cemmap
centre for microdata methods and practice

Higher-order properties of approximate estimators

Dennis Kristensen
Bernard Salanié

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP45/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Higher-order Properties of Approximate Estimators

Dennis Kristensen[*] **Bernard Salanié**[†]

September 16, 2013[‡]

Abstract

Many modern estimation methods in econometrics approximate an objective function, for instance, through simulation or discretization. These approximations typically affect both bias and variance of the resulting estimator. We provide a higher-order expansion of such "approximate" estimators that takes into account the errors due to the use of approximations. This expansion allows us to establish general conditions under which the approximate estimator is first-order equivalent to the exact estimator. Moreover, we use the expansion to propose adjustments of the approximate estimator that remove its first-order bias and adjust its standard errors. These adjustments apply to a broad class of approximate estimators that includes all known simulation-based procedures. We also propose another approach to reduce the impact of approximations, based on a Newton-Raphson adjustment. A Monte Carlo simulation on the mixed logit model shows that our proposed adjustments can yield spectacular improvements at a low computational cost.

[*]UCL, IFS and CREATES. E-mail: d.kristensen@ucl.ac.uk.

[†]Columbia University. E-mail: bsalanie@columbia.edu

[‡]This paper was previously circulated under the title "Higher-order Improvements of Approximate Estimators". We would like to thank participants at the "Conference on Dynamic Aspects in Economic Decision Making" (University of Copenhagen) and seminar audiences at Northwestern, Ohio State, Toulouse, UCL and Yale for helpful comments and suggestions. Kristensen acknowledges financial support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001, the Danish Social Sciences Research Council (through CREATES) and the European Research Council (Starting grant No. 312474). Parts of this research was conducted while Kristensen visited Princeton University and the University of Copenhagen, whose hospitality is gratefully acknowledged. Salanié gratefully acknowledges the Georges Meyer endowment for its support during a visit at the Toulouse School of Economics.

1 Introduction

The complexity of econometric models has grown steadily over the past three decades. The increase in computer power contributed to this development in various ways, and in particular by allowing econometricians to estimate more complicated models using methods that rely on approximations. A leading example is simulation-based inference, where a function of observables and parameters is approximated using simulations. In this case, the function is an integral such as a moment, as in the simulated method of moments (McFadden (1989), Duffie and Singleton (1993)) and in simulated pseudo-maximum likelihood (Laroque and Salanié (1989, 1993, 1994)). It may also be an integrated density/cdf, as in simulated maximum likelihood (Lee (1992, 1995)), Kolmogorov-Smirnov type statistics (Corradi and Swanson (2007)), or an integrated value function (Rust (1997)).[1] Then the approximation technique often amounts to Monte Carlo integration. Other numerical integration techniques may be preferred for low-dimensional integrals, e.g. Gaussian quadrature, or both techniques can be mixed (see for example Lee (2001)). Within the class of simulation-based methods, some nonparametric alternatives rely on kernel sums instead of integration (e.g. Fermanian and Salanié (2004); Creel and Kristensen (2012); Kristensen and Shin (2012)), or on sieve methods (Kristensen and Schjerning (2011); Norets (2012)). Other estimation methods involve numerical approximations, such as discretization of continuous processes, using a finite grid in the state space for dynamic programming models, and so on.

In all of these cases, we call the *approximator* the term that replaces the component of the objective function that we cannot evaluate exactly. Then the *exact estimator* is the infeasible estimator that reduces the approximation error to zero. In simulation-based inference, the exact estimator would be obtained with an infinite number of simulations; in dynamic programming models it would rely on an infinitely fine grid. We call the estimator that relies on a finite approximation an *approximate estimator*.

The use of approximations usually deteriorates the properties of the approximate estimator relative to those of the corresponding exact estimator: the former may suffer from additional biases and/or lose efficiency compared to the latter. When the approximation error is not stochastic, its main effect is to impart additional bias to the estimator. On the other hand, stochastic approximations not only create bias: they also reduce inefficiency. These are generic statements, of course. In some important special cases, such as the simulated method of moments, using approximations does not create additional bias, although it does reduce efficiency.

The effect of the approximation on the estimator can usually be reduced by choosing a sufficiently fine approximation; but this comes at the cost of increased computation time. In

[1] Simulation-based inference is surveyed in Gouriéroux and Monfort (1996), van Dijk, Monfort and Brown (1995) and Mariano, Schuerman and Weeks (2001) among others.

many applications this may be a seriously limiting factor; increased computer power helps, but it also motivates researchers to work on more complex models. It is therefore important to quantify the additional estimation errors that approximators generate, and also to account for these additional errors when drawing inference. As we will show, standard confidence intervals on the estimated parameters can be quite misleading unless they are properly adjusted for the errors induced by the approximation. As a first step in this direction, we analyze higher-order properties of the approximate estimator relative to the exact one in a very general setting. Moreover, we provide adjusted confidence intervals that correct for the leading bias and variance terms in a large class of approximate estimators.

Our findings apply to generalized method of moment estimators as well as M-estimators, both when the approximation is stochastic and when it is not. They encompass and extend results in the literature on simulation-based estimators, such as Lee (1995, 1999), Gouriéroux and Monfort (1996) and Laroque and Salanié (1989). Moreover, they can be used to analyze the behavior of estimators that rely on numerical approximation. Many structural estimates rely on such approximations, in asset pricing models (Tauchen and Hussey, 1991), DSGE models (Fernández-Villaverde, Rubio-Ramirez and Santos, 2006), or dynamic discrete choice models (Rust, 1997.) Our results also apply to many estimators used in empirical IO, which combine simulation and numerical approximation. To the best of our knowledge, this is the first paper to provide results for such a general class of models.

Our higher-order expansion can be used to adjust the approximate estimator to remove the leading term of the approximation errors for many common stochastic approximators; we call this "analytical adjustment". We also describe two alternative methods. One adjusts the estimator in a spirit similar to that of parametric bootstrap and jackknifing. The other is a two-step method where an initial approximate estimator is updated through one or several Newton-Raphson iterations based on the same objective function, but with a much finer degree of approximation.

To test the practical performance of our proposed adjustment methods, we run a simulation study on a mixed logit model. The mixed logit is one of the basic building blocks in much work in demand analysis, for instance; and it is simple enough that we can compute the true value of the biases and efficiency losses, as well as our estimated corrections. We show that uncorrected SML has non-negligible bias, even for large sample sizes; and that standard confidence intervals can be wildly off the mark. Our analytical adjustment removes most of the bias at almost no additional computational cost; and it yields very reliable confidence intervals. The Newton-Raphson correction also reduces the bias and improves confidence intervals, but it does so less effectively than the analytical adjustment.

The paper is organized as follows: Section 2 presents our framework and informally introduces the methods we propose to improve the properties of approximate estimators. In Section 3, we derive a higher-order expansion of the approximate estimator relative to the

exact one; this expansion allows us to identify the leading bias and variance terms. Then in Section 4 we build on the expansion to propose adjusted standard errors and confidence intervals. We describe our resampling and Newton-Raphson corrections in sections 5 and 6, respectively. Section 7 presents the results of a Monte Carlo simulation study using the mixed logit model as an example; and we discuss possible extensions of our results in section 8. All proofs and lemmas are in appendices.

2 Framework

At the most general level, our framework can be described as follows. Given a sample $\mathcal{Z}_n = \{z_1, ..., z_n\}$ of n observations, the econometrician proposes to estimate a parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^k$ through an estimating equation, $G_n(\theta, \gamma_0) = \mathcal{G}_n(\mathcal{Z}_n, \theta, \gamma_0(\mathcal{Z}_n, \hat{\theta}_n))$ that the estimator $\hat{\theta}_n$ is set to solve, $G_n(\hat{\theta}_n, \gamma_0) = 0$. The estimating equation depends on the data $\mathcal{Z}_n$ both directly and, possibly, via a nuisance parameter γ_0 that, possibly, is also a function of the unknown parameter. The parameter γ_0 could be finite-dimensional (e.g. it could be an estimated variance), but in most situations it is a function and so is infinite-dimensional. Our paper focuses on situations where the true function γ_0 is not known in closed form to the econometrician, and instead has to be approximated. In this case, a feasible estimator is obtained by solving the analog estimating equation $G_n(\theta, \hat{\gamma}_S) = 0$ w.r.t. θ, where $\hat{\gamma}_S$ is the chosen "approximator" that depends on some approximation scheme of order S (e.g. S simulations, or a discretization on a grid of size S). The resulting estimator, denoted $\hat{\theta}_{n,S}$, will be referred to as the " approximate" estimator. The main goal of the paper is then to analyze the (stochastic) differences between $\hat{\theta}_{n,S}$ and the "exact" estimator, $\hat{\theta}_n$.

Note that the above estimation problem is similar to two-step semiparametric estimation, where in the first step a nuisance parameter (γ_0) is replaced by its estimator (the approximator $\hat{\gamma}_S$), which in turn is used to obtain an estimator ($\hat{\theta}_{n,S}$) of θ_0. Some themes of that literature (see e.g Andrews 1994) and Chen et al 2003) recur in our analysis.

2.1 Examples

We now present a few examples that fall within the above setting.

Example 1: Simulated maximum likelihood (SML). Suppose we want to estimate a parameterized conditional distribution $p(y|x;\theta)$. The natural choice is the maximum-likelihood estimator, which maximizes $L_n(\theta, \gamma_0) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i;\theta)$. Sometimes the density p cannot be written in closed form. For example, in models with unobserved heterogeneity, $p(y|x;\theta) = \int w(y|x,\varepsilon;\theta) f(\varepsilon) d\varepsilon$ for some densities w and f. If the integral cannot be computed analytically, a simulated version can be obtained by drawing $\varepsilon_{i,s}$, $s = 1, ..., S$, from the distribution of f and obtain a simulated version by $\hat{p}_S(z;\theta) = S^{-1}\sum_{s=1}^{S} w(y|x,\varepsilon_s;\theta)$. The re-

sulting estimator is a SMLE. This fits in our framework, with $z = (y, x)$, $\gamma_0(z;\theta) := p(y|x;\theta)$, and

$$G_n(\theta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log \gamma}{\partial \theta}(z_i,\theta).$$

More recently, Fermanian and Salanié (2004) proposed using kernel estimators as approximators. Suppose that $y = r(x,\varepsilon;\theta_0)$, with implied conditional density $\gamma_0(z;\theta) = p(y|x,\theta)$. Then generate samples, $y_s(x,\theta) = r(x,\varepsilon_s;\theta)$ for $s = 1,\ldots,S$, and approximate the density γ_0 with a kernel density estimator based on the y_s's: $\hat{\gamma}_S(z;\theta) = \sum_{s=1}^{S} K_h(y - y_s(x,\theta))/S$. For a similar approach in time series models, see Altissimo and Mele (2009), Brownlees, Kristensen and Shin (2011) and Kristensen and Shin (2012).

Example 2: Simulated pseudo-maximum likelihood (SPML). Suppose that we have the following conditional moment restriction, $E[y|x] = m(x;\theta)$, where, for some function w and some unobserved error ε, $m(x;\theta) = E[w(x,\varepsilon;\theta)|x]$. If the conditional expectation m cannot be evaluated analytically, Laroque and Salanié (1989) proposed simulated pseudo-maximum likelihood (SPML) estimators: Draw i.i.d. random variables ε_s, $s = 1,...,S$, and define $\hat{m}_S(x;\theta) = S^{-1}\sum_{s=1}^{S} w(x,\varepsilon_s;\theta)$. Then an SNLS estimator is obtained by replacing m with $\hat{m}_S$. The above idea can be extended to incorporate information regarding the conditional variance of y. Again, this fits into our framework with $\gamma_0(x;\theta) = m(x;\theta)$ and $G_n(\theta,\gamma_0) = \partial L_n(\theta,\gamma_0)/(\partial\theta)$, where $L_n(\theta,\gamma_0) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma_0(x_i;\theta))^2$.

Example 3: Simulated method of moments (SMM). The parameter of interest is identified through a set of moment conditions $E[m(z,\theta_0)] = 0$. Given a weighting matrix W_n, the GMM estimator would minimize $L_n(\theta,\gamma_0) = M_n(\theta)'W_nM_n(\theta)$, where $M_n(\theta) = \sum_{i=1}^{n} m(z_i,\theta)/n$. Here, γ_0 is simply the function m, which may be hard to evaluate, as in the multinomial probit example of McFadden (1989). Another example is the simulated method of moments (SMM) proposed by Duffie and Singleton (1993) to estimate dynamic models where a long string of simulations from the model, say $\{y_s(\theta) : s = 1,...,S\}$, are used to approximate unconditional moments of the model. The resulting estimator is of the minimum-distance type. Creel and Kristensen (2012) generalize the approach of Duffie and Singleton (1993) to the case where a set of conditional moments are used in the estimation; they propose to approximate conditional expectations by combining simulations with kernel regression techniques.

Example 4: Non-stochastic approximators. Evaluating the value function in dynamic programming models most often requires numerical approximations that involve simulations, interpolation or sieve methods (also referred to as parametric approximations); see Rust (1997), and more recently Kristensen and Schjerning (2011) and Norets (2009, 2012.). Here the approximated value function plays the role of γ_0.

Similarly, many models used in macroeconomics are so complex that estimation is based on an approximate model, often by linearizing equations close to a steady state. The quality of the model approximation can be improved at a larger computational cost by using a finer grid or by using, for example, more iterations of perturbations or projection methods as advocated by Judd, Kubler and Schmedders (2003). For a first-order theoretical analysis of the impact on the resulting approximate MLE, see Fernández-Villaverde, Rubio-Ramirez and Santos (2006) and Ackerberg, Geweke and Hahn (2009).

Another example includes numerical inversion of funtions. One example of this arises in the estimation procedure proposed by Berry, Levinsohn and Pakes (1995) (BLP) to estimate discrete choice models in industrial organization. Here, observed market shares (s) are modelled as functions of unobserved (ξ) and observed (z) product characteristics, $s = Q(\xi, z; \theta)$ for some choice proability function Q. The function Q is usually computed by Monte-Carlo integration over unobserved individual preference shocks—a stochastic approximation scheme. The BLP estimation procedure requires the econometrician to compute the unobserved product characteristics given observed market shares by inverting the market share function with respect to its first argument, $\xi(s, z; \theta) = Q^{-1}(s, z; \theta)$. Since Q^{-1} is not available in closed form, this is done using a numerical fixed-point algorithm. It leads to an approximate solution, $\hat{\xi}_S(s, z; \theta)$, where S captures the number of iterations and/or the tolerance level used in the algorithm. Judd and Su (2012) and Dubé, Fox and Su (2012) recently emphasized that the quality of the estimates of θ is very sensitive to errors in the computation of the fixed point. They propose using mathematical programming under equilibrium constraints instead, with an augmented parameter vector (θ, ξ). This also fits into our framework, with the Lagrange multipliers for the market share equations as γ_0.

Their are many other examples of numerical approximators. For instance, the numerical approximation of derivatives or integrals also impacts the statistical properties of estimators (see e.g. Hong, Mahajan and Nekipelov 2010); and our methods apply there too.

* * *

In all of the examples above, approximations reduce the quality of the estimator. Start with our first three examples, where stochastic approximations (i.e. simulations) are used to evaluate a mathematical expectation. The mean of course is an unbiased estimator of the expectation; but in many simulation-based estimation methods the objective function depends nonlinearly on the simulated mean, so that the approximate estimator based on S simulations has an additional bias, along with reduced efficiency. In many cases both are of order $1/S$; this holds for example when the approximator simulates an expectation through a simple average. The efficiency loss may not be a concern in large samples; but the additional

bias persists asymptotically. When using nonparametric techniques such as kernel smoothers or sieve methods in the approximation, the approximator itself is biased, and the objective function will be biased even if the approximator enters linearly.

The simulated method of moments (Example 3) is a special case. This approximate estimator has nicer properties since the objective function is linear in the simulated mean; while the asymptotic efficiency loss still is of order $1/S$, the simulations do not impart bias to the estimator, except when kernel smoothers are employed as in Creel and Kristensen (2012).

Non-stochastic approximations also lead to deteriorations of the properties of the resulting estimators. Take the problem of computing the density $p(y|x;\theta)$ in Example 1 for instance. If the dimensionality of the integration variable (ε) is small, then instead of simulations the numerical integration may be done by an S point Gaussian quadrature, as in Lee (2001). Because this is a non-stochastic approximation method, the resulting approximate estimator will suffer from additional biases, but its variance will not increase.

As this informal discussion illustrates, the approximate estimator $\hat{\theta}_{n,S}$ often is consistent only if S goes to infinity as n goes to infinity; and $\sqrt{n}$-consistency requires that S diverges fast enough. In other words (Section 3 will give more precise statements and regularity conditions), $\|\hat{\theta}_{n,S} - \hat{\theta}_n\| = o_P(1/\sqrt{n})$ as $n \to \infty$ for some sequence $S = S(n) \to \infty$, in which case there is no first-order difference between the exact and approximate estimator. For finite S and n, our higher-order expansion allows the researcher to better evaluate the properties of the approximate estimator.

To derive the higher-order expansion of the approximate estimator, we need to impose regularity conditions both on the estimating equation and on the approximators. We present these conditions in the next two subsections.

