Generalized Indirect Inference for Discrete Choice Models*

Marianne Bruins†
James A. Duffy†‡
Michael P. Keane†
Anthony A. Smith, Jr.§

July 2015

Abstract

This paper develops and implements a practical simulation-based method for estimating dynamic discrete choice models. The method, which can accommodate lagged dependent variables, serially correlated errors, unobserved variables, and many alternatives, builds on the ideas of indirect inference. The main difficulty in implementing indirect inference in discrete choice models is that the objective surface is a step function, rendering gradient-based optimization methods useless. To overcome this obstacle, this paper shows how to smooth the objective surface. The key idea is to use a smoothed function of the latent utilities as the dependent variable in the auxiliary model. As the smoothing parameter goes to zero, this function delivers the discrete choice implied by the latent utilities, thereby guaranteeing consistency. We establish conditions on the smoothing such that our estimator enjoys the same limiting distribution as the indirect inference estimator, while at the same time ensuring that the smoothing facilitates the convergence of gradient-based optimization methods. A set of Monte Carlo experiments shows that the method is fast, robust, and nearly as efficient as maximum likelihood when the auxiliary model is sufficiently rich.

Note. An earlier version of this paper was circulated as the unpublished manuscript Keane and Smith (2003). That paper proposed the method of generalized indirect inference (GII), but did not formally analyze its asymptotic or computational properties. The present work, under the same title but with two additional authors (Bruins and Duffy), rigorously establishes the asymptotic and computational properties of GII. It is thus intended to subsume the 2003 manuscript. Notably, the availability of the 2003 manuscript allowed GII to be used in numerous applied studies (see Section 3.3), even though the statistical foundations of the method had not been firmly established. The present paper provides these foundations and fills this gap in the literature.

*The authors thank Debopam Battacharya, Martin Browning and Liang Chen for helpful comments. Keane's work on this project has been funded by the Australian Research Council under grant FL110100247. The manuscript was prepared with LyX 2.1.3 and JabRef 2.7b.

†Nuffield College and Department of Economics, University of Oxford
‡Institute for New Economic Thinking at the Oxford Martin School
§Department of Economics, Yale University and National Bureau of Economic Research

Contents

1 Introduction

Many economic models have the features that (i) given knowledge of the model parameters, it is easy to simulate data from the model, but (ii) estimation of the model parameters is extremely difficult. Models with discrete outcomes or mixed discrete/continuous outcomes commonly fall into this category. A good example is the multinomial probit (MNP), in which an agent chooses from among several discrete alternatives the one with the highest utility. Simulation of data from the model is trivial: simply draw utilities for each alternative, and assign to each agent the alternative that gives them the greatest utility. But estimation of the MNP, via either maximum likelihood (ML) or the method of moments (MOM), is quite difficult.

The source of the difficulty in estimating the MNP, as with many other discrete choice models, is that, from the perspective of the econometrician, the probability an agent chooses a particular alternative is a high-dimensional integral over multiple stochastic terms (unobserved by the econometrician) that affect utilities the agent assigns to each alternative. These probability expressions must be evaluated many times in order to estimate the model by ML or MOM. For many years econometricians worked on developing fast simulation methods to evaluate choice probabilities in discrete choice models (see Lerman and Manski, 1981). It was only with the development of fast and accurate smooth probability simulators that ML or MOM-based estimation in these models became practical (see McFadden, 1989, and Keane, 1994).

A different approach to inference in discrete choice models is the method of "indirect inference." This approach (see Smith, 1990, 1993; Gourieroux, Monfort, and Renault, 1993; Gallant and Tauchen, 1996), circumvents the need to construct the choice probabilities generated by the economic model, because it is not based on forming the likelihood or forming moments based on choice frequencies. Rather, the idea of indirect inference (II) is to choose a statistical model that provides a rich description of the patterns in the data. This descriptive model is estimated on both the actual observed data and on simulated data from the economic model. Letting β denote the vector of parameters of the structural economic model, the II estimator is that $\hat{\beta}$ which makes the simulated data "look like" the actual data—in the sense (defined formally below) that the descriptive statistical model estimated on the simulated data "looks like" that same model estimated on the actual data. (The method of moments is thus a special case of II, in which the descriptive statistical model corresponds to a vector of moments.)

