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

Jean-Jacques Forneron, Boston University

March 2023

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

- **The challenge is the estimation itself...**

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

- **The challenge is the estimation itself...**
  In finite samples, the objective function can be:

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

- **The challenge is the estimation itself...**
  In finite samples, the objective function can be:
  1. **noisy:** sampling uncertainty

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

- **The challenge is the estimation itself...**
  In finite samples, the objective function can be:
  1. **noisy:** sampling uncertainty
  2. **non-smooth:** simulated method of moments, robust estimation, ...

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

- **The challenge is the estimation itself...**
  In finite samples, the objective function can be:
  1. **noisy:** sampling uncertainty
  2. **non-smooth:** simulated method of moments, robust estimation, ...
  3. **non-convex:** common in structural work

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

- **The challenge is the estimation itself...**
  In finite samples, the objective function can be:
  1. **noisy:** sampling uncertainty
  2. **non-smooth:** simulated method of moments, robust estimation, ...
  3. **non-convex:** common in structural work

- **This paper:**

  smoothed Gauss-Newton (sGN) method for non-smooth GMM

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

- **The challenge is the estimation itself. . .**
  In finite samples, the objective function can be:
  1. **noisy:** sampling uncertainty
  2. **non-smooth:** simulated method of moments, robust estimation, . . .
  3. **non-convex:** common in structural work

- **This paper:**

  smoothed Gauss-Newton (sGN) method for non-smooth GMM

  1. non-asymptotic analysis: optimization and statistical properties

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

- **The challenge is the estimation itself...**
  In finite samples, the objective function can be:
  1. **noisy:** sampling uncertainty
  2. **non-smooth:** simulated method of moments, robust estimation, ...
  3. **non-convex:** common in structural work

- **This paper:**

  smoothed Gauss-Newton (sGN) method for non-smooth GMM

  1. non-asymptotic analysis: optimization and statistical properties
  2. local/global convergence using **only econometric assumptions**

Motivation

- GMM estimation is an important part of the empirical toolkit:

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

  - asymptotics well understood (e.g. Newey and McFadden, 1994)

- **The challenge is the estimation itself. . .**
  In finite samples, the objective function can be:
  1. **noisy:** sampling uncertainty
  2. **non-smooth:** simulated method of moments, robust estimation, . . .
  3. **non-convex:** common in structural work

- **This paper:**

  smoothed Gauss-Newton (sGN) method for non-smooth GMM

  1. non-asymptotic analysis: optimization and statistical properties
  2. local/global convergence using **only econometric assumptions**
  3. after a finite number of iterations, converges exponentially fast

Overview of the Problem and Literature

- **Optimization:** we want after $b \geq 1$ iterations:

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

How large does b have to be? Depends on the objective function:

Overview of the Problem and Literature

- **Optimization:** we want after $b \geq 1$ iterations:

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

How large does b have to be? Depends on the objective function:

  - **Smooth + Strongly Convex**: $b = O(|\log(\text{err})|)$
    using gradient-descent, (quasi)-Newton, etc.

Overview of the Problem and Literature

- **Optimization:** we want after $b \geq 1$ iterations:

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

How large does b have to be? Depends on the objective function:

  - **Smooth + Strongly Convex**: $b = O(|\log(\text{err})|)$ using gradient-descent, (quasi)-Newton, etc.
  - **Smooth + Non Convex**: $b = O(\text{err}^{-d/s})$ $s =$ smoothness, $d = \dim(\theta)$: curse of dimensionality

Overview of the Problem and Literature

- **Optimization:** we want after $b \geq 1$ iterations:

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

How large does b have to be? Depends on the objective function:

  - **Smooth + Strongly Convex**: $b = O(|\log(\text{err})|)$
    using gradient-descent, (quasi)-Newton, etc.
  - **Smooth + Non Convex**: $b = O(\text{err}^{-d/s})$
    $s =$ smoothness, $d = \dim(\theta)$: curse of dimensionality

- **Global optimizers:**
  1. heuristic (no guarantees), e.g. *genetic algorithms*
  2. or slow, e.g. *simulated annealing* $b = O(\exp[\text{err}^{-2}])$, *grid search*

Overview of the Problem and Literature

- **Optimization:** we want after $b \geq 1$ iterations:

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

How large does b have to be? Depends on the objective function:

  - **Smooth + Strongly Convex**: $b = O(|\log(\text{err})|)$
    using gradient-descent, (quasi)-Newton, etc.
  - **Smooth + Non Convex**: $b = O(\text{err}^{-d/s})$
    s = smoothness, $d = \dim(\theta)$: curse of dimensionality

- **Global optimizers:**
  1. heuristic (no guarantees), e.g. *genetic algorithms*
  2. or slow, e.g. *simulated annealing* $b = O(\exp[\text{err}^{-2}])$, *grid search*

- **Smoothing the objective:**
  (e.g. McFadden, 1989; Nesterov and Spokoiny, 2017; Bruins et al., 2018)
  1. helps with local optimization
  2. introduces estimation bias, requires undersmoothing

Overview of the Problem and Literature, cont'd

- **Two-step approach:** (Robinson, 1988; Andrews, 1997)
  1. find consistent estimate $\tilde{\theta}_n$,
  2. one (or more) Newton-Raphson iteration(s) from $\tilde{\theta}_n$

  Problem: $\tilde{\theta}_n$ is difficult to compute

Overview of the Problem and Literature, cont'd

- **Two-step approach:** (Robinson, 1988; Andrews, 1997)
  1. find consistent estimate $\tilde{\theta}_n$,
  2. one (or more) Newton-Raphson iteration(s) from $\tilde{\theta}_n$

  Problem: $\tilde{\theta}_n$ is difficult to compute
- **quasi-Bayesian**: (Chernozhukov and Hong, 2003) use MCMC to
  1. compute posterior mean $\overline{\theta}_n = \hat{\theta}_n + o_p(n^{-1/2})$
  2. compute SEs, CIs

  rate of cv. for MCMC mostly requires log-concave posteriors

  (Mengersen and Tweedie, 1996; Brooks, 1998; Belloni and Chernozhukov, 2009)

Overview of the Problem

- **This paper – Econometric assumptions imply:**
  1. Local convexity (local identification, $n = \infty$)
  2. Separation of the global minimum (global identification, $n = \infty$)
  3. Concentration of the sample moments (uniform cv., $n < \infty$)

Overview of the Problem

- **This paper – Econometric assumptions imply:**
  1. Local convexity (local identification, $n = \infty$)
  2. Separation of the global minimum (global identification, $n = \infty$)
  3. Concentration of the sample moments (uniform cv., $n < \infty$)
- **The plan:**
  1. Algorithm, Intuition, Illustration
  2. Local/global cv. with $n = \infty$
  3. Local cv. with $n < \infty$, extensions
  4. Empirical Application

The Algorithm

Smoothed Gauss-Newton Algorithm (sGN)

1. **Inputs** (a) a learning rate $\gamma \in (0, 1)$, (b) a smoothing parameter $\varepsilon > 0$, (c) a weighting matrix W_n, and (d) a sequence $(\theta^b)_{b \geq 0}$ covering the parameter space Θ

Smoothed Gauss-Newton Algorithm (sGN)

1. **Inputs** (a) a learning rate $\gamma \in (0, 1)$, (b) a smoothing parameter $\varepsilon > 0$, (c) a weighting matrix W_n, and (d) a sequence $(\theta^b)_{b \geq 0}$ covering the parameter space Θ

2. **Iterations**:
   set $b = 0$, $\theta_0 = \theta^0$, repeat:
   - **Local step:**

$$\boldsymbol{\theta_{b+1} = \theta_b - \gamma \Big[G_{n,\varepsilon}(\theta_b)' W_n G_{n,\varepsilon}(\theta_b) \Big]^{-1} G_{n,\varepsilon}(\theta_b)' W_n \overline{g}_n(\theta_b)}$$

Smoothed Gauss-Newton Algorithm (sGN)

1. **Inputs** (a) a learning rate $\gamma \in (0,1)$, (b) a smoothing parameter $\varepsilon > 0$, (c) a weighting matrix W_n, and (d) a sequence $(\theta^b)_{b \geq 0}$ covering the parameter space Θ
2. **Iterations**:
   set $b = 0$, $\theta_0 = \theta^0$, repeat:
   - **Local step:**

   $$\boldsymbol{\theta_{b+1} = \theta_b - \gamma \Big[G_{n,\varepsilon}(\theta_b)' W_n G_{n,\varepsilon}(\theta_b) \Big]^{-1} G_{n,\varepsilon}(\theta_b)' W_n \overline{g}_n(\theta_b)}$$