2.2 Estimating Equation

We restrict our attention to the class of exact estimators $\hat{\theta}_n$ that (asymptotically) satisfy a first order condition of the form

$$G_n(\hat{\theta}_n, \gamma_0) = o_P\left(1/\sqrt{n}\right), \tag{1}$$

for some random functional $G_n(\theta, \gamma)$. The corresponding approximate estimator $\hat{\theta}_{n,S}$ similarly satisfies

$$G_n(\hat{\theta}_{n,S}, \hat{\gamma}_S) = o_P\left(1/\sqrt{n}\right). \tag{2}$$

Furthermore, we assume that $G_n(\theta, \gamma)$ takes the form of a sample average,

$$G_n(\theta, \gamma) = \frac{1}{n}\sum_{i=1}^{n} g(z_i; \theta, \gamma). \tag{3}$$

The above framework includes M-estimators and GMM-estimators, including all of the examples described in Section 2.1. In the case of M-estimators, $\hat{\theta}_{n,S} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} q\left(z_i;\theta,\hat{\gamma}_S\right)/n$ and we choose $g\left(z;\theta,\gamma\right) = \partial q\left(z;\theta,\gamma\right)/\left(\partial\theta\right)$. In the case of GMM estimators, $\hat{\theta}_{n,S} = \arg\max_{\theta\in\Theta} M_n(\theta,\hat{\gamma}_S)W_n M_n(\theta,\hat{\gamma}_S)$ with $W_n \xrightarrow{P} W$ and $M_n(\theta,\gamma) = \sum_{i=1}^{n} m\left(z_i;\theta,\gamma\right)/n$. Defining $g\left(z_i;\theta,\gamma\right) = H_0 W m\left(z_i;\theta,\gamma\right)$, where $H_0 = E\left[\partial m\left(z_i;\theta,\gamma\right)/\partial\theta\right]$, $\hat{\theta}_{n,S}$ solving (2) is (asymptotically) first-order equivalent to the original GMM estimator.

We assume that the function of interest $\gamma_0 : \mathcal{Z}\times\Theta \mapsto \mathbb{R}^p$ belongs to a linear function space Γ equipped with a norm $\|\cdot\|$. In most cases, the norm will be the L_q-norm induced by the probability measure associated with the data generating process, $\|\gamma\| = E\left[\|\gamma(z)\|^q\right]^{1/q}$ for some $q \geq 1$. We introduce the first-order derivative of $G_n\left(\theta,\gamma\right)$ w.r.t. θ,

$$H_n\left(\theta,\gamma\right) = \frac{1}{n}\sum_{i=1}^{n} h\left(z_i;\theta,\gamma\right), \quad \text{with} \quad h\left(z_i;\theta,\gamma\right) = \frac{\partial g\left(z_i;\theta,\gamma\right)}{\partial\theta}, \tag{4}$$

and the corresponding population versions,

$$G\left(\theta,\gamma\right) = E\left[g\left(z_i;\theta,\gamma\right)\right], \quad H\left(\theta,\gamma\right) = E\left[\frac{\partial g\left(z_i;\theta,\gamma\right)}{\partial\theta}\right].$$

We first impose conditions to ensure that the exact, but infeasible estimator is well-behaved:

A.1 (i) $\hat{\theta}_n \xrightarrow{P} \theta_0$ which lies in the interior of the parameter space Θ; (ii) $\{z_i\}$ is stationary and geometrically α-mixing; (iii) $E\left[\left\|g\left(z_i;\theta_0,\gamma_0\right)\right\|^{2+\delta}\right] < \infty$ for some $\delta > 0$; (iv) $G\left(\theta_0,\gamma_0\right) = 0$.

A.2 For all γ in a neighbourhood $\mathcal{N}$ of γ_0, $h\left(z;\theta,\gamma\right)$ satisfies: (i) $H_0 := H\left(\theta_0,\gamma_0\right)$ is positive definite; (ii) for some $\delta > 0$, $E[\sup_{\|\theta-\theta_0\|<\delta}\|h\left(z_i;\theta,\gamma_0\right)\|] < \infty$; (iii) for some $\delta,\lambda,\bar{H} > 0$, and for all $\gamma\in\mathcal{N}$,

$$E\left[\sup_{\|\theta-\theta_0\|<\delta}\left\|h\left(z_i;\theta,\gamma\right) - h\left(z_i;\theta,\gamma_0\right)\right\|\right] \leq \bar{H}\left\|\gamma-\gamma_0\right\|^{\lambda}.$$

Assumption A.1(i) requires that the infeasible estimator be consistent; Lemma 1 below provides a set of sufficient conditions. A.1(ii) rules out strongly persistent data, thereby allowing us to obtain standard rates of convergence for the resulting estimators. In particular, A1(ii) and A.1(iii) together imply that a central limit theorem (CLT) applies to $G_n\left(\theta_0,\gamma_0\right)$. The geometric mixing condition could be weakened, but this would lead to more complicated results; we refer the reader to Kristensen and Shin (2012) for some results on approximate estimators based on strongly persistent and/or non-stationary data (and thereby estimators with non-standard rates) in the context of SMLE.

Assumption A.2 imposes differentiability of $\theta \mapsto g(z;\theta,\gamma)$. In particular, when γ depends on θ (as is the case for all of our examples), it requires the approximator to be a smooth function of θ. Therefore A.2 rules out discontinuous and non-differentiable approximators, such as the simulated method of moment estimators for discrete choice models proposed in McFadden (1989) and Pakes and Pollard (1989) which involve indicator functions. These cases can be handled by introducing a smoothed version of the approximators as discussed in McFadden (1989); see also Fermanian and Salanié (2004). Alternatively, one could extend our results by resorting to empirical process theory, as done in the work by Armstrong et al (2012) on simulation-based estimators. The Lipschitz condition imposed on $h(z;\theta,\gamma)$ is used to ensure that $H_n(\theta,\hat{\gamma}_S) \xrightarrow{P} H(\theta,\gamma)$ uniformly in θ as $\hat{\gamma}_S \xrightarrow{P} \gamma$.

Since our focus is on higher-order properties of the approximate estimator, we also assume that it is consistent, so that we can conduct our analysis locally around θ_0:

A.3 $\hat{\theta}_{n,S} \xrightarrow{P} \theta_0$ as $n, S \to \infty$.

A set of sufficient conditions (similar to those in Newey and McFadden, 1994 for consistency of two-step semiparametric estimators) for Assumptions A.1 (i) and A.3 to hold are provided in the following lemma:

Lemma 1 *Suppose that* $\hat{\theta}_{n,S} = \arg\max_{\theta\in\Theta} Q_n(\theta,\hat{\gamma}_S)$ *where: (i)* Θ *is compact; (ii)* $\hat{\gamma}_S \xrightarrow{P} \gamma_0$*; (ii) either* $\sup_{\theta\in\Theta,\|\gamma-\gamma_0\|<\delta} |Q_n(\theta,\gamma) - Q(\theta,\gamma)| \xrightarrow{P} 0$ *or* $|Q_n(\theta,\gamma_1) - Q_n(\theta,\gamma_2)| \leq B_n \|\gamma_1 - \gamma_2\|$ *for all* γ_1, γ_2 *in a neighbourhood of* γ_0 *where* $B_n = O_P(1)$ *and* $\sup_{\theta\in\Theta} |Q_n(\theta,\gamma_0) - Q(\theta,\gamma_0)| \xrightarrow{P} 0$*; (iv)* $\theta \mapsto Q(\theta,\gamma_0)$ *is continuous and has a unique maximum at* θ_0*. Then A.1(i) and A.3 hold.*

As a first step in our higher-order analysis, we prove in appendix B (Lemma 6) that

$$\hat{\theta}_{n,S} - \hat{\theta}_n = -H_0^{-1}\{G_n(\theta_0,\hat{\gamma}_S) - G_n(\theta_0,\gamma_0)\} + o_P\left(1/\sqrt{n}\right). \tag{5}$$

If γ was a finite-dimensional pameter, we could use a Taylor expansion of $G_n(\theta_0,\hat{\gamma}_S) - G_n(\theta_0,\gamma_0)$ to analyze the additional estimation errors due to the approximator $\hat{\gamma}_S$. However, γ may be a function; for such a functional expansion to be well-defined and for the individual terms in the expansion to be well-behaved, we need to impose some further regularity conditions on $g(z_i;\theta_0,\gamma)$ as a functional of γ. In all of the following, $d\gamma \in \Gamma$ denotes a small change around γ_0.

A.4(m) There exist functionals $\nabla^k g\left(z;\theta_0\right)\left[d\gamma_1,...,d\gamma_k\right]$, $k=1,...,m$, which are linear in each component $d\gamma_i \in \Gamma$, $i=1,...,k$, and constants $\delta>0$ and $\bar{G}_i>0$, $i=0,1,2$, such that:

$$E\left[\left\|g\left(z;\theta_0,\gamma_0+d\gamma\right)-g\left(z;\theta_0,\gamma_0\right)-\sum_{k=1}^{m}\frac{1}{k!}\nabla^k g\left(z;\theta_0\right)\left[d\gamma,...,d\gamma\right]\right\|\right] \leq \bar{G}_0\left\|d\gamma\right\|^{m+1}, \tag{6}$$

where, for $k=2,...,m$ and for some $\nu>0$,

$$E\left[\left\|\nabla g\left(z;\theta\right)\left[d\gamma\right]\right\|^2\right] \leq \bar{G}_1\left\|d\gamma\right\|^2, \tag{7}$$

$$E\left[\left\|\nabla^k g\left(z;\theta_0\right)\left[d\gamma_1,...,d\gamma_k\right]\right\|^{2+\nu}\right] \leq \bar{G}_k\left(\left\|d\gamma_1\right\|\cdots\left\|d\gamma_k\right\|\right)^{2+\nu}. \tag{8}$$

Assumption A.4(m) restricts $g\left(z;\theta_0,\gamma\right)$ to be m times pathwise differentiable w.r.t. γ with differentials $\nabla^k g\left(z;\theta_0\right)\left[d\gamma_1,...,d\gamma_k\right]$, $k=1,...,m$. These differentials are required to be Lipschitz in $d\gamma_1,...,d\gamma_k$. For a given choice of m, this allows us to use an mth order expansion of $G_n\left(\theta,\gamma\right)$ w.r.t. γ to evaluate the impact of $\hat{\gamma}_S$. In particular, the difference between the approximate and the exact objective functions can be written as

$$G_n(\theta_0,\hat{\gamma}_S)-G_n(\theta_0,\gamma_0)=\sum_{k=1}^{m}\frac{1}{k!}\nabla^k G_n(\theta_0)[\hat{\gamma}_S-\gamma_0,...,\hat{\gamma}_S-\gamma_0]+R_{n,S}, \tag{9}$$

where $R_{n,S}=O_P(\|\hat{\gamma}_S-\gamma_0\|^{m+1})$ is the remainder term, and

$$\nabla^k G_n(\theta_0)\left[d\gamma_1,...,d\gamma_m\right]=\frac{1}{n}\sum_{i=1}^{n}\nabla^k g\left(z_i;\theta_0\right)\left[d\gamma_1,...,d\gamma_k\right]. \tag{10}$$

To evaluate the higher-order errors due to the approximation, we will study the mean and variance of each of the terms in the sum on the right hand side of Eq. (9).

For a wide range of objective functions, such as the ones of SPML and SMM, Assumption A.4 is satisfied. However, in some cases minor modifications of the estimating equations such as trimming will be required; here is an example.

Example 1: SML (continued). Sometimes g is pathwise differentiable but the bound in equation (6) does not apply. Take SML, where $g\left(z;\theta,\gamma\right)=\dot{\gamma}\left(z;\theta\right)/\gamma\left(z;\theta\right)$, $\dot{\gamma}\left(z;\theta\right)=\partial\gamma\left(z;\theta\right)/\left(\partial\theta\right)$, is the score of the log-likelihood and we approximate the density $\gamma\left(z;\theta\right)$. Then, suppressing dependence on $\left(z;\theta\right)$,

$$\nabla g\left[d\gamma\right]=\frac{1}{\gamma_0}d\dot{\gamma}-\frac{\dot{\gamma}_0}{\gamma_0^2}d\gamma, \quad \nabla^2 g\left[d\gamma,d\gamma\right]=-\frac{2}{\gamma_0^2}d\gamma d\dot{\gamma}+\frac{2\dot{\gamma}_0}{\gamma_0^3}d\gamma^2, \text{ etc,}$$

so that $\bar{G}_0$ in A.4 involves higher-order moments of $1/\gamma$. If the density $\gamma(z;\theta_0) \to 0$ as $\|z\| \to \infty$, these moments may not be finite. One way out is to introduce trimming, replacing the simple simulator $\hat{\gamma}_S(z;\theta)$ described above with $\hat{\gamma}_{a,S}(z;\theta) = \hat{\gamma}_S(z;\theta)\,\tau_a(\hat{\gamma}_S(z;\theta))$ where $\tau_a(w)$ is a smooth trimming function that satisfies $\tau_a(w) = 1$ for $w \geq 2a$ and $\tau_a(w) = 0$ for $w \leq a$. Then $\bar{G}_{a,0} = O\left(a^{-(m+1)}\right)$ is finite for any $a > 0$, and the remainder term satisfies $R_{n,S} = O_P(a^{-(m+1)} \left\|\hat{\gamma}_{a,S} - \gamma_0\right\|^{m+1})$. By letting $a = a_S \to 0$ at a suitable rate as $S \to \infty$, it is now possible to control the remainder term while the expansion remains valid; see Creel and Kristensen (2012) and Kristensen and Shin (2012) for more details in the context of SMM and SMLE, respectively.

2.3 Approximators

To analyze the impact of approximations, we need to further specify how the approximator behaves. Let us first introduce two alternative ways of implementing the approximation: Either one common approximator is used across all observations, or a new approximator is used for each observation. To differentiate between the two approximation schemes, we will refer to the approximate estimator based on the first scheme as an *estimator based on common approximators* (ECA) and to the second one as an *estimator based on individual approximators* (EIA). In the first case, the approximate sample moment takes the form

$$\text{ECA}: G_n(\theta, \hat{\gamma}_S) = \frac{1}{n}\sum_{i=1}^{n} g(z_i;\theta,\hat{\gamma}_S), \tag{11}$$

with one single approximator, $\hat{\gamma}_S$, being used in the computation of the moment conditions. In the second case,

$$\text{EIA}: G_n(\theta, \hat{\gamma}_S) = \frac{1}{n}\sum_{i=1}^{n} g\left(z_i;\theta,\hat{\gamma}_{i,S}\right), \tag{12}$$

and n approximators, $\hat{\gamma}_{1,S}, ...\hat{\gamma}_{n,S}$, are used in the computation. We stress that the ECA and EIA are both targeting the same infeasible estimator; the only difference lies in how the approximators are used in the computation of the objective function.

Take simulation-based estimation for instance. Earlier papers (e.g. Laroque and Salanié 1989, McFadden 1989) used EIAs, and most papers on cross-sectional or panel data still do. ECAs were proposed by Lee (1992) for cross-sectional discrete choice models, but they have been more useful in dynamic models where one long trajectory of the model is simulated and used to compute simulated moments (see Example 3) or densities (see Example 1).

When the number of approximators remains fixed, as in ECAs, the resulting approximate estimator is similar to semiparametric two-step estimators where in the first step a function is nonparametrically estimated, see e.g. Andrews (1994) and Chen et al (2003). In contrast, EIAs employ n approximators—one for each observation. Thus, the dimension of $\hat{\gamma}_S =$

$\left(\hat{\gamma}_{1,S}, \ldots, \hat{\gamma}_{n,S}\right)$ increases with sample size, and EIAs give rise to an incidental parameters problem. Some of our results for this situation are similar to those found in the literature on higher-order properties and bias-correction of estimators in an incidental parameters setting, see e.g. Hahn and Newey (2004) and Arellano and Hahn (2007).

To provide a streamlined set of regularity conditions that apply to both of these approximation schemes, we let $J \geq 1$ denote the number of approximators used in the computation of $\hat{\theta}_{n,S}$. For ECAs and EIAs, $J = 1$ and $J = n$, respectively. In what follows, it is crucial to separate assumptions on the bias of the approximator

$$b_S(z;\theta) := E[\hat{\gamma}_{j,S}(z;\theta)|x] - \gamma_0(z;\theta) \tag{13}$$

from assumptions on its stochastic component (which is by definition zero for non-stochastic approximators):

$$\psi_{j,S}(z;\theta) := \hat{\gamma}_{j,S}(z;\theta) - E\left[\hat{\gamma}_{j,S}(z;\theta)|z\right]. \tag{14}$$

A.5(p) The approximator(s) lies in Γ and satisfies:

(i) for any fixed value z, the J ($= 1$ or $= n$) random functions $\hat{\gamma}_{1,S}(z;\theta), ..., \hat{\gamma}_{J,S}(z;\theta)$ are identically distributed, mutually independent and independent of $\mathcal{Z}_n$.

(ii) The bias b_S is of order $\beta > 0$:

$$b_S(z;\theta) = S^{-\beta}\bar{b}(z;\theta) + o(S^{-\beta}).$$

(iii) For $2 \leq q \leq p$, the stochastic component of the approximator satisfies:

$$E\left[\left\|\psi_{j,S}(z;\theta)\right\|^q\right] = S^{-\alpha_q} v_q(z;\theta) + o(S^{-\alpha_q}),$$

for some constant $\alpha_q > 0$.

Note that the $o(\cdot)$ terms in (ii)-(iii) are w.r.t. the function norm on Γ. Assumption A.5 is sufficiently general to cover all of the examples in Section 2 under suitable regularity conditions. First consider A.5(iii). It requires that the approximator have p moments and that each of these vanish at a given rate as $S \to \infty$. We will choose p in conjunction with the order of the expansion m of Assumption A.4, since we wish to evaluate the mean and variance of each of the higher-order terms. For example, in order to ensure that the variance of $\nabla^k G_n(\theta_0)[\hat{\gamma}_S, ..., \hat{\gamma}_S]$ exists and to evaluate its rate of convergence, we will require A.5(p) to hold with $p = 2k$.

2.3.1 Non-stochastic approximators

First, consider an approximation that does not involve any randomness, as with numerical integration, discretization, or numerical inversion of a function. A.5(i) clearly has no bite when non-stochastic approximators are used. Then by construction the conditional variance of the approximator is zero, so that $\alpha_p = +\infty$ for all $p \geq 2$. Non-stochastic approximation imparts a bias, which in leading cases obeys assumption A.5(ii) for some $\beta > 0$. As we will see, our general theory also applies in this case. However, with numerical approximators the bias b_S is often hard to correct for using analytical adjustments. This is one reason why we propose alternative approaches in sections 5 and 6.

