - \partial_\theta G(\theta)| \leq \epsilon L_2 + \frac{\eta}{\epsilon^2}$. In the BLP application, $\eta = 10^{-12}$ and $\epsilon^{-2} = 1/36 \cdot 10^{12}$ are of the same order of magnitude so that the approximation error, for second-order derivatives is likely to be large.

# Appendix G Additional Material for Section D.2

## G.1 General overview of Algorithms properties

The following describes three of the algorithms in Table D7: 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 G.

**Notation:** $Q_n$ is a continuous objective function to be minimized over $\Theta$, 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 D7 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.[4] 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 $k$ 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

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

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.

## G.2 Implementation of the algorithms

**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 $\theta_c$ 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 $\theta_{p+1}$ with $\theta_r$, re-order the points, compute the new $\theta_c$, 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 $\theta_{p+1}$ with $\theta_r$. 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 $\theta_e$ replaces $\theta_{p+1}$. Otherwise, $\theta_r$ replaces $\theta_{p+1}$. Once $\theta_{p+1}$ is replaced, re-order the points, compute the new $\theta_c$, 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 $Q_n(\theta_s) < Q_n(\theta_{p+1})$, $\theta_s$ replaces $\theta_{p+1}$, then re-order the points, compute the new $\theta_c$, 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 $\theta_c$, and go to **R**.

Clearly, the choice of initial simplex can affect the convergence of the algorithm. Typically, one provides a starting value $\theta_1$ 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 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.[5] 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:

$$\underline{\lambda}\|\theta - \hat{\theta}_n\|^2 \leq Q_n(\theta) - Q_n(\hat{\theta}_n) \leq \overline{\lambda}\|\theta - \hat{\theta}_n\|^2, \tag{G.5}$$

i.e. $Q_n$ is locally strictly convex.[6] 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 $\|\tilde{\theta}_k - \hat{\theta}_n\| \leq [\overline{\lambda}/\underline{\lambda}]^{1/2}(\inf_{1\leq\ell\leq k} \|\theta_\ell - \hat{\theta}_n\|)$.

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 $\varepsilon_2$ 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.

[5] 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.

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

**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 $\eta_\ell$ proportional to $T_\ell$.

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 $\theta_\ell$, regardless of $u_\ell$. Bélisle (1992) gave sufficient condition for $\tilde{\theta}_k \overset{a.s.}{\to} \hat{\theta}_n$ 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 $T_\infty = +\infty$, $f_\infty$ 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 $\theta_k$ is outside the $\varepsilon_1$-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 $\varepsilon_1$, $\varepsilon_2$ 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 (G.5) 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 $\varepsilon_2$ 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 $T_k = T_1/\log(k)$.[7]

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