   - **Global step:**

   $$\text{if } \|\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}^{b+1})\|_{W_n} < \|\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_{b+1})\|_{W_n}, \text{ set } \boldsymbol{\theta}_{b+1} = \boldsymbol{\theta}^{b+1}$$

Smoothed Gauss-Newton Algorithm (sGN)

1. **Inputs** (a) a learning rate $\gamma \in (0,1)$, (b) a smoothing parameter $\varepsilon > 0$, (c) a weighting matrix W_n, and (d) a sequence $(\theta^b)_{b\geq 0}$ covering the parameter space Θ

2. **Iterations**:
   set $b = 0$, $\theta_0 = \theta^0$, repeat:

   - **Local step:**

   $$\boldsymbol{\theta_{b+1} = \theta_b - \gamma\Big[G_{n,\varepsilon}(\theta_b)' W_n G_{n,\varepsilon}(\theta_b)\Big]^{-1} G_{n,\varepsilon}(\theta_b)' W_n \overline{g}_n(\theta_b)}$$

   - **Global step:**

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

   - Increment $b := b + 1$, until a stopping critera is met

Smoothed Gauss-Newton Algorithm (sGN)

1. **Inputs** (a) a learning rate $\gamma \in (0,1)$, (b) a smoothing parameter $\varepsilon > 0$, (c) a weighting matrix W_n, and (d) a sequence $(\theta^b)_{b \geq 0}$ covering the parameter space Θ
2. **Iterations**:
   set $b = 0$, $\theta_0 = \theta^0$, repeat:
   - **Local step:**

   $$\boldsymbol{\theta_{b+1} = \theta_b - \gamma \left[G_{n,\varepsilon}(\theta_b)' W_n G_{n,\varepsilon}(\theta_b) \right]^{-1} G_{n,\varepsilon}(\theta_b)' W_n \overline{g}_n(\theta_b)}$$

   - **Global step:**

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

   - Increment $b := b + 1$, until a stopping critera is met
3. **Output**

$$\tilde{\theta}_n = \text{argmin}_{0 \leq j \leq b_{\max}} \|\overline{g}_n(\theta_j)\|_{W_n}$$

Algorithm, continued

- Jacobian computed by convolution smoothing:

$$\overline{g}_{n,\varepsilon}(\theta) = \mathbb{E}_{\sim Z}[\overline{g}_n(\theta + \varepsilon Z)], \quad G_{n,\varepsilon}(\theta) = \partial_\theta \overline{g}_{n,\varepsilon}(\theta), \quad Z \sim \mathcal{N}(0, I)$$

- Unbiased Monte Carlo estimate:

$$\hat{G}_{n,\varepsilon}(\theta) = \frac{1}{\varepsilon L} \sum_{\ell=0}^{L-1} [\overline{g}_n(\theta + \varepsilon Z_\ell) - \overline{g}_n(\theta)] Z'_\ell,$$

In the paper: quasi-Newton Monte Carlo approach

Algorithm, continued

- Jacobian computed by convolution smoothing:

$$\overline{g}_{n,\varepsilon}(\theta) = \mathbb{E}_{\sim Z}[\overline{g}_n(\theta + \varepsilon Z)], \quad G_{n,\varepsilon}(\theta) = \partial_\theta \overline{g}_{n,\varepsilon}(\theta), \quad Z \sim \mathcal{N}(0, I)$$

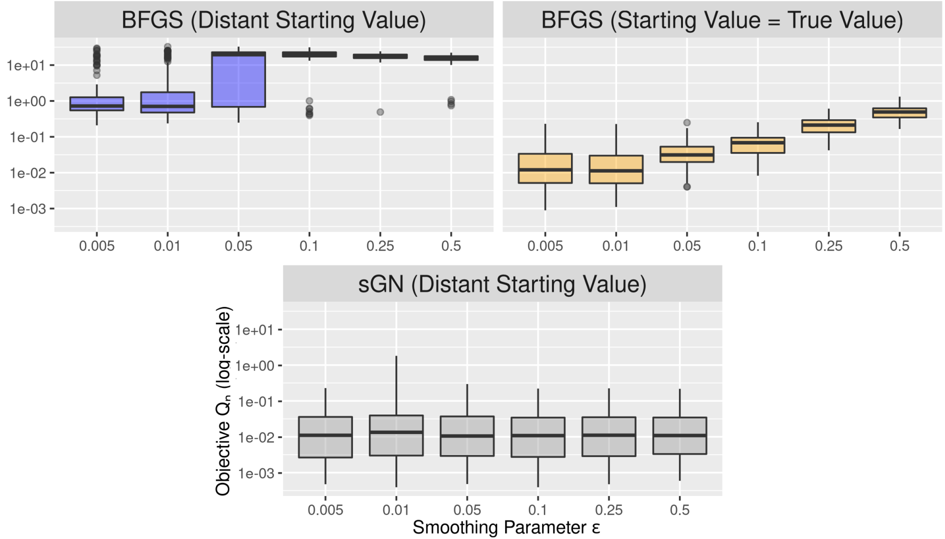
- Unbiased Monte Carlo estimate:

$$\hat{G}_{n,\varepsilon}(\theta) = \frac{1}{\varepsilon L} \sum_{\ell=0}^{L-1} [\overline{g}_n(\theta + \varepsilon Z_\ell) - \overline{g}_n(\theta)] Z'_\ell,$$

  In the paper: quasi-Newton Monte Carlo approach

- Smoothing only Jacobian implies:
  - if $\overline{g}_n(\hat{\theta}_n) = 0$ then $\theta_b = \hat{\theta}_n \Rightarrow \theta_{b+j} = \hat{\theta}_n$, $\forall j \geq 0$ (no bias)
  - allows for 'large bandwidth' $\varepsilon = O(n^{-1/4})$, optimal for optimization

Illustration: Dynamic Discrete Choice model ($d = 15$ pars)



DGP: $y_{it} = \mathbb{1}\{x_{it}'\beta + u_{it} > 0\}$, $u_{it} = \rho u_{it-1} + e_{it}$, $e_{it} \sim \mathcal{N}(0,1)$, $\theta = (\beta, \rho)$, $n = 250$, $T = 10$. DGP and benchmark based on Bruins et al. (2018). BFGS = smoothed moments (generalized indirect inference)

Properties of the Algorithm

Local Convergence, $n = \infty$

- Gauss-Newton (GN) iterations:

$$\theta_{b+1} = \theta_b - \gamma \left[G(\theta_b)' W G(\theta_b)\right]^{-1} G(\theta_b)' W g(\theta_b)$$

- Take $\theta^\dagger$ s.t. $g(\theta^\dagger) = 0$.
- Suppose G is Lipschitz continuous, and

$$\|\theta - \theta^\dagger\| \leq R_G \Rightarrow \sigma_{\min}[G(\theta)] \geq \underline{\sigma} > 0.$$

- Then for any $\gamma \in (0,1)$, $\overline{\gamma} \in (0, \gamma)$ and

$$\|\theta_0 - \theta^\dagger\| \leq \min(R_G, \underline{\sigma}[\gamma - \overline{\gamma}][\gamma L_G \sqrt{\kappa_W}]^{-1}) := R$$

  we have:

$$\|\theta_b - \theta^\dagger\| \leq (1 - \overline{\gamma})^b \|\theta_0 - \theta^\dagger\|, \quad \forall b \geq 1$$

- Local convergence implied by local identification and smoothness

Local Convergence, $n = \infty$

- Quick proof:

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

since $g(\theta^\dagger) = 0$.

Local Convergence, $n = \infty$

- Quick proof:

$$\theta_{b+1} - \theta^\dagger = (1-\gamma)(\theta_b - \theta^\dagger) \\ - \gamma \left[G(\theta_b)' W G(\theta_b) \right]^{-1} G(\theta_b)' W [g(\theta_b) - g(\theta^\dagger) - G(\theta_b)(\theta_b - \theta^\dagger)]$$

since $g(\theta^\dagger) = 0$.