2.3.2 Stochastic approximators

Next, let us examine stochastic approximation schemes, which encompass simulation-based inference methods. Most simulation-based estimators in a dynamic setting use the ECA scheme: only one approximator is used for all observations, c.f. the discussion in Example 3, and so A.5(i) is automatically satisfied. The typical EIA scheme draws S independent batches of size n; this again satisfies A.5(i). Our assumptions do not rule out dependence between the simulated values within each simulated sample: in time series models for instance, we only neeed to draw from the assumed distribution of the innovations. Note that A.5(i) is stated for some *fixed* value of z; the requirement that the simulations be independent of data is satisfied by most standard simulation schemes[2]. For parametric approximators in simulation-based inference, the bias b_S is typically zero and A.5(ii) holds with $\beta = \infty$.

Monte Carlo schemes are of course the most prominent example of stochastic approximators; and they have specific properties that allow for a more precise analysis of the approximation error appearing in the resulting estimator. We will therefore specialize some of our results to the following class of Monte Carlo approximators:

A.6(p) The approximator $\hat{\gamma}_{j,S}(z;\theta)$ takes the form

$$\hat{\gamma}_{j,S}(z;\theta) = \frac{1}{S}\sum_{s=1}^{S} w_S(z, \varepsilon_{j,s};\theta), \tag{15}$$

and that for each $j = 1, ..., J$,

[2]There is one situation where the independence assumption is violated: sequential approximation schemes used in dynamic latent variable models such as particle filters, see e.g. Brownlees, Kristensen and Shin (2011) and Olsson and Rydén (2008). Then the approximator of the conditional density of the current observation depends on that used for the previous observation.

- $\{\varepsilon_{js}\}_{s=1}^{S}$ is stationary and geometrically β-mixing[3]
- $\{\varepsilon_{js}\}_{s=1}^{S}$ and $\{\varepsilon_{ks}\}_{s=1}^{S}$ are independent for $j \neq k$, and they are all independent of the sample.

The function $w_S(z, \varepsilon_{js}; \theta)$ satisfies

$$\bar{w}_S(z;\theta) := E[w_S(z, \varepsilon_{js}; \theta)|x] = \gamma_0(z;\theta) + S^{-\beta}\bar{b}(z;\theta) + o\left(S^{-\beta}\right)$$

and

$$E[\|w_S(z, \varepsilon_{js}; \theta) - \bar{w}_S(z;\theta)\|^p] = O(S^{\mu_p}) \text{ for some } \mu_p < p/2.$$

To our knowledge, the class of approximators that satisfies A.6 includes all simulation-based approximators proposed in the literature. The bias and variance of approximators that obey (15) follow directly from those of the simulators w_S: it is easy to see that Assumption A.6 implies A.5 with the same rate β for the bias term and with $\alpha_p = p/2 - \mu_p > 0$ in A.5(iii).

In parametric simulation-based estimation, the simulating function $w_S \equiv w$ in Assumption A.6 is typically independent of the number of simulations, and the approximator has no bias: $b_S \equiv 0$ and so $\beta = \infty$. Moreover, Assumption A.6(iii) typically holds with $\mu_p = 0$, and A.5(iii) with $\alpha_p = p/2$.

The class of approximators in Assumption A.6 allows for nonparametric techniques such as the methods proposed in Fermanian and Salanié (2004), Creel and Kristensen (2012), Kristensen and Scherning (2012) and Norets (2009, 2012.) These also have a bias component, but A.5 still applies. As an example, consider the NPSML estimator: $w_S(y, x, \varepsilon_s; \theta) = K_h(y_s(x, \theta) - y)$ where the bandwidth $h = h(S) \to 0$ as $S \to \infty$. Let $d = \dim(y)$ and suppose that we use a kernel of order r. The bias component satisfies

$$\bar{w}_S(y, x; \theta) = p(y|x;\theta) + h^r \frac{\partial^r p(y|x;\theta)}{\partial y^r} + o(h^r),$$

Furthermore, it is easily checked that $E[|K_h(y_s(x, \theta) - x)|^p|x] = O\left(-h^{d(p-1)}\right)$ for all $p \geq 2$ under suitable regularity conditions. Thus, with a bandwidth of order $h \propto S^{-\delta}$ for some $\delta > 0$, A.5 holds with $\beta = r\delta$ and $\mu_p = \delta d(p-1)$ for $p \geq 2$.

As is well-known, the asymptotic mean integrated squared error is smallest when the bias and variance component are balanced. This occurs when $\delta^* = 1/(2r + d)$, leading to $\beta = r/(2r + d)$. We recover of course the standard nonparametric rate[4].

[3]This is only used in the proof of Theorem 4. It could be weakened to "strongly mixing" elsewhere, but we maintain the assumption of β-mixing throughout to streamline the assumptions.

[4]While the standard nonparametric rate is optimal for the approximation of the individual densities that make up the likelihood, this rate does not yield the best NPSML estimators. This is akin to results for semi-parametric two-step estimators where undersmoothing of the first-step nonparametric estimator is normally

3 Effects of Approximations

We are now ready to derive the leading bias and variance terms of the estimator due to approximation errors. In the following, when we discuss biases and variances and, for example, write $E[\hat{\theta}_{n,S}]$, we refer to the means and variances of the leading terms of a valid stochastic expansion of the estimators. This is a standard approach in the higher-order analysis of estimators; see, for example, Rothenberg (1984) and Newey and Smith (2004, section 3).

3.1 Higher-order Expansion

To state the asymptotic expansion in a compact manner, we introduce some additional notation and moments which will make up the leading bias and variance terms. Let $g_i := g(z_i;\theta_0,\gamma_0)$, $\nabla g_i[d\gamma] := \nabla g\left(z_i;\theta_0,\gamma_0\right)[d\gamma]$ and $\nabla^2 g_i[d\gamma,d\gamma] := \nabla^2 g\left(z_i;\theta_0,\gamma_0\right)[d\gamma,d\gamma]$ for any function $d\gamma$. As we will see, the leading terms in the bias of the approximate estimator are

$$B_{S,1} = -H_0^{-1}E\left[\nabla g_i[b_S]\right] \quad \text{and} \quad B_{S,2} = -\frac{1}{2}H_0^{-1}E\left[\nabla^2 g_i[\psi_S,\psi_S]\right], \tag{16}$$

where b_S and ψ_S are defined in eqs. (13)-(14). The first bias term $B_{S,1}$ is zero for unbiased approximators, as in parametric simulation-based inference. The second one, $B_{S,2}$, is zero for non-stochastic approximators of the type found in numerical approximation schemes.

The leading variance term due to the presence of approximations is $\nabla G_n(\theta_0)[\hat{\gamma}_S - \gamma]$. It can be decomposed into two terms. The first one is

$$D_{n,S} = \frac{1}{n}\sum_{i=1}^{n} d_{i,S}, \quad \text{with} \quad d_{i,S} = \nabla g_i[b_S] - E\left[\nabla g_i[b_S]\right],$$

which is common to the two approximation schemes. The form of the second variance component, $E_{n,S}$, on the other hand, depends on whether EIA or ECA is implemented. In both cases, $E_{n,S}$ is a sample average:

$$E_{n,S} = \frac{1}{n}\sum_{i=1}^{n} e_{i,S};$$

but $e_{i,S} := \nabla g_i[\psi_{i,S}]$ for EIA, while $e_{i,S} := \nabla g_i[\psi_S]$ for ECA.

The variance components $\psi_{i,S}$ vary across observations for EIAs; as a consequence, one can directly apply a CLT to $E_{n,S}$ to obtain that, for any fixed $S \geq 1$ as $n \to \infty$, $\sqrt{n}E_{n,S} \to^d N\left(0,\Omega_S^E\right)$ where $\Omega_S^E = \mathrm{Var}(e_{i,S})$.

On the other hand, ECAs only have one ψ_S, which is common across observations; and

required for the parametric estimator to be $\sqrt{n}$-consistent; see Kristensen-Salanié (2010) for details. For example, the optimal rate for NPSML estimation turns out to be $\delta^{**} = 1/(r+d+2)$. This involves undersmoothing, except when standard second-order kernels are employed ($r = 2$). Then the rate that minimizes the AMISE of the kernel estimator is also optimal for the MSE of $\hat{\theta}_{n,S}$: $\delta^* = \delta^{**} = 1/(4+d)$.

getting a CLT takes more work and additional assumptions. We start by rewriting $E_{n,S}$ as

$$E_{n,S} = \frac{1}{n}\sum_{i=1}^{n}\{e_{i,S} - \nabla G[\psi_S]\} + \nabla G[\psi_S], \quad \text{with } \nabla G[\psi_S] := E[\nabla g_i[\psi_S]|\psi_S].$$

The first term is $O_P\left(S^{-\alpha_2/2}/\sqrt{n}\right)$, while the second term $\nabla G[\psi_S] = O_P\left(S^{-\alpha_2/2}\right)$ dominates it. In general, the large-sample distribution of $\nabla G[\psi_S]$ is not known in closed-form. However, if we strengthen Assumption A.5 to A.6, we can write

$$\nabla G[\psi_S] = \frac{1}{S}\sum_{s=1}^{S}\nabla G[\bar{w}_{s,S}], \quad \text{with } \bar{w}_{s,S} := E[w_S(z_i, \varepsilon_s; \theta_0)|\varepsilon_s] - E[w_S(z_i, \varepsilon_s; \theta_0)], \quad (17)$$

and a CLT can be applied as $S \to \infty$.

The above terms make up the first-order expansion of the effects of approximations on the estimators:

Theorem 2 *Assume A.1-A.3, A.4(2), and A.5(4). Also assume that* $\sqrt{n}S^{-3\beta} \to 0$ *and* $\sqrt{n}S^{-\alpha_3} \to 0$. *Then:*

$$\hat{\theta}_{n,S} - \theta_0 = B_{S,1} + B_{S,2} + H_0^{-1}\{G_n + D_{n,S} + E_{n,S}\} + o_P\left(1/\sqrt{n}\right), \quad (18)$$

where we abbreviate $G_n = G_n(\theta_0, \gamma_0)$. *The two sequences* $(G_n, D_{n,S})$ *and* $E_{n,S}$ *are asymptotically mutually independent. Moreover, the following limit results hold:*

- *For both EIA and ECA approximators, for any fixed* $S \geq 1$ *and as* $n \to \infty$,

$$\sqrt{n}\{G_n + D_{n,S}\} \xrightarrow{d} N\left(0, \Omega_S^{G+D}\right), \quad \text{with } \Omega_S^{G+D} = \sum_{i=-\infty}^{\infty}\text{Cov}(g_0 + d_{0,S}, g_i + d_{i,S});$$

  As $S \to \infty$, *the bias terms have orders* $B_{S,1} = O(S^{-\beta})$ *and* $B_{S,2} = O(S^{-\alpha_2})$; *and* $\Omega_S^{G+D} = \Omega^G + O(S^{-2\beta})$, *where* $\Omega^G = \sum_{i=-\infty}^{\infty}\text{Cov}(g_0, g_i)$.

- *For EIA approximators, for any fixed* $S \geq 1$ *and as* $n \to \infty$,

$$\sqrt{n}E_{n,S} \xrightarrow{d} N\left(0, \Omega_S^E\right), \text{ with } \Omega_S^E = \text{Var}(e_{i,S}) = O\left(S^{-\alpha_2}\right).$$

- *For ECA approximators, as* $S \to \infty$, $\text{Var}(E_{n,S}) = O_P(S^{-\alpha_2})$. *If in addition Assumption A.6(4) holds with a* $w_S \equiv w$ *that does not depend on* S, *then* $\alpha_2 = 1$ *and*

$$\sqrt{S}E_{n,S} \xrightarrow{d} N\left(0, \Omega^E\right), \text{ with } \Omega^E = \text{Var}(\nabla G[\bar{w}_s]),$$

  where $\bar{w}_s = \bar{w}_{s,S}$ *is defined in eq. (17).*

The above expansion allows us to analyze the effects due to approximation errors. In particular, both EIA's and ECA's are normally distributed in large samples with leading bias and variance terms due to approximations given by:

$$\begin{aligned} E[\hat{\theta}_{n,S} - \theta_0] &\simeq B_{S,1} + B_{S,2} = O(S^{-\beta}) + O(S^{-\alpha_2}), \\ n\text{Var}(\hat{\theta}_{n,S} - \theta_0) - H_0^{-1}\Omega^G H_0^{-1} &\simeq H_0^{-1}\left\{O(S^{-2\beta}) + n\text{Var}(E_{n,S})\right\} H_0^{-1}, \end{aligned}$$

where $n\text{Var}(E_{n,S}) = O\left(S^{-\alpha_2}\right)$ for EIAs and $n\text{Var}(E_{n,S}) = O\left(nS^{-\alpha_2}\right)$ for ECAs.

The bias and the variance of the approximator enter the two leading bias terms separately: the bias b_S drives $B_{S,1}$, and the stochastic component ψ_S drives $B_{S,2}$. When the approximator is a simple unbiased simulated average, $B_{S,1} = 0$ and the leading bias term $B_{S,2} = O\left(1/S\right)$; this is a well-known result for specific simulation-based estimators in cross-sectional settings, see e.g. Gouriéroux-Monfort (1996). Our theorem shows that this result holds more generally under weak regularity conditions.

EIA's and ECA's differ regarding the second variance term $E_{n,S}$. In the computation of the ECA, one common approximator is used across all observations; this introduces additional correlations across observations. In contrast, for EIA, $\psi_{i,S}$ and $\psi_{k,S}$ are independent for $i \neq k$. As a consequence, we expect the variance due to simulations to be larger for ECA's; and in leading simulation-based inference cases with $\beta = \infty$ and $\alpha_2 = 1$, we need S to go to infinity faster than n to keep the varinace from exploding. This suggests that one should prefer EIA to ECA; but it may be necessary to sacrifice statistical efficiency for the sake of computational efficiency. This is, for example, the case for SMM with dependent data where $\hat{\gamma}_S$ is a single simulated moment that can be used across all observations. Similarly, when $\hat{\gamma}_S$ is computationally costly to implement, such as the sieve approximator of a value function employed in Kristensen and Scherning (2012), it is convenient to use the same approximator across all observations.

3.2 Sharpness

The sharpness of the rates in Theorem 2 depends on the type of approximator being used and how it enters into the objective function; that is, the precise nature of the mapping $\gamma \mapsto g\left(z, \theta, \gamma\right)$.

Theorem 3 *Under the assumptions of Theorem 2, if the rates in Assumption 5 are sharp then*

- *For non-stochastic approximators, all rates listed in the Theorem are sharp.*
- *For EIA's with $\nabla^2 g_i[\gamma, \gamma] \neq 0$, the rates of $B_{S,1}$ and $B_{S,2}$ and $D_{n,S}$ and $E_{n,S}$ are sharp. If additionally Assumption A.(4)6 holds with $w_S \equiv w$, the same is true for ECA's.*

The proof of Theorem 3 follows from the arguments in the proof of Theorem 2 together with rate results for sample averages. Note that it does not cover nonparametric simulators, for which w_S depends on S through the bandwidth. For example, when $\hat{\gamma}_S$ is a kernel estimator, one can show that $\Omega_S^E = O(S^{-1})$, which is sharper than the rate stated in the theorem; see Creel and Kristensen (2012) and Kristensen and Shin (2012).

In some special cases, a term in the expansion is zero. In SMM for instance, the function g is linear in the approximator γ. Then $\nabla^2 g_i[d\gamma, d\gamma] = 0$, so that $B_{S,2} = 0$; and our rates are obviously not sharp.

3.3 First-order efficiency

Our results allow us to provide rates on the degree of approximation under which the approximate estimator is asymptotically first-order equivalent to the exact estimator; that is, which choices of the sequence $S = S_n$ guarantee $\|\hat{\theta}_{n,S_n} - \hat{\theta}_n\| = o_P\left(n^{-1/2}\right)$. In general, asymptotic equivalence for ECAs obtain if $n/S^{\min(\alpha_2, 2\beta)} \to 0$; for EIA's we have a weaker condition, replacing α_2 with $2\alpha_2$.

For parametric simulation-based estimators ($\beta = 0$, $\alpha_2 = 1$), this gives the standard result that n/S_n should go to zero for ECA's (Duffie and Singleton, 1993; Lee, 1995, Theorem 1), while $\sqrt{n}/S_n$ should go to zero for EIA's (Laroque and Salanié, 1989; Lee, 1995, Theorem 4).

As a more complicated example, consider the case where nonparametric kernel methods are used. Using standard arguments from the literature on semiparametric estimation, one can show in great generality that $\Omega_S^E = O(S^{-1})$ in the case of ECA (see Kristensen and Shin, 2012 for further details). Given this result, it easily follows from Theorem 2 that for the NPSMLE based on ECA's to be equivalent to the MLE, we need $\sqrt{n}h^r \to 0$, $n/S \to 0$ and $\sqrt{n}/\left(Sh^d\right)^2 \to 0$.

4 Analytical Adjustments

The expansion derived in the previous section naturally suggests corrections of the approximate estimators and standard errors to take into account the biases and variances due to approximations. The corrections are obtained by constructing consistent estimators of the first terms in the relevant formulæ of Theorem 2. It turns out that some cases are easier to deal with than others; we examine bias and variance in turn.

4.1 Bias Adjustment

The leading bias terms are $B_{S,1}$ and $B_{S,2}$. We here focus on the case where $\beta > \alpha_2$ (this includes parametric simulation-based estimation methods) so that $B_{S,1}$ is of lower order and the leading bias component is $B_{S,2} = -\frac{1}{2} H_0^{-1} E\left[\nabla^2 g(z_i; \theta_0)[\psi_S, \psi_S]\right]$. We then wish to adjust

the approximate estimator to remove this bias component. The two main approaches to bias adjustment in the literature are "corrective" and "preventive"; see Arellano and Hahn (2007) for a further discussion of these in a panel data setting. We discuss their uses in our setting in the following.