Indirect inference holds out the promise that it should be practical to estimate any economic model from which it is practical to simulate data, even if construction of the likelihood or population moments implied by the model is very difficult or impossible. But this promise has not been fully realized because of limitations in the II procedure itself. It is very difficult to apply II to models that include discrete (or discrete/continuous) outcomes for the following reason: small changes in the structural parameters of such models will, in general, cause the data simulated from the model to change discretely. Such a discrete change causes the parameters of a descriptive model fit to the simulated data to jump discretely, and these discontinuities are inherited by the criterion function minimized by the II estimator.

Thus, given discrete (or discrete/continuous) outcomes, the II estimator cannot be implemented using gradient-based optimization methods. One instead faces the difficult computational task of optimizing a d_β-dimensional step function using much slower derivative-free meth-

ods. This is very time-consuming and puts severe constraints on the size of the structural models that can be feasibly estimated. Furthermore, even if estimates can be obtained, one does not have derivatives available for calculating standard errors.

In this paper we propose a "generalized indirect inference" (GII) procedure to address this important problem (Sections 3 and 4). The key idea is to generalize the original II method by applying two different descriptive statistical models to the simulated and actual data. As long as the two descriptive models share the same vector of pseudo-true parameter values (at least asymptotically), the GII estimator based on minimizing the distance between the two models is consistent, and will enjoy the same asymptotic distribution as the II estimator.

While the GII idea has wider applicability, here we focus on how it can be used to resolve the problem of non-smooth objective functions of II estimators in the case of discrete choice models. Specifically, the model we apply to the simulated data does not fit the discrete outcomes in that data. Rather, it fits a "smoothed" version of the simulated data, in which discrete choice indicators are replaced by smooth functions of the underlying continuous latent variables that determine the model's discrete outcomes. In contrast, the model we apply to the actual data is fit to observed discrete choices (obviously, the underlying latent variables that generate actual agents' observed choices are not seen by the econometrician).

As the latent variables that enter the descriptive model applied to the simulated data are smooth functions of the model parameters, the non-smooth objective function problem is obviously resolved. However, it remains to show that the GII estimator based on minimizing the distance between these two models is consistent and asymptotically normal. We show that, under certain conditions on the parameter regulating the smoothing, the GII estimator has the same limiting distribution as the II estimator, permitting inferences to be drawn in the usual manner (Section 5).

Our theoretical analysis goes well beyond merely deriving the limiting distribution of the minimizer of the GII criterion function. Rather, in keeping with computational motivation of this paper, we show that the proposed smoothing facilitates the convergence of derivative-based optimizers, in the sense that the smoothing leads to a sample optimization problem that is no more difficult than the corresponding population problem, where the latter involves the minimization of a necessarily smooth criterion (Section 5). We also provide a detailed analysis of the convergence properties of selected line-search and trust-region methods. Our results on the convergence of these derivative-based optimizers seem to be new to the literature. (While our work here is in some respects related to the theory of k-step estimators, we depart significantly from that literature, for example by dropping the usual requirement that the optimizations commence from the starting values provided by some consistent initial estimator.)

Finally, we provide Monte Carlo evidence indicating that the GII procedure performs well on a set of example models (Section 6). We look at some cases where simulated maximum likelihood (SML) is also feasible, and show that efficiency losses relative to SML are small. We also show how judicious choice of the descriptive (or auxiliary) model is very important for the efficiency of the estimator. This is true not only here, but for II more generally.

Proofs of the theoretical results stated in the paper are given in Appendices B–E. An index of key notation appears in Appendix F. All limits are taken as $n \to \infty$.

2 The model

We first describe a class of discrete choice models that we shall use as test cases for the estimation method that we develop in this paper. As will become clear, however, the ideas underlying the method could be applied to almost any conceivable model of discrete choice, including models with mixed discrete/continuous outcomes, and even models in which individuals' choices solve forward-looking dynamic programming problems.

We henceforth focus mainly on panel data models with n individuals, each of whom selects a choice from a set of J discrete alternatives in each of T time periods. Let u_{itj} be the (latent) utility that individual i attaches to alternative j in period t. Without loss of generality, set the utility of alternative J in any period equal to 0. In each period, each individual chooses the alternative with the highest utility. Let y_{itj} be equal to 1 if individual i chooses alternative j in period t and be equal to 0 otherwise. Define $u_{it} \coloneqq (u_{it1}, \ldots, u_{it,J-1})$ and $y_{it} \coloneqq (y_{it1}, \ldots, y_{it,J-1})$. The econometrician observes the choices $\{y_{it}\}$ but not the latent utilities $\{u_{it}\}$.