- Now, if $\|\theta_b - \theta^\dagger\| \leq R_G$

$$\| \left[G(\theta_b)' W G(\theta_b) \right]^{-1} G(\theta_b)' W \| \leq \underline{\sigma}^{-1} \sqrt{\kappa_W}$$

and $\|g(\theta_b) - g(\theta^\dagger) - G(\theta_b)(\theta_b - \theta^\dagger)\| \leq L_G \|\theta_b - \theta^\dagger\|^2$

Local Convergence, $n = \infty$

- Quick proof:

$$\theta_{b+1} - \theta^\dagger = (1-\gamma)(\theta_b - \theta^\dagger) \\ - \gamma \left[G(\theta_b)' W G(\theta_b) \right]^{-1} G(\theta_b)' W [g(\theta_b) - g(\theta^\dagger) - G(\theta_b)(\theta_b - \theta^\dagger)]$$

since $g(\theta^\dagger) = 0$.

- Now, if $\|\theta_b - \theta^\dagger\| \leq R_G$

$$\| \left[G(\theta_b)' W G(\theta_b) \right]^{-1} G(\theta_b)' W \| \leq \underline{\sigma}^{-1} \sqrt{\kappa_W}$$

and $\|g(\theta_b) - g(\theta^\dagger) - G(\theta_b)(\theta_b - \theta^\dagger)\| \leq L_G \|\theta_b - \theta^\dagger\|^2$

- By recursion for $\|\theta_0 - \theta^\dagger\| \leq R$:

$$\|\theta_1 - \theta^\dagger\| \leq (1 - \overline{\gamma}) \|\theta_0 - \theta^\dagger\| \leq R$$

$$\vdots$$

$$\|\theta_{b+1} - \theta^\dagger\| \leq (1 - \overline{\gamma}) \|\theta_b - \theta^\dagger\| \leq R$$

Local $\to$ Global Convergence, $n = \infty$

- So far: local convergence, need $\|\theta_0 - \theta^\dagger\| \leq R$
- Global convergence not guaranteed otherwise
- Add Global Step:

$$\theta_{b+1} = \theta_b - \gamma \left[G(\theta_b)' W G(\theta_b)\right]^{-1} G(\theta_b)' W g(\theta_b)$$
$$\text{if } \|g(\theta^{b+1})\|_W < \|g(\theta_{b+1})\|_W \text{ set } \theta_{b+1} = \theta^{b+1}$$

Local $\rightarrow$ Global Convergence, $n = \infty$

- Three ingredients, rely on a different norm $\|\cdot\|_{G'WG}$:
  1. $\exists \bar{r}_g > 0$ s.t. for $\|\theta_b - \theta^\dagger\|_{G'WG} \leq \bar{r}_g$, $G = G(\theta^\dagger)$

$$\|g(\theta_{b+1})\|_W \leq (1-\bar{\gamma})\|g(\theta_b)\|_W$$

Local $\to$ Global Convergence, $n = \infty$

- Three ingredients, rely on a different norm $\|\cdot\|_{G'WG}$:
  1. $\exists \bar{r}_g > 0$ s.t. for $\|\theta_b - \theta^\dagger\|_{G'WG} \leq \bar{r}_g$, $G = G(\theta^\dagger)$

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

  2. and for any $\|\theta - \theta^\dagger\|_{G'WG} \leq \bar{r}_g$:

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

Local $\rightarrow$ Global Convergence, $n = \infty$

- Three ingredients, rely on a different norm $\|\cdot\|_{G'WG}$:
  1. $\exists \bar{r}_g > 0$ s.t. for $\|\theta_b - \theta^\dagger\|_{G'WG} \leq \bar{r}_g$, $G = G(\theta^\dagger)$

$$\|g(\theta_{b+1})\|_W \leq (1 - \bar{\gamma})\|g(\theta_b)\|_W$$

  2. and for any $\|\theta - \theta^\dagger\|_{G'WG} \leq \bar{r}_g$:

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

  3. under global identification, $\exists \underline{r}_g \in (0, \bar{r}_g)$:

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

Local $\rightarrow$ Global Convergence, $n = \infty$

- Three ingredients, rely on a different norm $\|\cdot\|_{G'WG}$:
  1. $\exists \bar{r}_g > 0$ s.t. for $\|\theta_b - \theta^\dagger\|_{G'WG} \leq \bar{r}_g$, $G = G(\theta^\dagger)$

$$\|g(\theta_{b+1})\|_W \leq (1-\bar{\gamma})\|g(\theta_b)\|_W$$

  2. and for any $\|\theta - \theta^\dagger\|_{G'WG} \leq \bar{r}_g$:

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

  3. under global identification, $\exists \underline{r}_g \in (0, \bar{r}_g)$:

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

- Combine to get global convergence:
  - Take $b = k + j$, with k s.t. $\sup_{\theta \in \Theta} \left(\inf_{0 \leq \ell \leq k} \|\theta - \theta^\ell\|_{G'WG}\right) \leq \underline{r}_g$,
  - then for any $j \geq 0$:

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

Local $\to$ Global Convergence, $n = \infty$

- Three ingredients, rely on a different norm $\|\cdot\|_{G'WG}$:
  1. $\exists \overline{r}_g > 0$ s.t. for $\|\theta_b - \theta^\dagger\|_{G'WG} \leq \overline{r}_g$, $G = G(\theta^\dagger)$
  $$\|g(\theta_{b+1})\|_W \leq (1-\overline{\gamma})\|g(\theta_b)\|_W$$
  2. and for any $\|\theta - \theta^\dagger\|_{G'WG} \leq \overline{r}_g$:
  $$(1-\overline{\gamma}/2)\|\theta - \theta^\dagger\|_{G'WG} \leq \|g(\theta)\|_W \leq (1+\overline{\gamma}/2)\|\theta - \theta^\dagger\|_{G'WG}$$
  3. under global identification, $\exists \underline{r}_g \in (0, \overline{r}_g)$:
  $$\inf_{\|\theta-\theta^\dagger\|_{G'WG} \geq \overline{r}_g} \|g(\theta)\|_W \geq (1+\overline{\gamma}/2)(1-\overline{\gamma})\underline{r}_g$$

- Combine to get global convergence:
  - Take $b = k + j$, with k s.t. $\sup_{\theta \in \Theta}\left(\inf_{0 \leq \ell \leq k} \|\theta - \theta^\ell\|_{G'WG}\right) \leq \underline{r}_g$,
  - then for any $j \geq 0$:
  $$\|\theta_b - \theta^\dagger\|_{G'WG} \leq (1+\overline{\gamma}/2)(1-\overline{\gamma})^j \underline{r}_g$$

- Fast convergence after k iterations

Choice of covering sequence $(\theta^\ell)_{\ell\geq 0}$

- For simplicity suppose $\Theta = [0,1]^d$.
- Take $r > 0$, we want

$$D_k = \sup_{\theta\in\Theta}\Big[\inf_{0\leq\ell\leq k-1}\|\theta - \theta^\ell\|\Big] \leq r$$

- Covering number arguments give a lower bound:

$$k \geq r^{-d}\frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(\mathcal{B})}, \quad \mathcal{B} = \text{ unit ball},$$

i.e. $D_k = O(k^{-1/d})$

Choice of covering sequence $(\theta^\ell)_{\ell \geq 0}$

- For simplicity suppose $\Theta = [0,1]^d$.
- Take $r > 0$, we want

$$D_k = \sup_{\theta \in \Theta} [\inf_{0 \leq \ell \leq k-1} \|\theta - \theta^\ell\|] \leq r$$

- Covering number arguments give a lower bound:

$$k \geq r^{-d} \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(\mathcal{B})}, \quad \mathcal{B} = \text{ unit ball},$$

i.e. $D_k = O(k^{-1/d})$

- Niederreiter (1983, Th3) implies for Sobol/Halton sequence:

$$D_k \leq O(\sqrt{d} \log[k] k^{-1/d}),$$

close in rate, up to logs

Choice of covering sequence $(\theta^\ell)_{\ell \geq 0}$

- For simplicity suppose $\Theta = [0,1]^d$.
- Take $r > 0$, we want

$$D_k = \sup_{\theta \in \Theta} [\inf_{0 \leq \ell \leq k-1} \|\theta - \theta^\ell\|] \leq r$$