The *corrective method* first computes the unadjusted estimator, $\hat{\theta}_{n,S}$, obtains a consistent estimator of the bias, say $\hat{B}_{S,2}$, and then combines the two to obtain a new, bias-adjusted (BA) estimator

$$\tilde{\theta}_{n,S}^{\mathrm{BA}} = \hat{\theta}_{n,S} - \hat{B}_{S,2}.$$

One example of this approach for can be found in Lee (1995) for the special case of SMLE and SNLS in limited dependent variable models. A natural estimator of $\hat{B}_{S,2}$ would be $\hat{B}_{S,2} = -\frac{1}{2}\hat{H}_n^{-1}\nabla^2\hat{G}_n(\hat{\theta}_{n,S})$ for some consistent estimator $\nabla^2\hat{G}_n(\theta)$ of $E\left[\nabla^2 g(z_i;\theta)[\psi_S,\psi_S]\right]$. We propose two different estimators, depending on whether EIA's or ECA's are used in the implementation of the approximate estimator: First, when EIAs are used, we can use

$$\nabla^2\hat{G}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla^2 g(z_i;\theta,\hat{\gamma}_S)[\hat{\psi}_{S,i},\hat{\psi}_{S,i}], \quad \hat{\psi}_{S,i}(z;\theta) := \hat{\gamma}_{S,i}(z;\theta) - \frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_{S,i}(z;\theta) \quad (19)$$

If ECA's are employed, the above choice for $\hat{\psi}_S$ is not possible. However, assuming that A.6 also holds, the following alternative estimator is available, which can be employed for both ECA's and EIA's:

$$\nabla^2\hat{G}_n(\theta) = \frac{1}{nS(S-1)}\sum_{i=1}^{n}\sum_{s=1}^{S}\nabla^2 g(z_i;\theta)[w_{i,s}-\hat{\gamma}_{i,S}, w_{i,s}-\hat{\gamma}_{i,S}]. \quad (20)$$

In the case of ECA, $w_{i,s} = w_s$ and $\hat{\gamma}_{i,S} = \hat{\gamma}_S$ do not change across observations.

Instead of adjusting the estimator, we can do *preventive correction* where we adjust the estimating equation $G_n(\theta,\hat{\gamma}_S)$ to remove the component leading to the bias $B_{S,2}$. By inspection of the proof of Theorem 2, it is easily seen that the relevant adjustment of $G_n(\theta,\hat{\gamma}_S)$ is $E\left[\nabla^2 g(z_i;\theta)[\psi_S,\psi_S]\right]/2$ leading to the following bias-adjusted estimator $\hat{\theta}_{n,S}^{\mathrm{BA}}$ defined as the solution to

$$G_n(\hat{\theta}_{n,S}^{\mathrm{BA}},\hat{\gamma}_S) - \frac{1}{2}\nabla^2\hat{G}_n(\hat{\theta}_{n,S}^{\mathrm{BA}}) = o_P(1/\sqrt{n}), \quad (21)$$

where $\nabla^2\hat{G}_n(\theta)$ is chosen either as in eq. (19) or (20). This approach was pursued in the context of SNLLS (see Example 2) by Laffont et al (1995).

After either preventive or corrective adjustment, the bias component $B_{S,2}$ should change to

$$\tilde{B}_{S,2} := -\frac{1}{2}H_0^{-1}E\left(\nabla^2 G_n(\theta_0)[\psi_S,\psi_S] - \nabla^2\hat{G}(\theta_0)\right). \quad (22)$$

The following theorem analyzes the properties of the bias adjusted estimator based on

$\nabla^2 \hat{G}_n(\theta)$ given in eq. (20). We expect similar results to hold for any bias adjusted EIA estimator that uses eq. (19).

Theorem 4 *Assume that A.1-A.3, A.4*(3)*, and A.6*(8) *hold together with*

$$\left\|\nabla^2 g(z;\theta_0)[e_{is}, e_{it}]\right\| \leq b(z) \left\|e_{is}(z)\right\| \left\|e_{it}(z)\right\|,$$

where $E\left[b^8(z)\right] < \infty$*. Then any corrected estimator* $\hat{\theta}^{\text{BA}}_{n,S}$ *solving eq. (21) with* $\nabla^2 \hat{G}_n(\theta)$ *defined in (20) has reduced bias:*

$$E[\hat{\theta}^{\text{BA}}_{n,S} - \hat{\theta}_n] = B_{S,1} + \tilde{B}_{S,2}$$

where the new bias term given in eq. (22) satisfies $\tilde{B}_{S,2} = O(S^{-2+\mu_2})$ *and the rate of* $B_{S,1}$ *is* $O(S^{-\beta})$ *as in Theorem 2.*

The theorem shows that under slightly stronger conditions[5] than in Theorem 2, $\tilde{B}_{S,2}$ has a faster rate of convergence than $B_{S,2}$, while the rate of the other leading terms is unchanged. More precisely, when comparing with Theorem 2, the bias term $B_{S,2} = O\left(S^{-\alpha_2}\right) = O\left(S^{-1+\mu_2}\right)$ has been replaced by $\tilde{B}_{S,2} = O(S^{-2+\mu_2})$.

With unbiased simulators, we have $\mu_2 = 0$ and $\beta = \infty$, and by Theorem 2 the leading bias term of the unadjusted estimator is of order $O\left(S^{-1}\right)$. Theorem 4 shows that for the adjusted estimator the leading term of the bias is of order $O\left(S^{-2}\right)$. The improvement is by a factor S and may be quite large.

More generally, the proposed adjustment will remove the largest bias component as long as $\alpha_2 < \beta$. Otherwise the bias term $O_P\left(S^{-\beta}\right)$ is of a larger order than $O_P\left(S^{-\alpha_2}\right)$ and the proposed bias adjustment does not remove the leading term anymore. In particular, when non-stochastic approximations are employed the above adjustment does not help. If we could estimate b_S, then $B_{S,1}$ could be taken care of easily by adjusting either estimator or estimating equation using $\nabla \hat{G}_n(\theta) := \sum_{i=1}^n \nabla g_i(\theta, \hat{\gamma}_S)[\hat{b}_S]/n$. However, estimating b_S can be a difficult task.

As an illustration, we now return to the SNLS example introduced in Section 2 to derive the bias adjustment. Section 7 shows the formulæ for the mixed logit example we use in our Monte Carlo study.

Example 2: SNLS (continued). Recall that for nonlinear least squares,

$$G_n(\theta, \gamma) = \frac{1}{n}\sum_{i=1}^n g_i(\theta, \gamma), \quad g_i(\theta, \gamma) = -2(y_i - \gamma_i(\theta))\dot{\gamma}_i(\theta),$$

[5]The higher order on A.6 is required to ensure that in the asymptotic expansion, the remainder term, $R_{n,S}$ is still dominated.

where $\dot{\gamma}_i(\theta) = \partial\gamma_i(\theta)/(\partial\theta)$. Its first and second-order pathwise differentials are easily found to be

$$\nabla g_i[d\gamma] = 2\dot{\gamma}_i d\gamma - 2(y_i - \gamma_i)d\dot{\gamma}, \quad \nabla^2 g_i[d\gamma, d\gamma] = -4d\dot{\gamma}d\gamma.$$

Therefore, denoting $r(x_i, \varepsilon_{i,s}; \theta) := w_S(x_i, \varepsilon_{i,s}; \theta) - \hat{\gamma}_S(x_i; \theta)$, the adjustment term $\nabla^2 \hat{G}_n(\theta)$ takes the form

$$\nabla^2 \hat{G}_n(\theta) = -\frac{2}{S(S-1)} \sum_{i=1}^{n} \sum_{s=1}^{S} r(x_i, \varepsilon_{i,s}; \theta)\, \dot{r}(x_i, \varepsilon_{i,s}; \theta).$$

Note that instead of adjusting $G_n(\theta, \hat{\gamma}_S)$, we could have corrected the nonlinear sum of squares instead and minimized

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\gamma}_{i,S}(\theta)\right)^2 - \nabla^2 \hat{Q}_n(\theta),$$

with the obvious definition

$$\nabla^2 \hat{Q}_n(\theta) = \frac{1}{nS(S-1)} \sum_{i=1}^{n} \sum_{s=1}^{S} r^2(x_i, \varepsilon_s; \theta).$$

This is exactly the correction proposed in Laffont et al. (1995); and as $\nabla^3 g_i \equiv 0$ in SNLS, all approximation biases are removed.

4.2 Adjusting Standard Errors

If the approximator is stochastic, the approximate estimator will not only be biased; it will also contain additional variance terms. For a given sample size n and number of simulations S, we should adjust inferential tools (such as standard errors and t-statistics) to account for these additional variances. This turns out to be quite straightforward in many cases. We focus here on parametric simulation-based inference; and we assume that A.6(4) holds with unbiased simulators that do not directly depend on S, so that $\beta = \infty$ and $\alpha_2 = 1$.

First note that, as part of Theorem 2, the overall variance of the approximate estimator is

$$\text{EIA} : \text{Var}(\hat{\theta}_{n,S}) \approx \frac{1}{n} H_0^{-1} \left\{\Omega_S^{G+D} + \Omega_S^E\right\} H_0^{-1}, \quad \text{ECA} : \text{Var}(\hat{\theta}_{n,S}) \approx H_0^{-1} \left\{\frac{1}{n}\Omega_S^{G+D} + \frac{1}{S}\Omega^E\right\} H_0^{-1},$$

where Ω_S^{G+D}, Ω_S^E and Ω^E are defined in the theorem. Implicitly, these depend on θ_0 and γ_0. In standard estimation procedures, one would usually estimate the above variance components by simply replacing θ_0 and γ_0 by $\hat{\theta}_{n,S}$ and $\hat{\gamma}_S$, respectively, in the expressions of the Ω matrices, and by replacing any population means by their sample counterparts. However, as

explained in the previous section, replacing γ_0 by $\hat{\gamma}_S$ will generate biases. Similarly, if $\hat{\theta}_{n,S}$ has not been bias adjusted, replacing θ_0 by $\hat{\theta}_{n,S}$ will add biases to the variance estimator.

We therefore propose to use a bias-adjusted variance estimator to improve on the basic variance estimators. We assume in the following that $\hat{\theta}_{n,S}$ has already been bias adjusted so that we only need to adjust any biases due to $\hat{\gamma}_S$. Furthermore, for notational simplicity we here assume that the data is i.i.d.[6]. First, consider $\Omega^G = E\left[g\left(z;\theta_0,\gamma_0\right)g\left(z;\theta_0,\gamma_0\right)'\right]$. A naive estimator would be $\hat{\Omega}^G = \frac{1}{n}\sum_{i=1}^n \hat{g}_i\hat{g}_i'$, where $\hat{g}_i = g(z_i;\hat{\theta}_{n,S},\hat{\gamma}_S)$. However, this variance estimator will suffer from a bias of the same order as Ω_S^E, and so will lead to imprecise standard errors. We therefore propose to bias adjust this estimator in the same way that we bias-adjusted $G_n\left(\theta,\hat{\gamma}_S\right)$:

$$\hat{\Omega}^G = \frac{1}{n}\sum_{i=1}^n \hat{g}_i\hat{g}_i' - \hat{\Delta}_{n,S}^{\Omega}$$

where, with $\hat{\psi}_{i,S} := \hat{\gamma}_{i,S} - \bar{\gamma}_S$,

$$\hat{\Delta}_{n,S}^{\Omega} = \frac{1}{n}\sum_{i=1}^n \left\{\nabla^2\hat{g}_i[\hat{\psi}_{i,S},\hat{\psi}_{i,S}]\hat{g}_i' + 2\nabla\hat{g}_i[\hat{\psi}_{i,S}]\nabla\hat{g}_i[\hat{\psi}_{i,S}]' + \hat{g}_i\nabla^2\hat{g}_i[\hat{\psi}_{i,S},\hat{\psi}_{i,S}]'\right\},$$

or, when Assumption A.6 is satisfied,

$$\hat{\Delta}_{n,S}^{\Omega} = \frac{1}{nS\left(S-1\right)}\sum_{i=1}^n\sum_{s=1}^S \left\{\nabla^2\hat{g}_i[\hat{e}_{s,i},\hat{e}_{s,i}]\hat{g}_i' + 2\nabla\hat{g}_i[\hat{e}_{s,i}]\nabla\hat{g}_i[\hat{e}_{s,i}]' + \hat{g}_i\nabla^2\hat{g}_i[\hat{e}_{s,i},\hat{e}_{s,i}]'\right\},$$

where $\hat{e}_{s,i} = w_{s,i} - \hat{\gamma}_{i,S}$. The analysis of this estimator proceeds as in the proof of Theorem 4.

Next, consider Ω_S^E. Since this term is already of order $1/S$, we do not need to bias adjust it; we can simply use the naive estimator given by

$$\hat{\Omega}_S^E = \frac{1}{nS^2}\sum_{i=1}^n\sum_{s=1}^S \nabla\hat{g}_i[\hat{\psi}_{i,s}]\nabla\hat{g}_i[\hat{\psi}_{i,s}]'.$$

Finally, the naive estimator of H_0 takes the form

$$\hat{H} = \frac{1}{n}\sum_{i=1}^n \frac{\partial\hat{g}_i}{\partial\theta}.$$

One could bias-adjust this estimator as we did for $\hat{\Omega}^G$. However, note that the approximate

[6]Otherwise long-run variance estimators have to be used.

estimator satisfies:

$$0 = \frac{1}{n}\sum_{i=1}^{n} g(z_i;\theta_0,\hat{\gamma}_{i,S}) + \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(z_i;\bar{\theta}_{n,S},\hat{\gamma}_{i,S})}{\partial\theta}\right\}(\hat{\theta}_{n,S}-\theta_0).$$

So in order to get a precise approximation of the distribution of $\hat{\theta}_{n,S}-\theta_0$, we want to use an estimator that mimics the behaviour of $\frac{1}{n}\sum_{i=1}^{n}\partial g(z_i;\bar{\theta}_{n,S},\hat{\gamma}_{i,S})/(\partial\theta)$. This is exactly what $\hat{H}$ does; and we can still use it as an estimator of H_0.

Variance estimation for ECA estimators proceeds in a similar manner. The only difference is the term Ω^E. Under the additional assumption A.6, $\Omega^E = \lim_{S\to\infty}\sum_{s=-S}^{S}\mathrm{Cov}\left(\tilde{e}_0,\tilde{e}_s\right)/S$ where $\tilde{e}_s = \nabla G\left(\theta_0,\gamma_0\right)[\bar{w}_s]$. First, observe that $\nabla\hat{G}[\gamma] = \frac{1}{n}\sum_{i=1}^{n}\nabla g(z_i,\hat{\theta}_{n,S},\hat{\gamma}_S)[\gamma]$ is a consistent estimator of $\nabla G\left(\theta_0,\gamma_0\right)[\gamma]$ for any given γ. If the simulations are independent, a natural estimator of Ω^E is then

$$\text{ECA}: \hat{\Omega}_S^E = \frac{1}{S^2}\sum_{s=1}^{S}\nabla\hat{G}[w_s-\hat{\gamma}_S]\nabla\hat{G}[w_s-\hat{\gamma}_S]'.$$

This estimator is similar to the one proposed in Newey (1994) for semiparametric two-step estimators. If the simulations are dependent, as is often the case in time series models, a HAC estimator has to be employed.

5 Bias Adjustment by Resampling

As an alternative to analytical bias corrections, resampling methods could be used[7]. They will in general handle the biases due to both the stochastic and the non-stochastic component of the approximator; and the researcher is not required to derive an expression of the bias. On the other hand, they are computationally more demanding than the analytical bias correction proposed in the previous section, and may lead to an increase in finite-sample variance.

To motivate the bias adjustment, recall from Theorem 2 that $E[\hat{\theta}_{n,S}-\hat{\theta}_n] \simeq b_1 S^{-\beta} + b_2 S^{-\alpha_2}$. As before, the goal is to obtain an estimator of (parts of) the leading bias terms and use this for bias correction. We here propose to do this by resampling methods: First, compute two approximators of order S^* which we denote $\hat{\gamma}_{S^*}^{[1]}$ and $\hat{\gamma}_{S^*}^{[2]}$. Let $\hat{\theta}_{n,S^*}^{[m]}$ be the estimator based on the same data sample $\mathcal{Z}_n$ but using the mth approximator $\hat{\gamma}_{S^*}^{[m]}$, $m=1,2$.

We then propose the following jackknife (JK) type estimator:

$$\hat{\theta}_{n,S}^{\text{JK}} := 2\hat{\theta}_{n,S} - \frac{1}{2}\left(\hat{\theta}_{n,S^*}^{[1]} + \hat{\theta}_{n,S^*}^{[2]}\right), \tag{23}$$

[7]See Hahn and Newey (2004) and Dhaene and Jochmans (2012) for bias correction using Jackknife in the context of panel models, while we refer to Phillips and Yu (2005) for a time series application.

and we easily see that

$$\begin{aligned} E\left[\hat{\theta}_{n,S}^{\text{JK}} - \hat{\theta}_n\right] &= 2E\left[\hat{\theta}_{n,S} - \hat{\theta}_n\right] - \frac{1}{2}\left(E\left[\hat{\theta}_{n,S^*}^{[1]} - \hat{\theta}_n\right] + E\left[\hat{\theta}_{n,S^*}^{[2]} - \hat{\theta}_n\right]\right) \\ &\simeq b_1\left[2S^{-\beta} - (S^*)^{-\beta}\right] + b_2\left[2S^{-\alpha} - (S^*)^{-\alpha_2}\right], \end{aligned}$$

where higher-order terms have been ignored. We would now ideally choose S^* such that both of the above bias terms cancel out. However, we can only remove either of the two: By choosing either

$$S^* = \frac{S}{2^{1/\beta}} \text{ or } S^* = \frac{S}{2^{1/\alpha_2}}, \tag{24}$$

we will remove the first or the second term respectively. Obviously, S^* should be chosen so as to remove the bias component that dominates in the expansion.