The vector of latent utilities u_{it} is assumed to follow a stochastic process

$$u_{it} = f(x_{it}, y_{i,t-1}, \ldots, y_{i,t-l}, \epsilon_{it}; \beta), \qquad t = 1, \ldots, T, \tag{2.1}$$

where x_{it} is a vector of exogenous variables.[1] For each individual i, the vector of disturbances $\epsilon_{it} \coloneqq (\epsilon_{it1}, \ldots, \epsilon_{it,J-1})$ follows a Markov process $\epsilon_{it} = g(\epsilon_{i,t-1}, \eta_{it}; \beta)$, where $\{\eta_{it}\}_{t=1}^{T}$ is a sequence of i.i.d. random vectors (of dimension $J-1$) having a specified distribution (which does *not* depend on β). The functions f and g depend on a set of k structural parameters $\beta \in \mathrm{B}$. The sequences $\{\eta_{it}\}_{t=1}^{T}$, $i = 1, \ldots, n$, are independent across individuals and independent of x_{it} for all i and t. The initial values ϵ_{i0} and y_{it}, $t = 0, -1, \ldots, 1-l$, are fixed exogenously.

Although the estimation method proposed in this paper can (in principle) be applied to any model of this form, we focus on four special cases of the general model. Three of these cases (Models 1, 2, and 4 below) can be feasibly estimated using simulated maximum likelihood, allowing us to compare its performance with that of the proposed method.

Model 1. $J = 2$, $T > 1$, and $u_{it} = bx_{it} + \epsilon_{it}$, where x_{it} is a scalar, $\epsilon_{it} = r\epsilon_{i,t-1} + \eta_{it}$, $\eta_{it} \sim_{\text{i.i.d.}} N[0,1]$, and $\epsilon_{i0} = 0$. This is a two-alternative dynamic probit model with serially correlated errors; it has two unknown parameters b and r.

Model 2. $J = 2$, $T > 1$, and $u_{it} = b_1 x_{it} + b_2 y_{i,t-1} + \epsilon_{it}$, where x_{it} is a scalar and ϵ_{it} follows the same process as in Model 1. The initial value y_{i0} is set equal to 0. This is a two-alternative dynamic probit model with serially correlated errors and a lagged dependent variable; it has three unknown parameters b_1, b_2, and r.

Model 3. Identical to Model 2 except that the econometrician does not observe the first $s < T$ of the individual's choices. Thus there is an "initial conditions" problem (see Heckman, 1981).

[1] The estimation method proposed in this paper can also accommodate models in which the latent utilities in any given period depend on lagged values of the latent utilities.

Model 4. $J = 3$, $T = 1$, and the latent utilities obey:

$$u_{i1} = b_{10} + b_{11}x_{i1} + b_{12}x_{i2} + \eta_{i1}$$
$$u_{i2} = b_{20} + b_{21}x_{i1} + b_{22}x_{i3} + c_1\eta_{i1} + c_2\eta_{i2},$$

where $(\eta_{i1}, \eta_{i2}) \sim_{\text{i.i.d.}} N[0, I_2]$. (Since $T = 1$ in this model, the time subscript has been omitted.) This is a static three-alternative probit model; it has eight unknown parameters $\{b_{1k}\}_{k=0}^2$, $\{b_{2k}\}_{k=0}^2$, c_1, and c_2.

The techniques developed in this paper may also be applied to models with a mixture of discrete and continuous outcomes. A leading example is the Heckman selection model:

Model 5. A selection model with two equations: The first equation determines an individual's wage and the second determines his/her latent utility from working:

$$w_i = b_{10} + b_{11}x_{1i} + c_1\eta_{1i} + c_2\eta_{i2}$$
$$u_i = b_{20} + b_{21}x_{2i} + b_{22}w_i + \eta_{i2},$$

Here x_{1i} and x_{2i} are exogenous regressors and $(\eta_{i1}, \eta_{i2}) \sim_{\text{i.i.d.}} N[0, I_2]$. The unknown parameters are $\{b_{1k}\}_{k=0}^1$, $\{b_{2k}\}_{k=0}^2$, c_1, and c_2. Let $y_i \coloneqq I(u_i \geq 0)$ be an indicator for employment status. The econometrician observes the outcome y_i but not the latent utility u_i. In addition, the econometrician observes a person's wage w_i if and only if he/she works (i.e. if $y_i = 1$).