- Covering number arguments give a lower bound:

$$k \geq r^{-d} \frac{\text{vol}(\Theta)}{\text{vol}(\mathcal{B})}, \quad \mathcal{B} = \text{ unit ball},$$

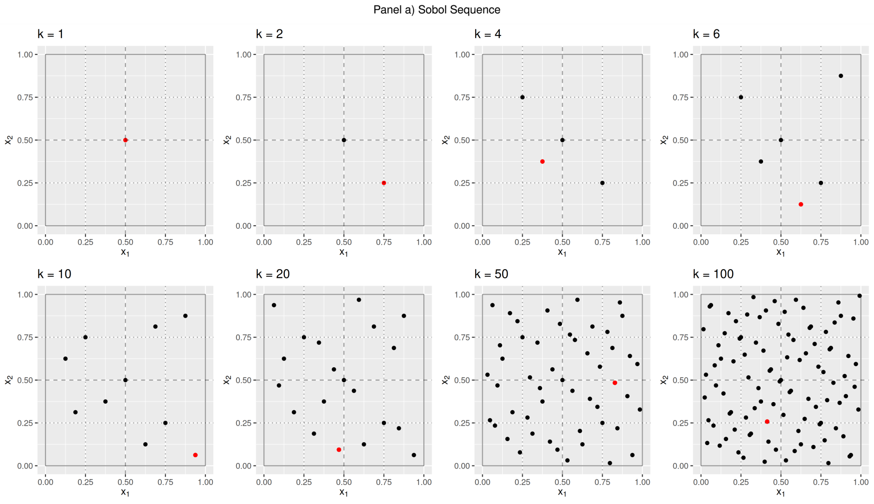
  i.e. $D_k = O(k^{-1/d})$
- Niederreiter (1983, Th3) implies for Sobol/Halton sequence:

$$D_k \leq O(\sqrt{d} \log[k] k^{-1/d}),$$

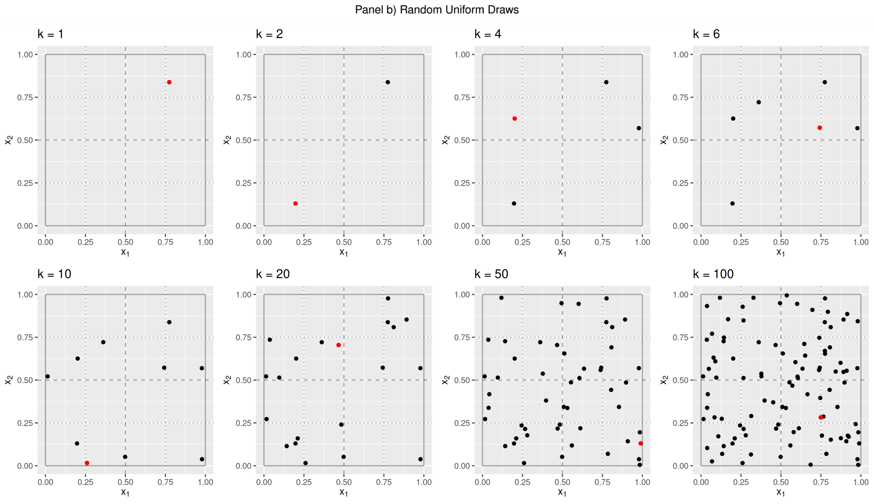
  close in rate, up to logs
- Compare with $\theta^\ell \overset{iid}{\sim} \mathcal{U}_\Theta$:

$$D_k = O_p(k^{-1/2d})$$

Choice of covering sequence $(\theta^\ell)_{\ell \geq 0}$



Choice of covering sequence $(\theta^\ell)_{\ell\geq 0}$



Local $\rightarrow$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k

Local $\rightarrow$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?

Local $\rightarrow$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?
- Forneron and Zhong (2022): large n – suppose $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$, $\forall \theta \in \Theta$. Questions:

Local $\rightarrow$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?
- Forneron and Zhong (2022): large n – suppose $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$, $\forall \theta \in \Theta$. Questions:
  Can GN alone be globally convergent?

Local $\rightarrow$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?
- Forneron and Zhong (2022): large n – suppose $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$, $\forall \theta \in \Theta$. Questions:

  Can GN alone be globally convergent? **Yes.**

Local $\to$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?
- Forneron and Zhong (2022): large n – suppose $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$, $\forall \theta \in \Theta$. Questions:
  Can GN alone be globally convergent? **Yes.** For over-identified models?

Local $\to$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?
- Forneron and Zhong (2022): large n – suppose $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$, $\forall \theta \in \Theta$. Questions:
  Can GN alone be globally convergent? **Yes.** For over-identified models? **Yes.***

Local $\rightarrow$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?
- Forneron and Zhong (2022): large n – suppose $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$, $\forall \theta \in \Theta$. Questions:
  Can GN alone be globally convergent? **Yes.** For over-identified models? **Yes.*** Allowing for misspecification?

Local $\to$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?
- Forneron and Zhong (2022): large n – suppose $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$, $\forall \theta \in \Theta$. Questions:
  Can GN alone be globally convergent? **Yes.** For over-identified models? **Yes.*** Allowing for misspecification? **Yes.****

Local $\to$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?
- Forneron and Zhong (2022): large n – suppose $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$, $\forall \theta \in \Theta$. Questions:

  Can GN alone be globally convergent? **Yes.** For over-identified models? **Yes.*** Allowing for misspecification? **Yes.****

  - Convexity not required
  - Rank conditions exclude local optima
  - Restrictions on tuning parameter γ

Local $\rightarrow$ Global Convergence, $n = \infty$

- Curse of dimensionality only appears through k
- Can we do better, without convexity?
- Forneron and Zhong (2022): large n – suppose $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$, $\forall \theta \in \Theta$. Questions:

  Can GN alone be globally convergent? **Yes.** For over-identified models? **Yes.*** Allowing for misspecification? **Yes.****

  - Convexity not required
  - Rank conditions exclude local optima
  - Restrictions on tuning parameter γ
- Applications: BLP, Impulse Response Matching

Finite Samples

Non-smooth moments, $n < \infty$

- Allow for discontinuous moments, assume:

$$[\mathbb{E}(\sup_{\|\theta_1-\theta_2\|\leq\delta} \|g(\theta_1;x_i) - g(\theta_2;x_i)\|^2)]^{1/2} \leq L_g\delta^{\psi}, \quad \psi \in (0,1]$$

and x_i are iid $\Rightarrow$ probability bounds for sample/smoothed moments

Non-smooth moments, $n < \infty$

- Allow for discontinuous moments, assume:

$$[\mathbb{E}(\sup_{\|\theta_1-\theta_2\|\leq\delta} \|g(\theta_1;x_i) - g(\theta_2;x_i)\|^2)]^{1/2} \leq L_g \delta^{\psi}, \quad \psi \in (0,1]$$

and x_i are iid $\Rightarrow$ probability bounds for sample/smoothed moments

- e.g. for any $c_n \geq 1$

$$\|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\| \leq c_n n^{-1/2} C_\Theta L_g \|\theta_1 - \theta_2\|^{\psi},$$
$$\|[G_{n,\varepsilon}(\theta_1) - G_\varepsilon(\theta_1)] - [G_{n,\varepsilon}(\theta_2) - G_\varepsilon(\theta_2)]\| \leq c_n \varepsilon^{-1} n^{-1/2} C_\Theta M_Z L_g \|\theta_1 - \theta_2\|^{\psi},$$

unif. in θ_1, θ_2, and $\varepsilon > 0$ with prob. $1 - C/c_n$

Non-smooth moments, $n < \infty$

- Allow for discontinuous moments, assume:

$$[\mathbb{E}(\sup_{\|\theta_1-\theta_2\|\leq\delta} \|g(\theta_1;x_i) - g(\theta_2;x_i)\|^2)]^{1/2} \leq L_g\delta^{\psi}, \quad \psi \in (0,1]$$

  and x_i are iid $\Rightarrow$ probability bounds for sample/smoothed moments

- e.g. for any $c_n \geq 1$

$$\|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\| \leq c_n n^{-1/2} C_\Theta L_g \|\theta_1 - \theta_2\|^{\psi},$$
$$\|[G_{n,\varepsilon}(\theta_1) - G_\varepsilon(\theta_1)] - [G_{n,\varepsilon}(\theta_2) - G_\varepsilon(\theta_2)]\| \leq c_n \varepsilon^{-1} n^{-1/2} C_\Theta M_Z L_g \|\theta_1 - \theta_2\|^{\psi},$$

  unif. in θ_1, θ_2, and $\varepsilon > 0$ with prob. $1 - C/c_n$

- use these bounds in the local cv. proof

Non-smooth moments, $n < \infty$

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

Non-smooth moments, $n < \infty$

- Let
$$\hat{\theta}_n = \theta^\dagger - (G' W_n G)^{-1} G' W_n \overline{g}_n(\theta^\dagger),$$
where $G = G(\theta^\dagger)$
- Take $G_b = G_{n,\varepsilon}(\theta_b)$ and re-arrange terms to get:
$$\begin{aligned}\theta_{b+1} - \hat{\theta}_n &= (1-\gamma)(\theta_b - \hat{\theta}_n) \\ &\quad - \gamma(G_b' W_n G_n)^{-1} W_n G_b' [\overline{g}_n(\theta_b) - G_b(\theta_b - \hat{\theta}_n)]\end{aligned}$$