One can generalize the above and compute M approximators, $\hat{\gamma}_{S_m}^{[m]}$, $m = 1, ..., M$, of order $S_m < S$, and for each of those the corresponding approximate estimator, $\hat{\theta}_{n,S_m}^{[m]}$. For a given set of weights p_m, $m = 1, ..., M$, we then define the adjusted estimator as

$$\hat{\theta}_{n,S}^{\text{JK}} = M\hat{\theta}_{n,S} - \sum_{m=1}^{M} p_m \hat{\theta}_{n,S_m}^{[m]}. \tag{25}$$

Dhaene and Jochmans (2012, Corollary 1) demonstrate in a panel data context that the optimal procedure to remove the leading bias term is to choose $M = 2$ and $p_m = 1/2$. We expect that a similar result extends to parametric simulation-based estimators in our setting. On the other hand, the generalized adjustment as given in (25) can be used to remove further higher-order bias components by appropriate choice of weights and appproximation orders, c.f. Dhaene and Jochmans's section 3. While we do not pursue this here, we conjecture that the generalized adjustment would enable us to remove both B_1 and B_2.

One way to reduce the computational cost is to jackknife the objective function directly, thereby avoiding having to compute two separate estimators, $\hat{\theta}_{n,S^*}^{[1]}$ and $\hat{\theta}_{n,S^*}^{[2]}$; the jackknifed objective function is constructed along the same lines as in the discussion of analytical bias correction.

6 Newton-Raphson Adjustment

The previous sections developed two bias adjustment methods. We here propose a simple method that can reduce both bias and variance of the approximate estimator in a simple manner. The proposed method works with non-stochastic approximations as well as with stochastic approximations by extending the well-known idea that a consistent estimator can be made asymptotically efficient by applying one Newton-Raphson (NR) step of the log-likelihood function to it. E.g. if $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ_0 in a model with

log-likelihood $L_n(\theta)$, then a single NR-step yields a consistent and asymptotically efficient estimator. We apply this idea to our setting by starting from some initial approximate estimator based on a small degree of approximation S, say $\bar{\theta}_{n,S}$. We then define the corrected estimator through one or possibly several Newton-Raphson iterations of an approximate objective function that uses a much finer approximation, $S^* \gg S$. With $H_n(\theta,\gamma) = \partial G_n(\theta,\gamma)/(\partial\theta)$, define

$$\hat{\theta}_{n,S}^{(k+1)} = \hat{\theta}_{n,S}^{(k)} - H_n^{-1}(\hat{\theta}_{n,S}^{(k)}, \hat{\gamma}_{S^*}) G_n(\hat{\theta}_{n,S}^{(k)}, \hat{\gamma}_{S^*}), \quad k = 1,2,3,... \tag{26}$$

where $\hat{\theta}_{n,S}^{(1)} = \bar{\theta}_{n,S}$ is some initial estimator and we use the S^*th order approximator, $\hat{\gamma}_{S^*}$, in the iterations.

Note that the cost of computing this new estimator from the first one is (very) roughly S^*/S times the cost of one iteration in the minimization of $Q_n(\theta, \hat{\gamma}_{S^*})$. Since the minimization easily can require a hundred iterations or so, we can therefore take S^* ten or twenty times larger than S without adding much to the cost of the estimation procedure.[8] Also, one iteration is enough if S^* goes to infinity at least as fast as S. Our final proposal will also work for general approximation-based estimators. We show that starting from either $\bar{\theta}_{n,S} = \hat{\theta}_{n,S}^{\mathrm{AB}}$, $\hat{\theta}_{n,S}^{\mathrm{JK}}$ or even the initial, unadjusted estimator, $\hat{\theta}_{n,S}$, one or more Newton-Raphson iterations based on the approximate objective function with a finer approximation $S^* > S$ produce an estimator that has the presumably higher precision of $\hat{\theta}_{n,S^*}$. The resulting estimator based on k iterations, $\hat{\theta}_{n,S}^{(k+1)}$, is defined in (26).

To evaluate the performance of $\hat{\theta}_{n,S}^{(k+1)}$ relative to $\bar{\theta}_{n,S^*}$, we first note that

$$\|\hat{\theta}_{n,S}^{(k+1)} - \hat{\theta}_n\| \leq \|\hat{\theta}_{n,S}^{(k+1)} - \bar{\theta}_{n,S^*}\| + \|\bar{\theta}_{n,S^*} - \hat{\theta}_n\|.$$

Combining this with Robinson (1988, Theorem 2), we obtain the following theorem:

Theorem 5 *Assume that A.1-A.3, A.4*(3) *and A.5*(6) *hold. Let the initial estimate $\bar{\theta}_{n,S}$ be chosen as either $\hat{\theta}_{n,S}$, $\hat{\theta}_{n,S}^{\mathrm{AB}}$, or $\hat{\theta}_{n,S}^{\mathrm{JK}}$. Then the NR-estimator $\hat{\theta}_{n,S}^{(k+1)}$ defined in (26) satisfies:*

$$\|\hat{\theta}_{n,S}^{(k+1)} - \hat{\theta}_n\| = O_P\left(\|\bar{\theta}_{n,S} - \hat{\theta}_n\|^{2^k}\right) + O_P\left(\|\bar{\theta}_{n,S^*} - \hat{\theta}_n\|\right) \tag{27}$$

as n, S and S^* go to infinity with $S^* > S$.

The above result formalizes the intuition that a (large enough) number of NR-steps with the score and Hessian evaluated at γ_{S^*} yields an estimator that is equivalent to the extremum estimator obtained from full optimization of the objective function based on γ_{S^*}. This holds irrespective of the convergence rate of the initial estimator. However, the number of NR

[8] In many cases, a large part of the dimensionality of θ only comes into play within some linear indexes $\theta' x$; then the trade off is even more favourable since the computation of the second derivative H_n is much simplified.

iterations, k, needed to obtain this result does depend on the precision of the initial estimator. For unadjusted parametric simulation-based estimators in the EIA scheme for instance, we know from Theorem 2 that $\|\bar{\theta}_{n,S} - \hat{\theta}_n\| = O_P(1/S)$. Then the first term on the right-hand side of the inequality in Theorem 5 is asymptotically dominated by the second term if $S^* = o(S^{2^k})$. Taking $k = 1$ and having S^*/S converge to some positive number would be enough in this case.

The above iterative estimator requires computation of the Hessian, $H_n(\theta, \hat{\gamma}_S)$. If this is not feasible or computationally burdensome, an approximation can be employed, e.g. numerical derivatives. This however will slow down the convergence rate and the result of Theorem 5 has to be adjusted, cf. Robinson (1988, Theorem 5). In particular, more iterations are required to obtain a given level of precision.

7 Simulation Study

To explore the performance of our proposed approaches, we set up a small Monte Carlo study of a mixed logit model: the econometrician observes i.i.d. draws of $z_i = (x_i, y_i)$ for $i = 1, \ldots, n$, with x_i a centered normal of variance τ^2 and

$$y_i = \mathbb{1}(b + (a + su_i)x_i + e_i > 0)$$

where e_i is standardized type I extreme value and u_i is a centered normal with unit variance, independent of e_i.

We take the true model to have parameters $a = 1, s = 1, b = 0$. In this specification, the mean probability of $y = 1$ is one-half. For $\tau = 1$ (resp. $\tau = 2$) the generalized R^2 is 0.11 (resp. 0.21); in the corresponding simple logit model, which has $s = 0$, the R^2 would be 0.17 (resp. 0.39.)

The mixed logit, in its multinomial form, has become a workhorse in studies of consumer demand (see e.g. the book by Train (2009)); it also figures prominently on the demand side of models of empirical industrial organization. It is usually estimated by simulation-based methods. In empirical IO, the simulated method of moments is commonly used because of endogeneity concerns; since they are absent here, we focus on SML instead.

This is still a very simple model; thus we can use Gaussian quadrature to compute the integral

$$\Pr(y = 1|x) = \int \frac{\phi(u)}{1 + \exp(-(b + (a + su)x))} du. \tag{28}$$

Since Gaussian quadrature achieves almost correct numerical integration in such a regular, one-dimensional case, we can rely on it to do (almost) exact maximum likelihood estimation. By the same token, it is easy to compute the asymptotic variance of the exact ML estimator

$\hat{\theta}_n$, and the leading term $B_{S,2}$ of the bias of the SML estimator. Simple calculations[9] give the numbers in Table 1.

τ	$\sqrt{n}\hat{\sigma}$			S times bias		
	a	s	b	a	s	b
1	7.2	17.1	2.4	−9.0	−23.2	−0.0
2	6.7	10.8	2.8	−8.2	−13.3	−0.0

Table 1: Rescaled asymptotic standard errors and simulation biases

The columns labeled $\sqrt{n}\hat{\sigma}$ give the square roots of the diagonal terms of the inverse of the Fisher information matrix. As can be seen from the values of $\sqrt{n}\hat{\sigma}$, it takes a large number of observations to estimate this model reliably. To take an example, assume that the econometrician would be happy with a modestly precise 95% confidence interval of half-diameter 0.2 for the mean slope a. With $\tau = 1$ it would take about $(7.2 * 1.96/0.2)^2 \simeq 5,200$ observations; and still about $4,500$ for $\tau = 2$, even though the generalized R^2 almost doubles. With such sample sizes, the estimate of the size of the heterogeneity s would still be very noisy: its 95% confidence intervals would have half-diameters 0.48 and 0.32, respectively for $\tau = 1$ and $\tau = 2$. We also found that the correlation between the estimators of a and of s is always large and positive—of the order of 0.8. Thus the confidence region for the pair (a, s) is in fact a rather elongated ellipsoid. On the other hand, the estimates of b are reasonably precise, which is not very surprising as b shifts the mean probability of $y = 1$ strongly.

The figures in the columns labeled " S times bias" refer to the expansions of $\hat{\theta}_{nS} - \hat{\theta}_n$ in our theorems. We will be using SML under the EIA scheme (independent draws across observations). Then we know that the leading term of the bias due to the simulations is $B_{S,2}$ and is of order $1/S$. The figures give our numerical evaluation of $SB_{S,2}$, using our formulæ and Gaussian quadrature again. As appears clearly from Table 1, once again the heterogeneity coefficient s is the harder to estimate, followed by a, while there is hardly any bias on b. With $S = 100$ simulations and $\tau = 1$ for instance, the bias on a is -0.09, and the bias on s is -0.23.

We ran experiments for several sets of parameter values, sample sizes n, explanatory power (through τ), and numbers of draws S. Since the results are similar, we only present here those we obtained for a sample of $10,000$ observations when the true model has $a = 1, s = 1, b = 0$, and the covariate has standard error $\tau = 1$ or $\tau = 2$.

We present below the results for $S = 50$, 100, and 200 simulations. We ran 5,000 simulations in each case, starting from initial values of the parameters drawn randomly from uniform distributions: $a \sim U[0.5, 1.5]$, $b \sim U[-0.5, 0.5]$, and $s \sim U[0.5, 1.5]$. For each simulated sample, we estimated the model using (i) uncorrected SML, (ii) SML with Newton-Raphson

[9] We used adaptive Gaussian quadrature.

(NR), and (iii) SML with analytic adjustment (AA) for both bias and variance[10]. The AA was done on the objective function. For the NR correction, we use only $k = 1$ step, with $S^* = 10 \times S$ draws.

For each method, we also used several ways of computing the standard errors of the estimates: from the most popular one, which consists of inverting the outer product if the scores without correcting for the simulations, to the better-grounded sandwidch formula which we introduced in Section 6.

We faced very few numerical difficulties. The optimization algorithm sometimes stopped very close to the bounds we had imposed for the heterogeneity parameter, $0.1 \leq s \leq 5$. In some cases it failed to find an optimum, especially for uncorrected SML with 50 draws. Finally, the second derivative of the simulated log-likelihood was sometimes not invertible in one of our sandwich formulæ. Altogether, we had to discard 1% to 1.5% of the 5,000 samples, depending on the run. The tables and graphs below only refer to the remaining samples. We focus on a and s since there is little to correct for in the SML estimates of b. We report (Huber) robust means, standard errors and RMSEs. "AA" refers to our analytical bias adjustment.

Tables 2 and 3 report our results for the mean error of our various SML methods. Each row corresponds to a value of the number of simulations S. All numbers in the last three columns of these tables were computed by averaging the " error terms" $(\hat{\theta}_{n,S} - \theta_0)$ over the 5,000 samples (minus the small number that were eliminated due to numerical issues). The standard error of these averages is about 0.001, so that several of the biases from the corrected estimates are close to insignificant.

τ	S	SML	SML+Newton	SML+AA
1	50	−0.133	−0.089	0.004
	100	−0.078	−0.039	0.000
	200	−0.041	−0.014	0.000
2	50	−0.133	−0.051	0.010
	100	−0.069	−0.016	0.007
	200	−0.033	0.003	0.006

Table 2: Mean error on a

The "SML" columns in the tables report the biases of the uncorrected SML estimator. The leading term appears to be a good approximation to the actual size of the bias in these simulations, and the measured bias is close to proportional to $1/S$. This suggests that our analytical bias adjustment, which focuses on correcting for the leading term of the bias, should work very well. As the last columns show, AA in fact does eliminate most of the bias.

[10] In a previous version we also reported results for the resampling method of section 5. Since they were dominated by the other methods, we drop them from the tables here.

τ	S	SML	SML+Newton	SML+AA
1	50	−0.364	−0.217	0.011
	100	−0.206	−0.093	0.000
	200	−0.109	−0.033	0.001
2	50	−0.214	−0.064	0.021
	100	−0.110	−0.018	0.015
	200	−0.051	0.010	0.013

Table 3: Mean error on s

The Newton step with ten times more simulations reduces the bias, as expected; but it does not do it as effectively as our analytical bias adjustment.

The discussion above only bears on bias, but one may legitimately be concerned about the possibility that our adjustment procedures introduce more noise into the estimates and perhaps even increase their mean square errors. Tables 4 and 5 show that this concern is unfounded. Correcting the estimates using analytical adjustment or a Newton step reduces the RMSE in all cases. Most often, the reduction in bias dominates and AA works better than Newton. However, for larger number of simulations when $\tau = 2$, bias reduction matters less; and since the Newton method is more effective at reducing dispersion, its RMSE becomes smaller than that of the AA method. This suggests that combining AA and a Newton step could yield an even larger reduction in the RMSE.

τ	S	SML	SML+Newton	SML+AA
1	50	0.139	0.095	0.041
	100	0.083	0.043	0.028
	200	0.046	0.019	0.020
2	50	0.136	0.053	0.032
	100	0.072	0.020	0.022
	200	0.036	0.010	0.016

Table 4: RMSE on a

τ	S	SML	SML+Newton	SML+AA
1	50	0.378	0.234	0.107
	100	0.219	0.103	0.074
	200	0.121	0.044	0.053
2	50	0.219	0.068	0.056
	100	0.115	0.025	0.039
	200	0.056	0.019	0.029

Table 5: RMSE on s

Both the bias and the increased variance imparted by the simulations affect the properties of standard tests. Figures 1 and 2 document this for t-tests that a and s, respectively, equal their true values. For such a large sample, we would expect the distributions of the t-statistics to be very close to a standard centered normal—the dashed curve in each panel; and 95% of the mass should lie between the two dashed vertical lines. What we observe for the uncorrected SML estimator ("SML") is quite different: the bias in the estimate skews the distribution to the left, spectacularly so for small number of simulations; and the increased variance flattens the distribution.

Resorting to one Newton-Raphson step (the "SML+Newton" curves) corrects part of the bias and reduces the variance; but except for large number of simulations, the distribution of the resulting t-statistics is still markedly different from $N(0,1)$. Using the AA bias-correction and using the proper formula for the variance-covariance matrix (the "SML+AA" curves), on the other hand, produces distributions that are essentially undistinguishable from $N(0,1)$.

Tables 6 and 7 give the actual coverage probabilities implied by figures 1 and 2. When using uncorrected SML, the nominally 95% confidence intervals undercover very badly, so that the null hypothesis is rejected up to three-quarters of the time when it is in fact true. Our corrections, on the other hand, yield tests that have close to exact coverage.