3 Generalized indirect inference

We propose to estimate the model in Section 2 via a generalization of indirect inference. First, in Section 3.1 we exposit the method of indirect inference as originally formulated. In Section 3.2 we explain the difficulty of applying the original approach to discrete choice models. Then, Section 3.3 presents our generalized indirect inference estimator that resolves this difficulty.

3.1 Indirect inference

Indirect inference exploits the ease and speed with which one can typically simulate data from even complex structural models. The basic idea is to view both the observed data and the simulated data through the "lens" of a descriptive statistical (or auxiliary) model characterized by a set of d_θ auxiliary parameters θ. The $d_\beta \leq d_\theta$ structural parameters β are then chosen so as to make the observed data and the simulated data look similar when viewed through this lens.

To formalize these ideas, assume the observed choices $\{y_{it}\}$, $i = 1, \ldots, n$, $t = 1, \ldots, T$, are generated by the structural discrete choice model described in (2.1), for a given value β_0 of the structural parameters. An auxiliary model can be estimated using the observed data to obtain parameter estimates $\hat{\theta}_n$. Formally, $\hat{\theta}_n$ solves:

$$\hat{\theta}_n \coloneqq \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}_n(y, x; \theta) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i; \theta), \tag{3.1}$$

where $\mathcal{L}_n(y, x; \theta)$ is the average log-likelihood function (or more generally, some statistical criterion function) associated with the auxiliary model, $y \coloneqq \{y_{it}\}$ is the set of observed choices, and $x \coloneqq \{x_{it}\}$ is the set of observed exogenous variables.

Let $\eta^m \coloneqq \{\eta_{it}^m\}$ denote a set of simulated draws for the values of the unobservable components of the model, with these draws being independent across $m \in \{1, \ldots, M\}$. Then given x and a parameter vector β, the structural model can be used to generate M corresponding sets of simulated choices, $y^m(\beta) \coloneqq \{y_{it}^m(\beta)\}$. (Note that the same values of x and $\{\eta^m\}$ are used for all β.) Estimating the auxiliary model on the mth simulated dataset thus yields

$$\hat{\theta}_n^m(\beta) \coloneqq \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}_n(y^m(\beta), x; \theta). \tag{3.2}$$

Let $\overline{\theta}_n(\beta) \coloneqq \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_n^m(\beta)$ denote the average of these estimates. Under appropriate regularity conditions, as the observed sample size n grows large (holding M and T fixed), $\overline{\theta}_n(\beta)$ converges uniformly in probability to a non-stochastic function $\theta(\beta)$, which Gourieroux, Monfort, and Renault (1993) term the *binding function*.

Loosely speaking, indirect inference generates an estimate $\hat{\beta}_n$ of the structural parameters by choosing β so as to make $\hat{\theta}_n$ and $\overline{\theta}_n(\beta)$ as close as possible, with consistency following from $\hat{\theta}_n$ and $\overline{\theta}_n(\beta_0)$ both converging to the same pseudo-true value $\theta_0 \coloneqq \theta(\beta_0)$. To implement the estimator we require a formal metric of the distance between $\hat{\theta}_n$ and $\overline{\theta}_n(\beta)$. There are three approaches to choosing such a metric, analogous to the three classical approaches to hypothesis testing: the Wald, likelihood ratio (LR), and Lagrange multiplier (LM) approaches.[2]

The Wald approach to indirect inference chooses β to minimize the weighted distance between $\overline{\theta}_n(\beta)$ and $\hat{\theta}_n$,

$$Q_n^{\mathrm{W}}(\beta) \coloneqq \|\overline{\theta}_n(\beta) - \hat{\theta}_n\|_{W_n}^2,$$

where $\|x\|_A^2 \coloneqq x^{\mathsf{T}} A x$, and W_n is a sequence of positive-definite weight matrices.