Non-smooth moments, $n < \infty$

- Let
$$\hat{\theta}_n = \theta^\dagger - (G' W_n G)^{-1} G' W_n \overline{g}_n(\theta^\dagger),$$
where $G = G(\theta^\dagger)$
- Take $G_b = G_{n,\varepsilon}(\theta_b)$ and re-arrange terms to get:
$$\begin{aligned} \theta_{b+1} - \hat{\theta}_n &= (1-\gamma)(\theta_b - \hat{\theta}_n) \\ &\quad - \gamma(G_b' W_n G_n)^{-1} W_n G_b' [\overline{g}_n(\theta_b) - G_b(\theta_b - \hat{\theta}_n)] \end{aligned}$$
- Then focus on $\overline{g}_n(\theta_b) - G_{n,\varepsilon}(\theta_b)(\theta_b - \hat{\theta}_n)$ using the unif. bounds

Non-smooth moments, $n < \infty$

- Let
$$\hat{\theta}_n = \theta^\dagger - (G' W_n G)^{-1} G' W_n \overline{g}_n(\theta^\dagger),$$
where $G = G(\theta^\dagger)$
- Take $G_b = G_{n,\varepsilon}(\theta_b)$ and re-arrange terms to get:
$$\begin{aligned} \theta_{b+1} - \hat{\theta}_n &= (1-\gamma)(\theta_b - \hat{\theta}_n) \\ &\quad - \gamma(G_b' W_n G_n)^{-1} W_n G_b' [\overline{g}_n(\theta_b) - G_b(\theta_b - \hat{\theta}_n)] \end{aligned}$$
- Then focus on $\overline{g}_n(\theta_b) - G_{n,\varepsilon}(\theta_b)(\theta_b - \hat{\theta}_n)$ using the unif. bounds
- Key term:
$$\|G_{n,\varepsilon}(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n)\| \leq C_1 (c_n n^{-1/2})^{1+\psi} \left(1 + \frac{c_n n^{-1/2}}{\varepsilon} + \frac{\varepsilon}{(c_n n^{-1/2})^\psi}\right) := \Gamma_{n,\varepsilon}$$
mesures stability of Gauss-Newton at $\theta = \hat{\theta}_n$

Non-smooth moments, $n < \infty$

- Pick $c_n \geq 1$. Uniformly in $\|\theta_b - \hat{\theta}_n\| \leq R_n = R - O(\varepsilon + c_n n^{-1/2}\varepsilon^{-1})$:

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

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

Non-smooth moments, $n < \infty$

- Pick $c_n \geq 1$. Uniformly in $\|\theta_b - \hat{\theta}_n\| \leq R_n = R - O(\varepsilon + c_n n^{-1/2}\varepsilon^{-1})$:

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

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

- where:

$$\Delta_{n,\varepsilon}(x) = \frac{C_2}{\underline{\sigma}_{n,\varepsilon}} \left(\Gamma_{n,\varepsilon} + \frac{[c_n n^{-1/2}]^2}{\varepsilon} x^{\psi} + \frac{c_n n^{-1/2}}{\varepsilon} x \right)$$

  with $\sqrt{n}\Gamma_{n,\varepsilon} = o(1)$ if $\varepsilon = o(1)$, $\sqrt{n}\varepsilon \to \infty$

Non-smooth moments, $n < \infty$

- Pick $c_n \geq 1$. Uniformly in $\|\theta_b - \hat{\theta}_n\| \leq R_n = R - O(\varepsilon + c_n n^{-1/2}\varepsilon^{-1})$:

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

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

- where:

$$\Delta_{n,\varepsilon}(x) = \frac{C_2}{\underline{\sigma}_{n,\varepsilon}} \left(\Gamma_{n,\varepsilon} + \frac{[c_n n^{-1/2}]^2}{\varepsilon} x^{\psi} + \frac{c_n n^{-1/2}}{\varepsilon} x \right)$$

  with $\sqrt{n}\Gamma_{n,\varepsilon} = o(1)$ if $\varepsilon = o(1)$, $\sqrt{n}\varepsilon \to \infty$

- For any $\tau \in (0, 1)$, n large enough, ε small enough:

$$\|\theta_b - \hat{\theta}_n\| \leq (1 - \overline{\gamma} + \tau\overline{\gamma})^b \|\theta_0 - \hat{\theta}_n\| + \frac{\gamma}{\overline{\gamma}(1 - \tau)} R_n,$$

Non-smooth moments, $n < \infty$

- Pick $c_n \geq 1$. Uniformly in $\|\theta_b - \hat{\theta}_n\| \leq R_n = R - O(\varepsilon + c_n n^{-1/2}\varepsilon^{-1})$:

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

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

- where:

$$\Delta_{n,\varepsilon}(x) = \frac{C_2}{\underline{\sigma}_{n,\varepsilon}}\left(\Gamma_{n,\varepsilon} + \frac{[c_n n^{-1/2}]^2}{\varepsilon}x^{\psi} + \frac{c_n n^{-1/2}}{\varepsilon}x\right)$$

  with $\sqrt{n}\Gamma_{n,\varepsilon} = o(1)$ if $\varepsilon = o(1)$, $\sqrt{n}\varepsilon \to \infty$

- For any $\tau \in (0, 1)$, n large enough, ε small enough:

$$\|\theta_b - \hat{\theta}_n\| \leq (1 - \overline{\gamma} + \tau\overline{\gamma})^b\|\theta_0 - \hat{\theta}_n\| + \frac{\gamma}{\overline{\gamma}(1 - \tau)}R_n,$$

- with probability $1 - (1 + C)/c_n$, where $R_n = O(\Gamma_{n,\varepsilon})$, and then: $\sqrt{n}(\theta_b - \hat{\theta}_n) = O_p(\sqrt{n}\Gamma_{n,\varepsilon}) = o_p(1)$ if $b = O(\log[n])$

Non-smooth moments, $n < \infty$

- Global convergence similar to $n = \infty$, main differences:
  - norm equivalence now involves $\|\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\|_{W_n}$
  - need n large enough for tight enough equivalence
- Global rate of cv. slightly slower than $n = \infty$

Non-smooth moments, $n < \infty$

- Optimal choice of bandwidth: $\varepsilon \asymp \sqrt{c_n} n^{-1/4}$

Non-smooth moments, $n < \infty$

- Optimal choice of bandwidth: $\varepsilon \asymp \sqrt{c_n} n^{-1/4}$
- Tradeoff: convergence rate $(1 - \overline{\gamma})$ vs. sampling noise $\gamma \Delta_{n,\varepsilon}$

Non-smooth moments, $n < \infty$

- Optimal choice of bandwidth: $\varepsilon \asymp \sqrt{c_n} n^{-1/4}$
- Tradeoff: convergence rate $(1 - \overline{\gamma})$ vs. sampling noise $\gamma \Delta_{n,\varepsilon}$
- Extension 1.: heavy-ball (Polyak, 1964)

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

allows for $\overline{\gamma}(\alpha) > \gamma$, optimal $\alpha(\gamma)$ tabulated