τ	S	SML	SML+Newton	SML corrected
1	50	0.295	0.631	0.961
	100	0.687	0.867	0.952
	200	0.860	0.926	0.949
2	50	0.253	0.842	0.949
	100	0.697	0.935	0.954
	200	0.872	0.954	0.956

Table 6: Actual coverage probabilities for a

τ	S	SML	SML+Newton	SML corrected
1	50	0.232	0.637	0.967
	100	0.660	0.879	0.961
	200	0.866	0.933	0.956
2	50	0.236	0.879	0.951
	100	0.690	0.940	0.953
	200	0.873	0.951	0.952

Table 7: Actual coverage probabilities for s

Two other considerations are worth mentioning:

- *Ease of implementation:* the analytical bias adjustment wins on that count, since it

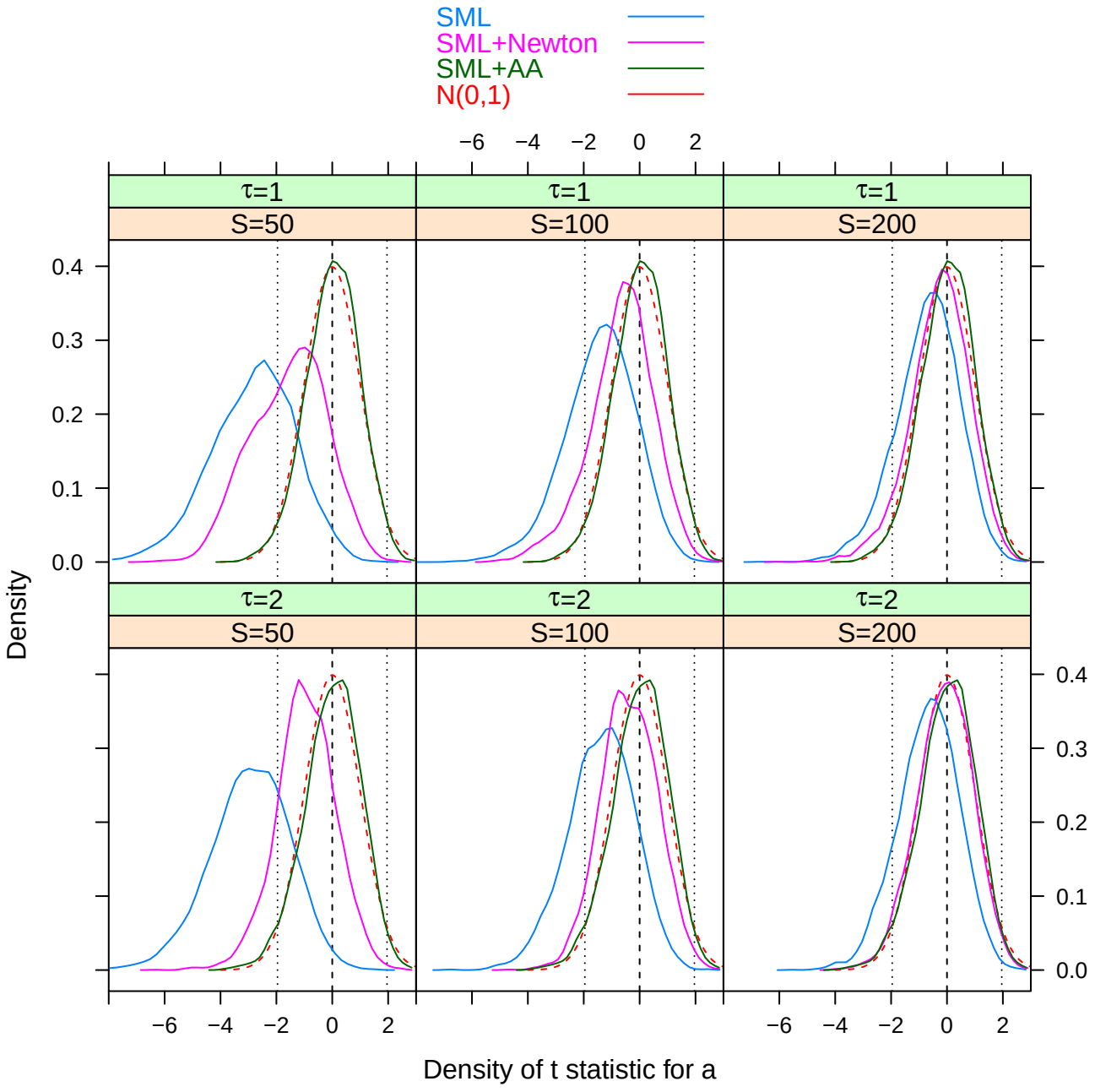

Figure 1: Distributions of the t statistics for $a = 1$

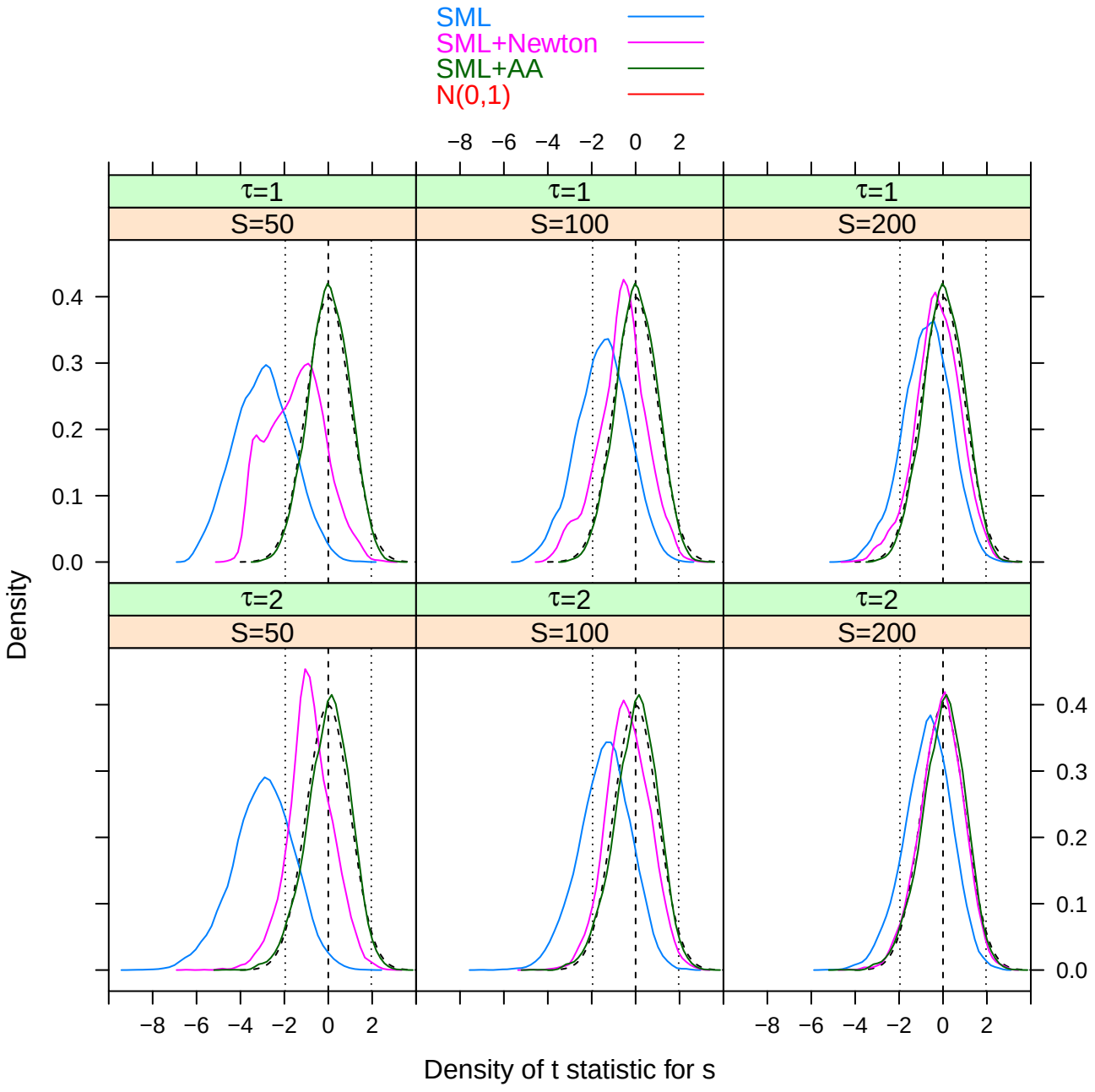

Figure 2: Distributions of the t statistics for $s = 1$

is usually easy to get a formula for the Δ term and to program it. The Newton method may be more troublesome in models with more than a few parameters, as it requires a reasonably accurate evaluation of the matrix of second derivatives. In our experiment, we relied on the fact that the minimization algorithm itself proceeds by Newton-Raphson steps; after multiplying by ten the number of simulations, we let the algorithm do exactly one iteration of its line search. This appears to work very well, and is very easy to implement.

- *Computer time:* The analytical bias adjustment wins this comparison hands down. For SML for instance, the evaluation of the corrected objective function requires computing the variance of the simulated choice probabilities in addition to their mean, as well as their derivatives—a very small computational cost. Newton adjustment was about five times more costly in our example; it may be more or less time-consuming in other applications, depending on the structure of the model and the care needed to approximate the Hessian.

Like any Monte Carlo study, ours can only be illustrative; yet our results are very encouraging. Our analytical corrections for both bias and variance spectacularly improve inference. Using one Newton step, while less effective, can also be a good way to reduce errors.

8 Conclusion

We developed in this paper a unifying framework for the analysis of approximate estimators. We derived a higher-order expansion and we used it to propose methods for reducing the bias and the efficiency loss that result from the approximation. Simulations on the mixed logit model confirm that the proposed methods work well in finite samples.

We restricted ourselves to estimators where objective function and approximator (as functions of θ) were both smooth. In principle, one could import the arguments of Chen et al (2003) to handle non-smooth cases as is done in Armstrong et al (2013). Another approach would be to employ a slight generalization of Robinson (1988, Theorem 1) which in our setting would yield

$$||\hat{\theta}_{n,S} - \tilde{\theta}_n|| = O_P\left(\sup_{\|\theta-\theta_0\|\leq\delta} \|G_n(\theta, \hat{\gamma}_S) - G_n(\theta, \gamma)\|\right) + o_P\left(1/\sqrt{n}\right),$$

for some $\delta > 0$. If one could then strengthen the pointwise bias and variance results derived here to hold uniformly over $\|\theta - \theta_0\| \leq \delta$, all our results would remain valid.

Also, we require the approximators to be mutually independent, which rules out certain recursive approximation schemes such as particle filtering. Establishing results for this more

complicated case would be highly useful. One could here try to use the results of Chen and White (2002) who analyze random dynamic function systems.

We only allowed for one source of approximation in γ. More general situations could have several such terms, possibly with quite different properties. As an example, we could have evaluated a quantity γ_1 using simulations, and another term γ_2 by discretizing over a grid and interpolating. We could still write a Taylor expansion, and evaluate the corresponding terms. While we have not formally explored this extension, we feel that we can venture some conjectures. The Newton method would still work, using here both a larger number of simulations and a more precise grid in computing the Newton correction. The analytical bias-adjustment method would only work if all sources of approximations were "stochastic" (unlike γ_2 in our example); and then one would focus on the approximation whose size goes to zero most slowly.

Finally, one could interpret an approximate estimator as the exact estimator of a misspecified model. Suppose for instance that we use maximum likelihood to estimate a model with pdf $f(z, \theta)$; and that we suspect that the data may have been generated by a model whose pdf $f^*(z, \theta_0)$ is close to the set of pdfs $(f(\cdot, \theta))$. We can transport all of our results to this problem, with f as γ_S and f^* as γ. In practice we do not know f^* of course; but our methods can be used to explore the likely consequences of any type of (local) misspecification of concern.

References

Ackerberg, D., J. Geweke and J. Hahn (2009) Comments on "Convergence Properties of the Likelihood of Computed Dynamic Models". *Econometrica* 77, 2009–2017.

Altissimo, F. and A. Mele (2009) Simulated Nonparametric Estimation of Dynamic Models. *Review of Economic Studies* 76, 413–450.

Arellano, M. and J. Hahn (2007) Understanding Bias in Nonlinear Panel Models: Some Recent Developments. In *Advances in Economics and Econometrics*, Volume III (eds. R. Blundell, W.K. Newey and T. Persson). Cambridge: Cambridge University Press.

Andrews, D.W.K. (1994) Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity. *Econometrica* 62, 43–72.

Andrews, D.W.K. (2002) Higher-order Improvements of a Computationally Attractive k-step Bootstrap for Extremum Estimators. *Econometrica* 70, 119–162.

Armstrong, T, Gallant, A. R., Hong, H. and L. Huiyu (2012) The Asymptotic Distribution of Estimators with Overlapping Simulation Draws, mimeo Stanford.

Berry, S., Levinsohn, J., and Pakes, A. (1995) Automobile Prices in Market Equilibrium. *Econometrica* 63, 841–890.

Brownlees, C.T., D. Kristensen and Y. Shin (2011) Smooth Filtering and Likelihood Inference in Dynamic Latent Variables Models. Manuscript, Department of Economics, UCL.

Chen, X., O. Linton and I. Van Keilegom (2003) Estimation of Semiparametric Models When the Criterion Function Is Not Smooth. *Econometrica* 71, 1591–1608.

Chen, X. and H. White (2002) Asymptotic Properties of Some Projection-Based Robbins-Monro Procedures in a Hilbert Space. *Studies in Nonlinear Dynamics & Econometrics* 6(1), Article 1.

Creel, M. and D. Kristensen (2012) Estimation of Dynamic Latent Variable Models Using Simulated Nonparametric Moments. *Econometrics Journal* 15, 490-515.

Corradi, V. and N.R. Swanson (2007) Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historical Data. *Journal of Econometrics* 136, 699–723.

Dhaene, G. and K. Jochmans (2012) Split-panel Jackknife Estimation of Fixed-effect Models. Manuscript, Katholieke Universiteit Leuven.

van Dijk, H., A. Monfort and B. Brown (1995) *Econometric Inference Using Simulation Techniques*. John Wiley.

Dubé, J.P., J. Fox, and C-L. Su (2012) Improving the Numerical Performance of Static and Dynamic Aggregate Discrete Choice Random Coefficients Demand Estimation. *Econometrica* 80, 2231–2267.

Duffie, D. and K. J. Singleton (1993) Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica* 61, 929–952.

Fermanian, J.-D. and B. Salanié (2004) A Nonparametric Simulated Maximum Likelihood Estimation Method. *Econometric Theory* 20, 701–734.

Fernández-Villaverde, J. and J.F. Rubio-Ramirez (2005) Estimating Dynamic Equilibrium Economies: Linear versus Nonlinear Likelihood. *Journal of Applied Econometrics* 20, 891–910.

Fernández-Villaverde, J., J.F. Rubio-Ramirez and M. Santos (2006) Convergence Properties of the Likelihood of Computed Dynamic Models. *Econometrica* 74, 93–119.

Gouriéroux, C. and A. Monfort (1996) *Simulation-Based Econometric Methods*. Oxford: Oxford University Press.

Hahn, J. and W.K. Newey (2004) Jackknife and Analytical Bias Reduction for Nonlinear Panel Models. *Econometrica* 72, 1295–1319.

Hajivassiliou, V.A. (2000) Some Practical Issues in Maximum Simulated Likelihood. In *Simulation-based Inference in Econometrics* (eds. R. Mariano, T. Schuermann and M.J. Weeks), 71–99. Cambridge: Cambridge University Press.

Hong, H., A. Mahajan and D. Nekipelov (2010) Extremum Estimation and Numerical Derivatives. Manuscript, Stanford University.

Judd, K., F. F. Kubler and K. Schmedder (2003) Computational Methods for Dynamic Equilibria with Heterogeneous Agents. In *Advances in Economics and Econometrics* (eds. M. Dewatripont, L.P. Hansen, and S. Turnovsky). Cambridge University Press.

Judd, K. and C. Su (2012) Constrained Optimization Approaches to Estimation of Structural models. *Econometrica* 80, 2213–2230.

Kristensen, D. and B. Salanié (2010) Higher Order Improvements for Approximate Estimators. CAM Working Paper 2010-04, University of Copenhagen.

Kristensen, D. and B. Schjerning (2011) Implementation and Estimation of Discrete Markov Decision Models by Sieve Approximations. Manuscript, University of Copenhagen.

Kristensen, D. and Y. Shin (2012) Estimation of Dynamic Models with Nonparametric Simulated Maximum Likelihood. *Journal of Econometrics* 167, 76–94.

Kristensen, D. and A. Rahbek (2005) Asymptotics of the QMLE for a Class of ARCH(q) Models. *Econometric Theory* 21, 946–961.

Laffont, J.-J., H. Ossard and Q. Vuong (1995) Econometrics of First-Price Auctions. *Econometrica* 63, 953–980.

Laroque, G. and B. Salanié (1989) Estimation of Multimarket Fix-Price Models: An Application of Pseudo-maximum Likelihood Methods. *Econometrica* 57, 831–860.

Laroque, G. and B. Salanié (1993) Simulation-based Estimation of Models with Lagged Latent Variables. *Journal of Applied Econometrics* 8, 119–133.

Lee, L.-F. (1992) On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models. *Econometric Theory* 8, 518–552.

Lee, L.-F. (1995) Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models. *Econometric Theory* 11, 437–483.

Lee, L.-F. (1999) Statistical Inference with Simulated Likelihood Functions. *Econometric Theory* 15, 337–360.

Lee, L.-F. (2001) Interpolation, Quadrature, and Stochastic Integration. *Econometric Theory* 17, 933–961.

Mariano, R., T. Schuerman and M. Weeks (2000) *Simulation-based Inference in Econometrics.* Cambridge University Press.

McFadden, D.F. (1989) A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration. *Econometrica* 57, 995–1026.

Newey, W.K. (1991) Uniform Convergence in Probability and Stochastic Equicontinuity, *Econometrica* 59, 1161–1167.

Newey, W.K. (1991) Kernel Estimation of Partial Means and a General Variance Estimator. *Econometric Theory* 10, 233–253.

Newey, W.K. and D. McFadden (1994) Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics*, Vol. 4 (eds. R.F. Engle and D.L. McFadden), Chapter 36. Elsevier Science B.V.

Newey, W.K. and R. Smith (2004) Higher-order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica* 72, 219–255.

Norets, A. (2009) Inference in Dynamic Discrete Choice Models with Serially Correlated Unobserved State Variables. *Econometrica* 77, 1665–1682.

Norets, A. (2012) Estimation of Dynamic Discrete Choice Models Using Artificial Neural Network Approximations. *Econometric Reviews* 31, 84–106.

Nze, P.A. and P. Doukhan (2004) Weak Dependence: Models and Applications to Econometrics. *Econometric Theory* 20, 995–1045.

Olsson, J. and T. Rydén (2008) Asymptotic Properties of Particle Filter-Based Maximum Likelihood Estimators for State Space Models. *Stochastic Processes and their Applications* 118, 649–680.

Pakes, A. and D. Pollard (1989) Simulation and the Asymptotics of Optimization Estimators. *Econometrica* 57, 1027–57.

Rio, E. (1994) Inégalités de moments pour les suites stationnaires et fortement mélangeantes. *Comptes rendus de l'Académie des Sciences* 318, 355–360.

Robinson, P.M. (1988) The Stochastic Difference Between Econometric Statistics. *Econometrica* 56, 531–548.

Rothenberg, T.J. (1984) Approximating the Distributions of Econometric Estimators and Test Statistics. In *Handbook of Econometrics*, vol. 2, eds. K. Arrow and M. Intriligator. North Holland.

Rust, J. (1997) Using Randomization to Break the Curse of Dimensionality. *Econometrica* 65, 487–516.

Tauchen, G. and R. Hussey (1991) Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models. *Econometrica* 59, 371–396.

Train, K. (2009) *Discrete Choice Methods with Simulation*, Cambridge University Press.