Non-smooth moments, $n < \infty$

- Optimal choice of bandwidth: $\varepsilon \asymp \sqrt{c_n} n^{-1/4}$
- Tradeoff: convergence rate $(1 - \overline{\gamma})$ vs. sampling noise $\gamma \Delta_{n,\varepsilon}$
- Extension 1.: heavy-ball (Polyak, 1964)

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

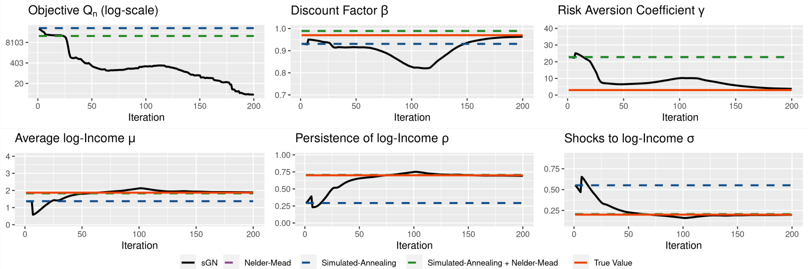
  allows for $\overline{\gamma}(\alpha) > \gamma$, optimal $\alpha(\gamma)$ tabulated
- Extension 2.: quasi-Newton Monte Carlo estimator of $G_{n,\varepsilon}$

Simulated Example: Estimation of an Aiyagari model

SMM estimation of an Aiyagari model

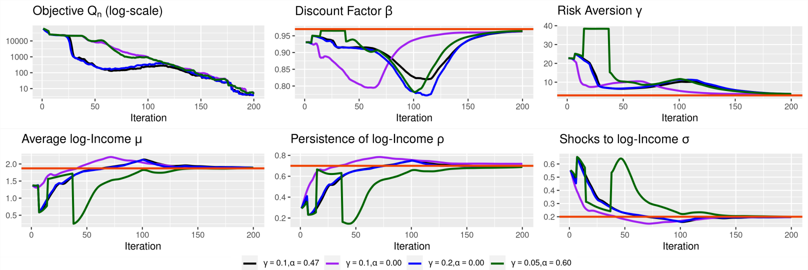
- Optimal consumption choice with borrowing constraint
- Non-smooth: discretize GDP, value function iterations
- Moments = sample quantiles
- Computationally demanding, compare with global & local optimizers
- Set $n = 10000$, $T = 2$
- Estimate $\theta = (\beta, \gamma, \mu, \rho, \sigma)$, preferences and log-income process

Results vs. optimizers (one sample)



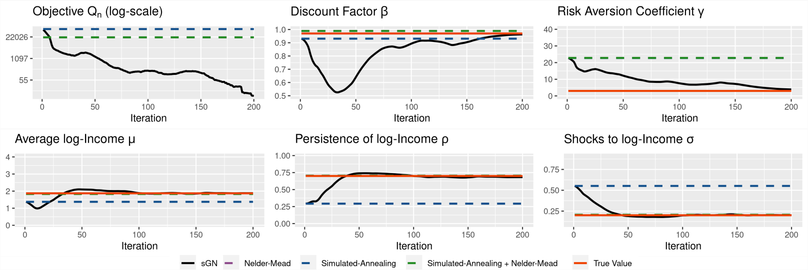
Legend: $n = 10000$, $T = 2$. $\gamma = 0.1$, $\alpha = 0.47$. sGN (black): Algorithm 1. Simulated-Annealing (dashed blue): 5000 iterations from θ_0. Simulated-Annealing + Nelder-Mead (dashed green): run Nelder-Mead after 5000 Simulated-Annealing iterations.

Range of optimization parameters (one sample)



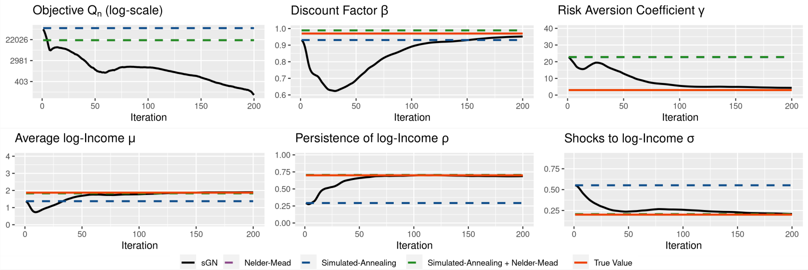
Legend: $n = 10000$, $T = 2$. $\varepsilon = 0.1$. sGN (black): Algorithm 1. Simulated-Annealing (dashed blue): 5000 iterations from θ_0. Simulated-Annealing + Nelder-Mead (dashed green): run Nelder-Mead after 5000 Simulated-Annealing iterations.

$2x$ smoothing parameter (one sample)



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

$5x$ smoothing parameter (one sample)



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

Empirical Application: Joint Retirement Decision

Empirical Application: Interdependent Durations

- Replication of Honoré and de Paula (2018, HP)
- Model of joint retirement decision (husband + wife)
- Likelihood intractable: indirect inference, discrete outcomes
- Estimation is difficult, HP use a ‘loop of procedures’:
  (a) *particle swarm*
  (b) *Powell's conjugate direction* method
  (c) *downhill simplex* (fminsearch)
  (d) *pattern search*
  (e) *particle swarm* focusing on specific parameters
- with fairly good starting values

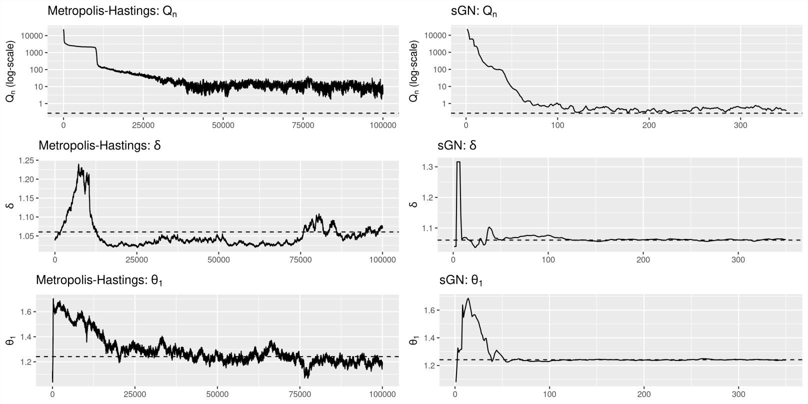
Empirical Application: Interdependent Durations

	Coefficients for Wives				Coefficients for Husbands			
	HP		sGN		HP		sGN	
δ	1.052 (0.039)	1.064 (0.042)	1.060 (0.039)	1.064 (0.037)	1.052 (0.039)	1.064 (0.042)	1.060 (0.039)	1.064 (0.037)
θ_1	1.244 (0.054)	1.244 (0.054)	1.241 (0.055)	1.233 (0.050)	1.169 (0.043)	1.218 (0.058)	1.181 (0.043)	1.192 (0.040)
$\geq$ 62 yrs-old	10.640 (5.916)	13.446 (5.694)	10.203 (7.818)	12.254 (5.692)	31.532 (11.356)	39.824 (11.372)	33.330 (8.131)	35.371 (7.672)
$\geq$ 65 yrs-old	10.036 (11.555)	12.326 (7.495)	10.480 (10.067)	11.974 (10.897)	25.696 (9.497)	29.254 (11.229)	25.203 (13.215)	26.240 (14.289)
...	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Q_n(\theta_0)$ $Q_n(\hat{\theta}_n)$	93.70 0.470	89.77 0.758	2.10^4 0.271	5.10^4 0.342	- -	- -	- -	- -
$\dim(\theta)$	12	30	12	30	-	-	-	-
Time	3h25m	5h34m	11min	11min	-	-	-	-

HP = Honoré and de Paula (2018), Paper: also compare with MCMC

sGN: random starting values, 250 iterations

Comparison with MCMC, distant starting value



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

Conclusion

- Global optimization is slow, difficult
- Econometric assumptions: faster rates possible
- Algorithm:
  - does not require undersmoothing (more robust)
  - automatic transition from global to local cv.
- Most applications: smoothing not tractable
  - quasi-Newton Monte Carlo approach
  - derive exponential bounds
  - computationally attractive (cf. empirical application)
- Beyond GMM:
  - global step extends to other M-estimations (e.g. MLE)
  - local step requires some structure

References

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

Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of mcmc-based estimators in large samples. *The Annals of Statistics*, 37(4):2011–2055.

Brooks, S. P. (1998). Mcmc convergence diagnosis via multivariate bounds on log-concave densities. *The Annals of Statistics*, 26(1):398–433.

Bruins, M., Duffy, J. A., Keane, M. P., and Smith Jr, A. A. (2018). Generalized indirect inference for discrete choice models. *Journal of econometrics*, 205(1):177–203.

Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Forneron, J.-J. and Zhong, L. (2022). Convexity not required: Estimation of smooth moment condition models. *Manuscript in preparation*.

Honoré, B. E. and de Paula, Á. (2018). A new model for interdependent durations. *Quantitative Economics*, 9(3):1299–1333.

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

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the hastings and metropolis algorithms. *The annals of Statistics*, 24(1):101–121.

Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566.

Newey, W. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 36:4, pages 2111–2234. North Holland.

Niederreiter, H. (1983). A quasi-monte carlo method for the approximate computation of the extreme values of a function. In *Studies in pure mathematics*, pages 523–529. Springer.

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17.

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

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

Convexity Not Required

Global convergence under a rank condition

- Forneron and Zhong (2022), suppose:

$$\sigma_{\min}[G_n(\theta_1)' W_n G_n(\theta_2)] \geq \underline{\sigma}^2 \underline{\lambda}_W > 0 \text{ for all } \theta_1, \theta_2 \in \Theta$$

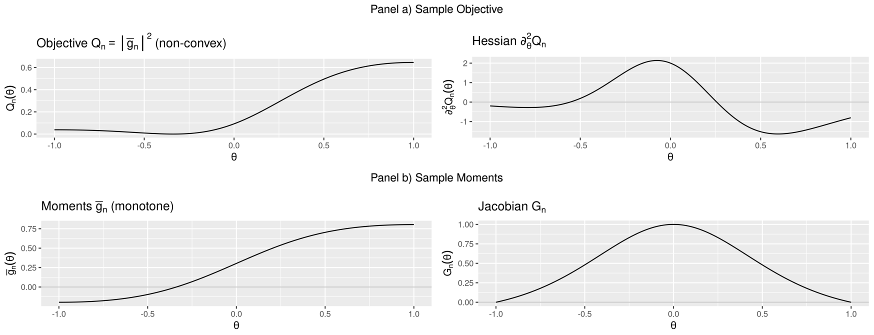
- for $\gamma \in (0,1)$ small enough, $\exists \overline{\gamma} \in (0,1)$, $0 < \underline{\gamma}, \overline{\gamma}, C$ and $C_n = O_p(1)$:

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

- Rank condition is sufficient for cv., $\|\overline{g}_n(\cdot)\|^2_{W_n}$ can be non-convex

Global convergence under a rank condition

- Example: $y_t = e_t - \theta e_{t-1}$, $e_t \overset{iid}{\sim} (0,1)$, $|\theta| < 1$.
- Minimum Distance: $y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + u_t$
- Minimize $\|\hat{\beta}_n - \beta(\theta)\|$.
- For $p = 1$, $\text{plim}_{n\to\infty} \hat{\beta}_n = \beta(\theta) = -\theta/(1+\theta^2)$



Global convergence under a rank condition

- For $p = 1$, $Q_n(\theta) = [\hat{\beta}_n + \theta/(1+\theta^2)]^2$ is **non-convex**
- However:

$$F_n(\theta) = \int_{\vartheta=0}^{\theta} [\hat{\beta}_n + \vartheta/(1+\vartheta^2)] d\vartheta = \theta\hat{\beta}_1 + \frac{1}{2}\log(1+\theta^2)$$

  is convex on $(-1, 1)$ and $\partial_\theta F_n = \overline{g}_n$
- Q_n and F_n have the same minimizer but:
  Minimizing Q_n is difficult, minimizing F_n is not
- Gauss-Newton is minimizing F_n (implicitly)

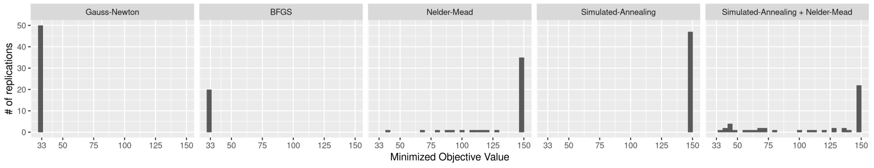
Global convergence under a rank condition

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

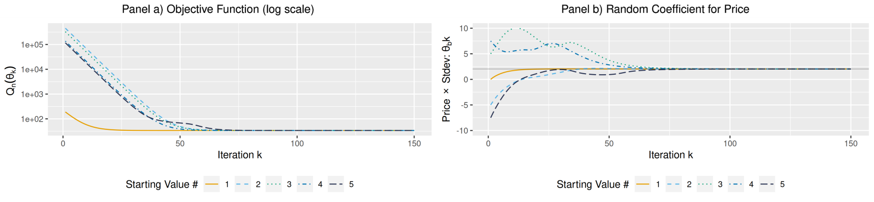
Legend: simulated data with sample size $n = 200$, $\theta^\dagger = -1/2$. For $p = 1$, $\overline{g}_n(\theta) = \hat{\beta}_1 - \theta/(1+\theta^2)$. For $p = 12$, $\overline{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$ where $\beta(\theta)$ is the p-limit of the AR(p) coefficients, evaluated at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

Application 1: BLP with Cereal Data

Comparison with other methods:

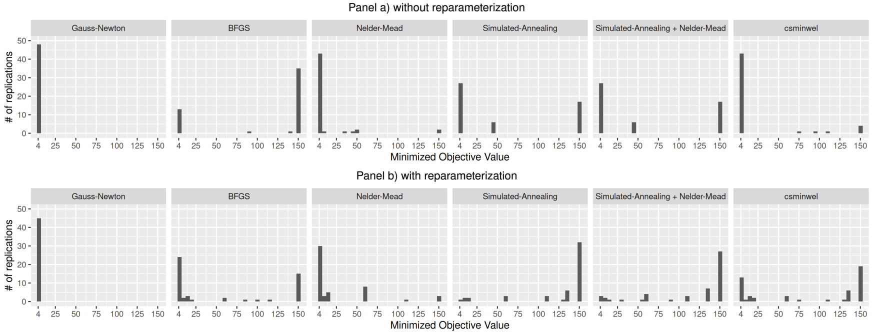


Paths for 5 starting values in $[-10, 10] \times \cdots \times [-10, 10]$:



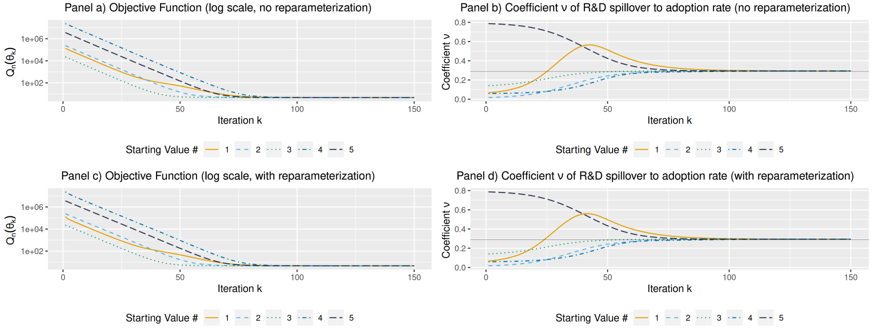
Application 2: Impulse Response Matching

Comparison with other methods:



Application 2: Impulse Response Matching

Paths for 5 starting values:



Local Convergence

Local Convergence ($n = \infty$)