Yoshihara, K. (1976) Limiting Behaviour of U-Statistics for Stationary, Absolutely Regular Processes. *Zeitschrift für Wahrenscheinlichkeittheorie und verwandte Gebeite* 35, 237–252.

A Proofs of the Main Results

Proof of Theorem 2. By Lemma 6,

$$\hat{\theta}_{n,S} - \hat{\theta}_n = -H_0^{-1} \left\{ G_n(\theta_0, \hat{\gamma}_S) - G_n(\theta_0, \gamma_0) \right\} + o_P\left(1/\sqrt{n}\right).$$

We now use the expansion given in (9) with $m = 2$; this yields

$$\left\| \hat{\theta}_{n,S} - \hat{\theta}_n \right\| = O_P\left(\left\| \nabla G_n(\theta_0) \left[\Delta\hat{\gamma}_S \right] + \frac{1}{2} \nabla^2 G_n(\theta_0) \left[\Delta\hat{\gamma}_S, \Delta\hat{\gamma}_S \right] + R_{n,S} \right\| \right) + o_P\left(1/\sqrt{n}\right), \tag{29}$$

where $\Delta\hat{\gamma}_{i,S} = \hat{\gamma}_{i,S} - \gamma_0$.

We first derive the rate of the remainder term $R_{n,S}$:

$$\begin{aligned} E\left[\|R_{n,S}\|\right] &= E\left\| G_n(\theta_0, \hat{\gamma}_S) - G_n(\theta_0, \gamma_0) - \nabla G_n(\theta_0)\left[\Delta\hat{\gamma}_S\right] - \frac{1}{2}\nabla^2 G_n(\theta_0)\left[\Delta\hat{\gamma}_S, \Delta\hat{\gamma}_S\right] \right\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} E\left\| g_i(\theta_0, \hat{\gamma}_{i,S}) - g_i(\theta_0, \gamma_0) - \nabla g_i(\theta_0)\left[\Delta\hat{\gamma}_{i,S}\right] - \frac{1}{2}\nabla^2 g_i(\theta_0)\left[\Delta\hat{\gamma}_{i,S}, \Delta\hat{\gamma}_{i,S}\right] \right\| \\ &\leq \frac{\bar{G}_0}{n}\sum_{i=1}^{n} E\left[\left\| \Delta\hat{\gamma}_{i,S} \right\|^3 \right], \end{aligned}$$

where we have used A.4(2). Applying first Minkowski's inequality and then the inequality $(a+b)^p \leq 2^{p-1}a^p + 2^{p-1}b^p$ (which holds for all $a, b > 0$ and $p \geq 1$), we obtain—dropping the i index:

$$\begin{aligned} E\left[\|\Delta\hat{\gamma}_S\|^3\right] &= E\left[\|\psi_S + b_S\|^3\right] \\ &\leq (E\left[\|\psi_S\|\right] + E\left[\|b_S\|\right])^3 \\ &\leq 4E\left[\|\psi_S\|^3\right] + 4E\left[\|b_S\|^3\right] = O\left(S^{-\alpha_3}\right) + O\left(S^{-3\beta}\right). \end{aligned}$$

The rates of the first and second order functional differentials of $G_n(\theta_0, \gamma)$ are given in Lemmas 9 and 10 depending on whether the ECA approximator of (11) or the EIA approximator of (12) is used. These rates together with the rate of $R_{n,S}$ and (29) yield the higher-order stochastic expansion of the EIA and ECA in equation (18).

The weak convergence of D_n and E_n follows by the CLT for stationary and mixing sequences. Finally, the rates of the leading bias and variance terms as $S \to \infty$ also follow from Lemmas 9 and 10. ■

Proof of Theorem 4. We only give a proof for the case of EIA's; the proof for ECA's follows along the same lines. One can easily show that $\sup_{\theta\in\Theta} ||\nabla^2\hat{G}_n(\theta)|| = o_P(1)$ as $n, S \to \infty$, and it now follows by the same arguments as in the proof of Theorem 2 that $\hat{\theta}^{\text{BA}}_{n,S}$ is consistent.

Next, we take a Taylor expansion:

$$o_P\left(n^{-1/2}\right) = \left\{G_n(\theta_0, \hat{\gamma}_S) - \frac{1}{2}\nabla^2\hat{G}_n(\theta_0)\right\} + \left\{H_n(\bar{\theta}_{n,S}, \hat{\gamma}_S) - \frac{1}{2}\nabla^2\hat{H}_n\left(\bar{\theta}_{n,S}\right)\right\}(\hat{\theta}^{\text{AB}}_{n,S} - \theta_0),$$

where $\nabla^2\hat{H}_n(\theta) = \partial\nabla^2\hat{G}_n(\theta)/(\partial\theta)$. From the proof of Theorem 2, $H_n(\bar{\theta}_{n,S}, \hat{\gamma}_S) = H_0 + o_P(1)$, while it is easily shown that $\nabla^2\hat{H}_n\left(\bar{\theta}_{n,S}\right) = o_P(1)$ as $n, S \to 0$, so that, by the same arguments as in the proof of Theorem 2,

$$\hat{\theta}^{\text{AB}}_{n,S} - \hat{\theta}_n = H_0^{-1}\left\{G_n(\theta_0, \hat{\gamma}_S) - \frac{1}{2}\nabla^2\hat{G}_n(\theta_0) - G_n(\theta_0, \gamma)\right\} + o_P\left(1/\sqrt{n}\right).$$

Suppressing any dependence on θ_0, use (9) to write

$$\begin{aligned} G_n(\hat{\gamma}_S) - \frac{1}{2}\nabla^2\hat{G}_n - G_n(\gamma) &= \frac{1}{2}\left\{\nabla^2 G_n[\psi_{n,S}, \psi_{n,S}] - \nabla^2\hat{G}_n\right\} + \nabla G_n[\hat{\gamma}_S - \gamma] \\ &\quad + \frac{1}{2}\left\{\nabla^2 G_n[\hat{\gamma}_S - \gamma, \hat{\gamma}_S - \gamma] - \nabla^2 G_n[\psi_{n,S}, \psi_{n,S}]\right\} + R_{n,S}. \end{aligned} \tag{30}$$

The rates of the second and third terms of (30) are derived in Lemma 10. To ensure that $R_{n,S}$ is negligible, we build on Lemma 11, which uses A.6 to deliver a better rate than that used the proof of Theorem 2.

The crucial term is the first term of (30). Now, recall that $\hat{\gamma}_i = S^{-1}\sum_{s=1}^{S} w_{is}$, and the definition of $\nabla^2\hat{G}_n$ in eq. (20). Using the bilinearity of $(d\gamma, d\gamma') \mapsto \nabla^2 g_i[d\gamma, d\gamma']$, and denoting $\bar{w}_i = E[w_{i,s}]$ and $e_{is} = w_{is} - \bar{w}_i$, the first term of (30) can be rewritten as

$$\begin{aligned}
&\nabla^2 G_n[\psi_{n,S}, \psi_{n,S}] - \nabla^2\hat{G}_n \\
= \;& \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s\neq t}\nabla^2 g_i[e_{is}, e_{it}] + \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S}\nabla^2 g_i[e_{is}, e_{is}] - \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S}\nabla g_i[w_{is}-\hat{\gamma}_i, w_{is}-\hat{\gamma}_i] \\
= \;& \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s\neq t}\nabla^2 g_i[e_{is}, e_{it}] + \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S}\left\{\nabla^2 g_i[e_{is}, e_{is}] - \nabla g_i[w_{is}-\hat{\gamma}_i, w_{is}-\hat{\gamma}_i]\right\} \\
= \;& \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s\neq t}\nabla^2 g_i[e_{is}, e_{it}] + \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s=1}^{S}\left\{\nabla^2 g_i[\hat{\gamma}_i-\bar{w}_i, e_{is}] + \nabla^2 g_i[e_{is}, \hat{\gamma}_i-\bar{w}_i]\right\} \\
= \;& \frac{1}{nS^2}\sum_{i=1}^{n}\sum_{s\neq t}\nabla^2 g_i[e_{is}, e_{it}] + \frac{2}{nS}\sum_{i=1}^{n}\nabla^2 g_i[\hat{\gamma}_i-\bar{w}_i, \hat{\gamma}_i-\bar{w}_i],
\end{aligned}$$

where the last equality uses the fact that $S^{-1}\sum_{s=1}^{S} e_{is} = \hat{\gamma}_i - \bar{w}_i$.

Start with the first term, and note that $E\left[\nabla^2 g_i[e_{is}, e_{it}]\right] = 0$ when $s \neq t$. Then apply Lemma 7 with $r = 1$ to $W_{i,S} := S^{-2}\sum_{s\neq t}\nabla^2 g_i[e_{is}, e_{it}]$, getting

$$\mathrm{Var}\left(\frac{1}{2nS^2}\sum_{i=1}^{n}\sum_{s\neq t}\nabla^2 g_i[e_{is}, e_{it}]\right) \leq \frac{C}{n}E\left[\|W_{i,S}\|^{2+\delta}\right]^{2/(2+\delta)}.$$

Now $W_{i,S}$ is a degenerate U-statistic since

$$E\left[\nabla^2 g(z_i)[e_{is}, e_{it}]|z_i, e_{it}\right] = E\left[\nabla^2 g(z_i)[e_{is}, e_{it}]|z_i, e_{is}\right] = 0.$$

Given the conditions imposed on $\{e_{i,s} : 1 \leq s \leq S\}$ in (A.6), we can employ U-statistic results for absolutely regular sequences: Yoshihara (1976, Lemma 3) states that $E\left[\|W_{i,S}\|^4 |z_i\right] = O\left(S^{-4}\right)$. By inspection of the proof of Yoshihara (1976, Lemma 3), it is easily checked that in fact, for some constant $C > 0$ we have $E\left[\|W_{i,S}\|^4 |z_i\right] \leq CS^{-4}M_S(z_i)$, where

$$M_S(z_i) := \sup_{s<t} E\left[\left\|\nabla^2 g(z_i)[e_{is}, e_{it}]\right\|^{4+\epsilon} |z_i\right]^{4/(4+\epsilon)}, \text{ for some } \epsilon > 0.$$

Thus, with $\delta = 2$ and using the Lipschitz condition on $\nabla^2 g$, we obtain

$$\begin{aligned}
E\left[\|W_{i,S}\|^4\right] &\leq CS^{-4}E\left[M_S\left(z_i\right)\right] \\
&\leq CS^{-4}E\left[\sup_{s<t} E\left[\left\|\nabla^2 g(z_i)[e_{is},e_{it}]\right\|^{4+\epsilon}|z_i\right]^{4/(4+\epsilon)}\right] \\
&\leq CS^{-4}E\left[b^4(z_i)\sup_{s<t} E\left[\|e_{is}\left(z\right)]\|^{4+\epsilon}\|e_{it}(z)]\|^{4+\epsilon}|z_i\right]^{4/(4+\epsilon)}\right] \\
&\leq CS^{-4}E\left[b^4(z_i)E\left[\|e_{is}(z)]\|^{8+\epsilon}|z_i\right]^{4/(8+\epsilon)}\right] \\
&\leq CS^{-4}\sqrt{E\left[b^8(z_i)\right]}E\left[\|e_{is}\|^{8+2\epsilon}\right]^{4/(8+2\epsilon)} \\
&= O\left(S^{-4+\mu_8/2}\right).
\end{aligned}$$

It follows that $\sum_{i=1}^n \sum_{s\neq t} \nabla^2 g_i[e_{is},e_{it}]/\left(nS^2\right) = O_P(n^{-1/2}S^{-1+\mu_8/4})$.

As for the second term, by definition $\hat{\gamma}_i - \bar{w}_i = \psi_{i,S}$; and it follows from Lemma 8 that $E\left[\nabla^2 g_i[\psi_{i,S},\psi_{i,S}]\right] = O\left(S^{-\alpha_2}\right)$ and

$$\frac{1}{n}\sum_{i=1}^n\left(\nabla^2 g_i[\psi_{i,S},\psi_{i,S}] - E\left[\nabla^2 g_i[\psi_{i,S},\psi_{i,S}]\right]\right) = O_P\left(n^{-1/2}S^{-\alpha_4/2}\right).$$

Summing up, $\tilde{B}_2 = H_0^{-1}E\left[\nabla^2 G_n[\psi_{n,S},\psi_{n,S}] - \nabla^2\hat{G}_n\right]/2 = O\left(S^{-2+\mu_2}\right)$ while

$$\text{Var}\left(\nabla^2 G_n[\psi_{n,S},\psi_{n,S}] - \nabla^2\hat{G}_n\right) = O(n^{-1}S^{-2+\mu_8/2}) + O\left(n^{-1}S^{-2+\alpha_4}\right).$$

This completes the proof. ■

Proof of Theorem 5. To apply the general result in Robinson (1988, Theorem 2), we need to check that his conditions A.1 and A.3 are satisfied in our application. His condition A.1 requires consistency of the approximate estimator for a suitable choice of S, which our assumptions imply. Robinson's condition A.3 also holds, given the smoothness conditionswe imposed on $G_n(\theta,\hat{\gamma}_S)$ in our Assumption A.2. ■

B Lemmas

We first derive the expansion in (5):

Lemma 6 *Under Assumptions A.1 and A.2, eq. (5) holds.*

Proof. We first take a Taylor expansion of $G_n(\theta, \gamma_0)$ and $G_n(\theta, \hat{\gamma}_S)$ w.r.t. θ:

$$o_P\left(n^{-1/2}\right) = G_n(\hat{\theta}_n, \gamma_0) = G_n(\theta_0, \gamma_0) + H_n(\bar{\theta}_n, \gamma_0)(\hat{\theta}_n - \theta_0), \tag{31}$$

$$o_P\left(n^{-1/2}\right) = G_n(\hat{\theta}_{n,S}, \hat{\gamma}_S) = G_n(\theta_0, \hat{\gamma}_S) + H_n(\tilde{\theta}_{n,S}, \hat{\gamma}_S)(\hat{\theta}_{n,S} - \theta_0), \tag{32}$$

for some $\bar{\theta}_n$ ($\tilde{\theta}_{n,S}$) between $\hat{\theta}_n$ ($\hat{\theta}_{n,S}$) and θ_0. Since $\hat{\theta}_n$ ($\hat{\theta}_{n,S}$) is consistent, $\bar{\theta}_{n,S}$ ($\tilde{\theta}_{n,S}$) $\xrightarrow{P} \theta_0$. By standard arguments together with Assumption A.2,

$$\begin{aligned}
\left\|H_n\left(\tilde{\theta}_{n,S}, \hat{\gamma}_S\right) - H_0\right\| &\leq \left\|H_n\left(\tilde{\theta}_{n,S}, \hat{\gamma}_S\right) - H_n\left(\tilde{\theta}_{n,S}, \gamma_0\right)\right\| + \left\|H_n\left(\tilde{\theta}_{n,S}, \gamma_0\right) - H\left(\tilde{\theta}_{n,S}, \gamma_0\right)\right\| \\
&\quad + \left\|H\left(\tilde{\theta}_{n,S}, \gamma_0\right) - H\left(\theta_0, \gamma_0\right)\right\| \\
&\leq \sup_{\|\theta-\theta_0\|\leq\delta} \|H_n(\theta, \hat{\gamma}_S) - H_n(\theta, \gamma_0)\| + \sup_{\|\theta-\theta_0\|\leq\delta} \|H_n(\theta, \gamma_0) - H(\theta, \gamma_0)\| \\
&\quad + \left\|H\left(\tilde{\theta}_{n,S}, \gamma_0\right) - H\left(\theta_0, \gamma_0\right)\right\| \\
&= o_P(1),
\end{aligned}$$

and similar for $H_n(\bar{\theta}_n, \gamma_0)$. Going back to eqs. (31)-(32), we have now shown that

$$\hat{\theta}_{n,S} - \theta_0 = -H_0^{-1} G_n(\theta_0, \hat{\gamma}_S) + o_P\left(1/\sqrt{n}\right), \quad \hat{\theta}_n - \theta_0 = -H_0^{-1} G_n(\theta_0, \gamma_0) + o_P\left(1/\sqrt{n}\right).$$

Subtracting the second expansion from the first gives the result. ■

To establish the rates for the first and second order differentials, we first establish some useful auxiliary results:

Lemma 7 *Let $\{W_i\}$ be a sequence of random variables with $E[W_i] = 0$, $E\left[\|W_i\|^{2r+\delta}\right] < \infty$ for some $r \geq 1$ and $\delta > 0$.*

Assume that (W_i) is α-mixing with mixing coefficients α_i, $i = 1, 2, ...$, satisfying $\alpha_i \leq Ai^{-a}$ for some $A > 0$, and $a > 2r + 4r(r-1)/\delta - 2$.