- Take $R > 0$, such that $\sigma_{\min}[G(\theta)] \geq \underline{\sigma} > 0$ if $\|\theta - \theta^\dagger\| \leq R_G$ (exists under local identification + continuity of G)
- Now, let $G_b = G(\theta_b)$:

$$\begin{aligned}\theta_{b+1} - \theta^\dagger &= (1-\gamma)(\theta_b - \theta^\dagger) \\ &\qquad - \gamma(G_b' W G_b)^{-1} G_b' W[g(\theta_b) - g(\theta^\dagger) - G_b(\theta_b - \theta^\dagger)]\end{aligned}$$

  where $g(\theta^\dagger) = 0$
- If G is Lipschitz, when $\|\theta_b - \theta^\dagger\| \leq R_G$:

$$\begin{aligned}\|\theta_{b+1} - \theta^\dagger\| &\leq (1 - \gamma + \gamma \frac{L_G \sqrt{\kappa_W}}{\underline{\sigma}} \|\theta_b - \theta^\dagger\|)\|\theta_b - \theta^\dagger\| \\ &\leq (1 - \overline{\gamma})\|\theta_b - \theta^\dagger\|\end{aligned}$$

  if $\|\theta_b - \theta^\dagger\| \leq [\gamma - \overline{\gamma}]\frac{\underline{\sigma}}{\gamma L_G \sqrt{\kappa_W}} := R$
- Take $\|\theta_0 - \theta^\dagger\| \leq \min(R, R_G)$ and iterate

Local Convergence ($n < \infty$)

- Similar steps, additional terms, let $H_b = (G_b' W_n G_b)^{-1} G_b' W_n$:

$$
\begin{aligned}
\theta_{b+1} - \hat{\theta}_n &= (1-\gamma)(\theta_b - \hat{\theta}_n) && (1)\\
&- \gamma H_b[\overline{g}_n(\theta_b) - \overline{g}_n(\hat{\theta}_n) - G(\theta_b)(\theta_b - \hat{\theta}_n)] && (2)\\
&- \gamma H_b[G(\theta_b) - G_\varepsilon(\theta_b)](\theta_b - \hat{\theta}_n) && (3)\\
&- \gamma H_b[G_\varepsilon(\theta_b) - G_{n,\varepsilon}(\theta_b)](\theta_b - \hat{\theta}_n) && (4)\\
&- \gamma (G_b' W_n G_b)^{-1}[G_{n,\varepsilon}(\theta_b) - G_{n,\varepsilon}(\hat{\theta}_n)]' W_n \overline{g}_n(\hat{\theta}_n) && (5)\\
&- \gamma (G_b' W_n G_b)^{-1} G_{n,\varepsilon}(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) && (6)
\end{aligned}
$$

- (1): deterministic, (2): tail bounds (van der Vaart and Wellner, 1996, Ch2.14) + smoothness of $g(\cdot)$, (3): bounds with smoothing, (4): tail bounds with smoothing, (5): Lipschitz + stochastic bounds, (6): stochastic bounds
- Uniform bounds: holds for all θ with the same probability level
- $\Gamma_{n,\varepsilon} = (c_n n^{-1/2})^{1+\psi}(1 + \varepsilon^{-1} c_n n^{-1/2} + \varepsilon (c_n n^{-1/2})^{-\psi})$ comes from (6) and gives the smoothing bias, others give $\overline{\gamma}$ and $\Delta_{n,\varepsilon}$

Global Convergence

Quasi-Newton Monte Carlo Jacobian Update

quasi-Newton Monte Carlo algorithm

- If $G_{n,\varepsilon}$ not available in closed form, it can be approximated
- 1) **Input** $L \geq d$
- 2.0) **Initialization (**$b = 0$**)**
  - draw $Z_{-\ell} \sim \mathcal{N}(0, I)$, $\ell = 0, \dots, L-1$
  - compute $Y_{-\ell} = \varepsilon^{-1}[\overline{g}_n(\theta_0 + \varepsilon Z_{-\ell}) - \overline{g}_n(\theta_0)]$
- 2.1) **Update (**$b > 0$**)**
  - draw $Z_b \sim \mathcal{N}(0, I)$
  - compute $Y_b = \varepsilon^{-1}[\overline{g}_n(\theta_b + \varepsilon Z_b) - \overline{g}_n(\theta_b)]$
- 3) **Approximate**
  - de-mean $\tilde{Z}_{b-\ell} = Z_{b-\ell} - \sum_{\ell=0}^{L-1} Z_{b-\ell}/L$
  - compute $\hat{G}_L(\theta_b) = \sum_{\ell=0}^{L-1} Y_{b-\ell}\tilde{Z}_{b-\ell}(\sum_{\ell=0}^{L-1} \tilde{Z}_{b-\ell}\tilde{Z}_{b-\ell})^{-1}$
- Use $\hat{G}_L$ in the main algorithm

Acceleration

Acceleration

- Local convergence with $n < \infty$ looks like:

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

- Ideally $\overline{\gamma}$ is large and γ is small
- But we have $\overline{\gamma} < \gamma$: faster convergence implies more sensitive to sampling uncertainty
- Solution: accelerate:

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

- derive VAR(1)-type representation, well-chosen α implies $\overline{\gamma} > \gamma$: faster convergence without noise sensitivity

Acceleration: Optimal choice of α

Table 1: Values of γ and optimal choice of α

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

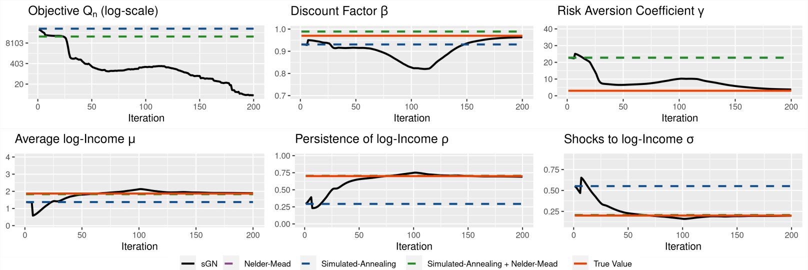
Simulated Example

SMM estimation of an Aiyagari model

- panel data, log-AR(1) income process, optimal consumption choice with borrowing constraint
- non-smooth: discretize GDP, value function iterations, moments = sample quantiles
- computationally demanding, compare with global & local optimizers
- set $n = 10,000$, $T = 2$ (large/short panel)

Results vs. optimizers (one sample)

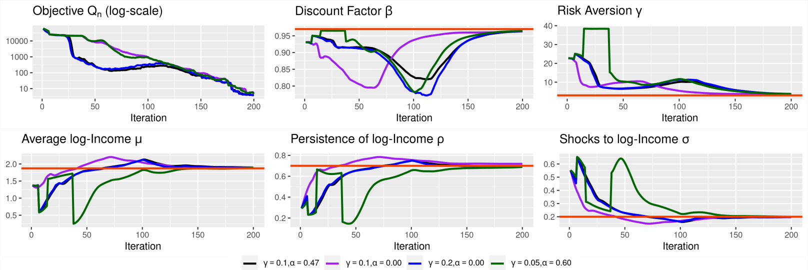
Figure 1: Aiyagari Model: local, global optimizers and sGN ($\varepsilon = 0.1$)



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

Results range of tuning parameters (one sample)

Figure 2: Aiyagari Model: sGN with different choices of tuning parameters ($\varepsilon = 0.1$)

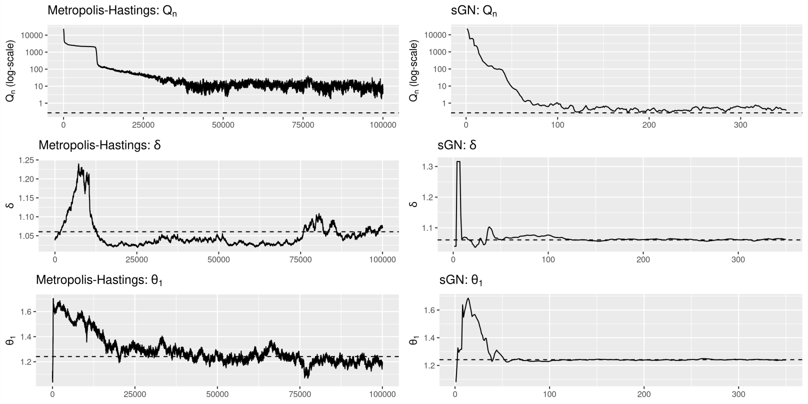


Legend: $n = 10000$, $T = 2$. $\gamma = 0.1$, $\alpha = 0.47$. sGN (black): Algorithm 1. Simulated-Annealing (dashed blue): 5000 iterations from θ_0. Simulated-Annealing + Nelder-Mead (dashed green): run Nelder-Mead after 5000 Simulated-Annealing iterations.

Empirical Example

Comparison with MCMC, distant starting value

Figure 3: Interdependent Duration Estimates: MCMC and sGN



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