Then there exists a constant $C = C(r, a, A) < \infty$ such that:

$$E\left[\left\|\frac{1}{n}\sum_{i=1}^n W_i\right\|^{2r}\right] \leq n^{-r} \times CE\left[\|W_i\|^{2+\delta}\right]^{2r/(2+\delta)} + o\left(n^{-r}\right).$$

Proof. From Rio (1994), we have for $r \geq 1$,

$$E\left[\left\|\frac{1}{n}\sum_{i=1}^n W_i\right\|^{2r}\right] \leq C_r\left[n^{-r} M_{2,\alpha,n}^r + n^{1-2r} M_{2r,\alpha,n}\right], \tag{33}$$

where the numbers $M_{p,\alpha,n}$ are defined in Rio (1994). By Nze and Doukhan (2004, p. 1040),

$$M_{p,\alpha,n} \leq \left[E\left\|W_i\right\|^{p+\delta}\right]^{p/(p+\delta)} \times \frac{(p+\delta)(p-1)}{\delta} \sum_{n=0}^{\infty} (n+1)^{p+p(p-1)/\delta-2} \alpha_n.$$

Given the bound we imposed on the mixing coefficients, there exists a constant $C(A,a)$ such that

$$\sum_{n=0}^{\infty} (n+1)^{p+p(p-1)/\delta-2} \alpha_n \leq C(A,a) \sum_{n=0}^{\infty} (n+1)^{p+p(p-1)/\delta-2-a} < \infty.$$

In particular, there exist constants $C(r,A,a)$ such that

$$M_{2,\alpha,n}^r \leq C(r,A,a)\left[E\left\|W_i\right\|^{2+\delta}\right]^{2r/(2+\delta)}, \quad \text{and} \quad M_{2r,\alpha,n} \leq C(r,A,a)\left[E\left\|W_i\right\|^{2r+\delta}\right]^{2r/(2r+\delta)}. \tag{34}$$

The result follows by noting that $n^{1-2r} = o(n^{-r})$ for $r > 1$, and that for $r = 1$ both terms in equation (33) are of order $n^{-1} = n^{-r}$. ■

Lemma 8 *Assume that $\{z_i\}$ satisfies Assumption A.1, and that $\hat{\gamma}_{j,S}$ satisfy Assumption A.5(4) for $j = 1, ..., J$. Let $m(z; d\gamma)$ be a functional satisfying:*

$$E\left[\left\|m(z;d\gamma)\right\|^{2r+\delta}\right] < \infty, \quad E\left[\left\|m(z;d\gamma)\right\|^{2+\delta}\right] \leq \bar{M}\left\|d\gamma\right\|^{k(2+\delta)}, \tag{35}$$

for some $r, k \geq 1$ and $\delta > 0$.

Then, with b_S and ψ_S given in A.5, the following hold:

(i) For EIA's, with $M_S^V := E\left[m\left(z_i;\psi_{i,S}\right)\right]$ and $M_S^B := E\left[m\left(z_i;b_{i,S}\right)\right]$,

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n}\left\{m\left(z_i;b_{i,S}\right) - M_S^B\right\}\right\|^{2r}\right] = O\left(n^{-r}\right) \times \left[E\left\|b_S\right\|^{k(2+\delta)}\right]^{2r/(2+\delta)},$$

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n}\left\{m\left(z_i;\psi_{i,S}\right) - M_S^V\right\}\right\|^{2r}\right] = O\left(n^{-r}\right) \times \left[E\left\|\psi_S\right\|^{k(2+\delta)}\right]^{2r/(2+\delta)}.$$

(ii) For ECA's, with $\bar{m}(\gamma) = E\left[m\left(z;\gamma\right)\right]$ for any fixed γ,

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n}\left\{m\left(z_i;b_S\right) - \bar{m}\left(b_S\right)\right\}\right\|^{2r}\right] = O\left(n^{-r}\right) \times \left[E\left\|\psi_S\right\|^{k(2+\delta)}\right]^{2r/(2+\delta)},$$

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n}\left\{m\left(z_i;\psi_S\right) - \bar{m}\left(\psi_S\right)\right\}\right\|^{2r}\right] = O\left(n^{-r}\right) \times \left[E\left\|\psi_S\right\|^{k(2+\delta)}\right]^{2r/(2+\delta)}.$$

(iii) The means satisfy:

$$\left\|M_S^B\right\| \leq \bar{M}E\left[\left\|b_{i,S}\right\|^k\right], \quad \left\|M_S^V\right\| \leq \bar{M}E\left[\left\|\psi_{i,S}\right\|^k\right], \quad E\left[\left\|\bar{m}\left(\psi_S\right)\right\|^{2r}\right] \leq \bar{M}E\left[\left\|\psi_S\right\|^{2kr}\right].$$

Proof. Define $W_{i,S} := m\left(z_i;\psi_{i,S}\right) - M_S\left(\psi_{i,S}\right)$. By assumptions (A.1) and (A.5), $\{W_{i,S}\}$ is a geometrically mixing process for any given value of S and so its mixing coefficients satisfy the mixing conditions imposed in Lemma 7. Furthermore, (35) implies that $E\left[\left\|W_{i,S}\right\|^{2r+\delta}\right] < \infty$. We can therefore apply Lemma 7

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n}\left\{m\left(z_i;\psi_{i,S}\right) - M_S\left(\psi_{i,S}\right)\right\}\right\|^{2r}\right] \leq Cn^{-r}\left[E\left\|m\left(z_i;\psi_{i,S}\right) - M_S\left(\psi_{i,S}\right)\right\|^{2+\delta}\right]^{2r/(2+\delta)} + o\left(n^{-r}\right)$$

where $C = C\left(r,a,A\right)$ only depends on r and the mixing coefficients of $\{z_i\}$ and $\left\{\psi_{i,S}\right\}$. By (35),

$$E\left[\left\|m\left(z;\psi_{i,S}\right)\right\|^{2+\delta}\right] \leq \bar{M}E\left[\left\|\psi_{i,S}\right\|^{k(2+\delta)}\right]n^{-r},$$

and

$$\left\|M_S\left(\psi_{i,S}\right)\right\| \leq E\left[\left\|m\left(z_i;\psi_{i,S}\right)\right\|\right] \leq \bar{M}E\left[\left\|\psi_{i,S}\right\|^k\right].$$

It is easily seen that the above inequalities still go through when replacing $\psi_{i,S}$ with $b_{i,S}$. This prove (i) and (iii).

To derive the second inequality of (ii), now redefine $W_{i,S}$ as $W_{i,S} := m\left(z_i;\psi_S\right) - \bar{m}\left(\psi_S\right)$. It is easily seen that conditionally on ψ_S, $(W_{i,S})$ satisfies the conditions of Lemma 7, so that

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n}W_{i,S}\right\|^{2r}\left|\psi_S\right.\right] \leq CE\left[\left\|W_{i,S}\right\|^{2+\delta}\left|\psi_S\right.\right]n^{-r} + o\left(n^{-r}\right),$$

where $C = C\left(r,a,A\right)$ does not depend on ψ_S. Next, observe that

$$E\left[\left\|W_{i,S}\right\|^{2+\delta}\right] \leq CE\left[\left\|m\left(z;\psi_S\right)\right\|^{2+\delta}\right] \leq C\bar{M}E\left[\left\|\psi_S\right\|^{k(2+\delta)}\right];$$

we conclude that

$$E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n}W_{i,S}\right\|^{2r}\right] = E\left[E\left[\left\|\frac{1}{n}\sum\nolimits_{i=1}^{n}W_{i,S}\right\|^{2r}\left|\psi_S\right.\right]\right] \leq CE\left[\left\|\psi_S\right\|^{k(2+\delta)}\right]n^{-r} + o\left(n^{-r}\right).$$

Finally,

$$E\left[\left\|\bar{m}\left(\psi_S\right)\right\|^{2r}\right] \leq E\left[\left\|m\left(z;\psi_S\right)\right\|^{2r}\right] \leq \bar{M}E\left[\left\|\psi_S\right\|^{2rk}\right].$$

The proof of the first inequality of (ii) follows along the same lines. ■

In the next three lemmas, we suppress the dependence on θ since it is kept fixed at the true value θ_0.

Lemma 9 *Under A.1-A.3, A.4(2), and A.6(4), the first and second order differentials of $G_n(\theta_0, \hat{\gamma}_S)$ for the ECA yield the rates given in Theorem 2.*

Proof. First consider the EIA case, in which the approximation of $G_n(\gamma)$ is on the form of eq. (12). The functional differentials of G_n are given by

$$\nabla G_n[d\gamma] = \frac{1}{n}\sum_{i=1}^{n} \nabla g_i[d\gamma], \quad \nabla^2 G_n\left[d\gamma, d\gamma'\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 g_i\left[d\gamma, d\gamma'\right],$$

and $d\gamma$ and $d\gamma'$ are the same for all observations $i = 1, \ldots, n$.

Given A.6(4), the application of the first-order differential to the bias component can be rewritten as

$$\nabla G_n[b_S] = S^{-\beta}\frac{1}{n}\sum_{i=1}^{n} \nabla g_i\left[\bar{b}\right] + \frac{1}{n}\sum_{i=1}^{n} \nabla g_i\left[b_S - S^{-\beta}\bar{b}\right].$$

Now,

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \nabla g_i\left[\bar{b}\right]\right] = E\left[\nabla g_i\left[\bar{b}\right]\right], \text{ and}$$

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \left\|\nabla g_i\left[b_S - S^{-\beta}\bar{b}\right]\right\|\right] \leq G_1\left\|b_S - S^{-\beta}\bar{b}\right\| = o\left(S^{-\beta}\right).$$

By Lemma 8(i) with $m(z; d\gamma) = \nabla g(z)[d\gamma]$, $k = 1$ and $r = 1$,

$$\mathrm{Var}\left(\nabla G_n[b_S]\right) \leq \frac{1}{n} C \left\|b_S\right\|^2 = O\left(\frac{S^{-2\beta}}{n}\right).$$

Since $d\gamma \mapsto \nabla g_i[d\gamma]$ is linear, the conditional mean of the stochastic component of the first-order term is

$$E\left[\nabla G_n[\psi_S]|\mathcal{Z}_n\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla g_i\left[E\left[\psi_S|z_i\right]\right] = 0.$$

Moreover, define $\nabla G[\gamma] = E[\nabla g_i[\gamma]]$ (where expectations are taken w.r.t. the observation z_i); then

$$\nabla G_n[\psi_S] = \nabla G[\psi_S] + \frac{1}{n}\sum_{i=1}^{n} \left\{\nabla g_i[\psi_S] - \nabla G[\psi_S]\right\}.$$

Recalling the definition of $\nabla G[\psi_S]$, it follows from Lemma 8.(ii) with $m(z; d\gamma) = \nabla g(z)[d\gamma]$ and $k = 2$ that the first term satisfies $\mathrm{Var}(\nabla G[\psi_S]) \leq ME\left[\left\|\psi_S\right\|^2\right] = O\left(S^{-\alpha_2}\right)$ while the second term is $O_P(n^{-1/2}S^{-\alpha_2})$.

Regarding the second order differential, its application to the bias component satisfies

$$\nabla^2 G_n[b_S, b_S] = S^{-2\beta} \frac{1}{n} \sum_{i=1}^{n} \nabla^2 g_i \left[\bar{b}, \bar{b}\right] + o_P\left(S^{-2\beta}\right);$$

moreover,

$$E\left[\frac{1}{n} \sum_{i=1}^{n} \nabla^2 g_i \left[\bar{b}, \bar{b}\right]\right] = E\left[\nabla^2 g_i \left[\bar{b}, \bar{b}\right]\right],$$

and, applying Lemma 8.(ii) with $m\left(z; d\gamma\right) = \nabla^2 g(z)\left[d\gamma, d\gamma\right]$, $k = 2$ and $r = 1$,

$$\text{Var}\left(\nabla^2 G_n[b_S, b_S]\right) \leq \frac{1}{n} C \left\|b_S\right\|^4 = O\left(n^{-1} S^{-4\beta}\right).$$

To bound the variance component, define $\nabla^2 G\left[\gamma, \gamma\right] = E\left[\nabla^2 g_i\left[\gamma, \gamma\right]\right]$, and write

$$\nabla^2 G_n[\psi_S, \psi_S] = \nabla^2 G\left[\psi_S, \psi_S\right] + \frac{1}{n} \sum_{i=1}^{n} \left(\nabla^2 g_i\left[\psi_S, \psi_S\right] - \nabla^2 G\left[\psi_S, \psi_S\right]\right).$$

Applying Lemma 8(ii) with $m\left(z; d\gamma\right) = \nabla^2 g(z)\left[d\gamma, d\gamma\right]$ and $r = 1, k = 2$, we obtain that $E\left\|\nabla^2 G_n\left[\psi_S, \psi_S\right]\right\| = O_P\left(S^{-2\alpha_2}\right)$.

Finally, by the same arguments as before, $E\left[\nabla^2 G_n[\psi_S, b_S]\right] = 0$ while $\text{Var}\left(\nabla^2 G_n[\psi_S, b_S]\right) = O(n^{-1} S^{-\alpha_4})$ and $\text{Var}\left(\nabla^2 G_n[\psi_S, b_S]\right) = O(n^{-1} S^{-\alpha_2 - 2\beta})$. ■

Lemma 10 *Under A.1-A.3, A.4(2) and A.5(4), the first and second order differentials of $G_n(\theta_0, \gamma_S)$ for the EIA in (11) yield the rates given in Theorem 2.*

Proof. For the EIA, the first and second order differentials are $\nabla G_n\left[d\gamma\right] = \sum_{i=1}^{n} \nabla g_i\left[d\gamma_i\right]/n$ and $\nabla^2 G_n)\left[d\gamma, d\gamma'\right] = \sum_{i=1}^{n} \nabla^2 g_i\left[d\gamma_i, d\gamma_i'\right]/n$, for any $d\gamma = (d\gamma_1, ..., d\gamma_n)$ and $d\gamma' = (d\gamma_1', ..., d\gamma_n')$. It is easily seen that the bias components are the same as those we derived for the ECA in Lemma 9, and so we only consider the variance components. With $\mathcal{Z}_n = (z_1, ..., z_n)$, the mean of the first-order variance component is zero,

$$E\left[\nabla G_n[\psi_S] | \mathcal{Z}_n\right] = \frac{1}{n} \sum_{i=1}^{n} \nabla g_i\left[E\left[\psi_{i,S} | z_i\right]\right] = 0,$$

while its variance satisfies, using Lemma 8.(i) with $m\left(z, \gamma\right) = \nabla g(z)\left[\gamma\right]$ (in particular, $M_S^V = 0$),

$$\text{Var}\left(\nabla G_n[\psi_S]\right) \leq \frac{1}{n} C E\left[\left\|\psi_S\right\|^2\right] = O\left(n^{-1} S^{-\alpha_2}\right).$$

Applying Lemma 8(i) and (iii) with $m(z;d\gamma) = \nabla^2 g(z)[d\gamma, d\gamma]$ and $k = 2$, the mean and the variance of the second order differential satisfy

$$E\left[\nabla^2 G_n[\psi_S, \psi_S]\right] = E\left[\nabla^2 g_i\left[\psi_{i,S}, \psi_{i,S}\right]\right] \leq CE\left[\left\|\psi_{i,S}\right\|^2\right] = O\left(S^{-\alpha_2}\right),$$

and $\mathrm{Var}\left[\nabla^2 G_n[\psi_S, \psi_S]\right] = O(n^{-1}S^{-\alpha_4})$. The cross term satisfies $E\left[\nabla^2 G_n[\psi_S, b_S]\right] = 0$ while $\mathrm{Var}\left(\nabla^2 G_n[\psi_S, b_S]\right) = O(n^{-1}S^{-\alpha_2}S^{-2\beta})$, and so we can ignore this term since it is of lower order. ■

Lemma 11 *Assume that A.1-A.3, A.4*(3) *and A.6*(6) *hold. Then the rate of the remainder term $R_{n,S}$ can be sharpened to:*

$$R_{n,S} = O_P\left(S^{-3\beta}\right) + O_P\left(S^{-(2-\mu_4)}\right) + O\left(S^{-(2-\mu_3)}\right) + O\left(n^{-1/2}S^{-(3-\mu_6)/2}\right).$$

Proof. Since the third-order differential exists, the remainder term in (9) can be further expanded to obtain $R_{n,S} = \nabla^3 G_n\left[\Delta\hat{\gamma}_S, \Delta\hat{\gamma}_S, \Delta\hat{\gamma}_S\right]/6 + \bar{R}_{n,S}$ where, by A.4(3) and the same arguments used in the proof of Theorem 2, $E\left[\left\|\bar{R}_{n,S}\right\|\right] \leq \bar{G}_0 E\left[\left\|\Delta\hat{\gamma}_{i,S}\right\|^4\right] = O\left(S^{-4\beta}\right) + O\left(S^{-(2-\mu_4)}\right)$. Regarding the third order term, it is easy to check that the bias component is of order $O_P\left(S^{-3\beta}\right) + O_P\left(n^{-1/2}S^{-3\beta}\right)$, by arguments similar to those used in Lemma 9.

This leaves the variance component. In the case of EIA, the variance component can be written as $\nabla^3 G_n\left[\psi_S, \psi_S, \psi_S\right] = \sum_{i=1}^n \nabla^3 g_i\left[\psi_S, \psi_S, \psi_S\right]/n$. By Lemma 8, we obtain:

$$\nabla^3 G_n\left[\psi_S, \psi_S, \psi_S\right] - E\left[\nabla^3 G_n\left[\psi_S, \psi_S, \psi_S\right]\right] = O\left(n^{-1/2}S^{-(3-\mu_6)/2}\right);$$

given the independence between simulations,

$$\begin{aligned}
\left|E\left[\nabla^3 G_n\left[\psi_S, \psi_S, \psi_S\right]\right]\right| &\leq \frac{1}{S^3}\sum_{s,t,u=1}^{S}\left|E\left[\nabla^3 g_i\left[e_{i,s}, e_{i,t}, e_{i,u}\right]\right]\right| \\
&= \frac{\left|E\left[\nabla^3 g_i\left[e_{i,s}, e_{i,s}, e_{i,s}\right]\right]\right|}{S^2} \\
&\leq \frac{C}{S^2}E\left[e_{i,s}^3\right] = O(S^{-(2-\mu_3)}).
\end{aligned}$$

In the case of ECA, define $\nabla^3\bar{g}\left[\gamma,\gamma,\gamma\right] = E\left[\nabla^2 g_i\left[\gamma,\gamma,\gamma\right]\right]$ and write

$$\nabla^3 G_n[\psi_S, \psi_S, \psi_S] = \nabla^3\bar{g}\left[\psi_S, \psi_S, \psi_S\right] + \frac{1}{n}\sum_{i=1}^{n}\left\{\nabla^3 g_i\left[\psi_S, \psi_S, \psi_S\right] - \nabla^3\bar{g}\left[\psi_S, \psi_S, \psi_S\right]\right\}.$$

Applying Lemma 8.(ii) with $m(z;d\gamma) = \nabla^3 g(z)[d\gamma, d\gamma, d\gamma]$, the two terms are $O_P\left(S^{-(3/2-\mu_3)}\right)$ and $O_P(n^{-1/2}S^{-(3-\mu_6)/2})$ respectively. ■