The LR approach forms a metric implicitly by using the average log-likelihood $\mathcal{L}_n(y, x; \theta)$ associated with the auxiliary model. In particular, it seeks to minimize

$$Q_n^{\mathrm{LR}}(\beta) \coloneqq -\mathcal{L}_n(y, x; \overline{\theta}_n(\beta)) = -\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i; \overline{\theta}_n(\beta))$$

Finally, the LM approach does not work directly with the estimated auxiliary parameters $\overline{\theta}_n(\beta)$ but instead uses the score vector associated with the auxiliary model.[3] Given the estimated auxiliary model parameters $\hat{\theta}$ from the observed data, the score vector is evaluated using each of the M simulated data sets. The LM estimator then minimizes a weighted norm of the average

[2]This nomenclature is due to Eric Renault. The Wald and LR approaches were first proposed in Smith (1990, 1993) and later extended by Gourieroux, Monfort, and Renault (1993). The LM approach was first proposed in Gallant and Tauchen (1996).

[3]When the LM approach is implemented using an auxiliary model that is (nearly) correctly specified in the sense that it provides a (nearly) correct statistical description of the observed data, Gallant and Tauchen (1996) refer to this approach as efficient method of moments (EMM).

score vector across these datasets,

$$Q_n^{\mathrm{LM}}(\beta) \coloneqq \left\| \frac{1}{M} \sum_{m=1}^{M} \dot{\mathcal{L}}_n(y^m(\beta), x; \hat{\theta}_n) \right\|_{V_n}^2,$$

where $\dot{\mathcal{L}}_n$ denotes the gradient of $\mathcal{L}_n$ with respect to θ, and V_n is a sequence of positive-definite weight matrices.

All three approaches yield consistent and asymptotically normal estimates of β_0, and are first-order asymptotically equivalent in the exactly identified case in which $d_\beta = d_\theta$. In the over-identified case, when the weight matrices W_n and V_n are chosen optimally (in the sense of minimizing asymptotic variance) both the Wald and LM estimators are more efficient than the LR estimator. However, if the auxiliary model is correctly specified, all three estimators are asymptotically equivalent not only to each other but also to maximum likelihood (provided that M is sufficiently large).

3.2 Indirect inference for discrete choice models

Step functions arise naturally when applying indirect inference to discrete choice models because any simulated choice $y_{it}^m(\beta)$ is a step function of β (holding fixed the set of random draws $\{\eta_{it}^m\}$ used to generate simulated data from the structural model). Consequently, the sample binding function $\overline{\theta}_n(\beta)$ is discontinuous in β. Obviously, this discontinuity is inherited by the criterion functions minimized by the II estimators in Section 3.1.

Thus, given discrete outcomes, II cannot be implemented using gradient-based optimization methods. One must instead rely on derivative-free methods (such as the Nelder-Mead simplex method); random search algorithms (such as simulated annealing); or abandon optimization altogether, and instead implement a Laplace-type estimator, via Markov Chain Monte Carlo (MCMC; see Chernozhukov and Hong, 2003). But convergence of derivative-free methods is often very slow; while MCMC, even when it converges, may produce (in finite samples) an estimator substantially different from the optimum of the statistical criterion to which it is applied (see Kormiltsina and Nekipelov, 2012). Thus, the non-smoothness of the criterion functions that define II estimators render them very difficult to use in the case of discrete data.

Despite the difficulties in applying II to discrete choice models, the appeal of the II approach has led some authors to push ahead and apply it nonetheless. Some notable papers that apply II by optimizing non-smooth objective functions are Magnac, Robin, and Visser (1995), An and Liu (2000), Nagypál (2007), Eisenhauer, Heckman, and Mosso (2015), Li and Zhang (2015) and Skira (2015). Our work aims to make it much easier to apply II in these and related contexts.

3.3 A smoothed estimator (GII)

Here we propose a generalization of indirect inference that is far more practical in the context of discrete outcomes. The fundamental idea is that the estimation procedures applied to the observed and simulated data sets need not be identical, provided that they both provide consistent estimates of the same binding function. (Genton and Ronchetti, 2003, use a similar insight to develop robust estimation procedures in the context of indirect inference.) We exploit this idea

to smooth the function $\overline{\theta}_n(\beta)$, obviating the need to optimize a step function when using indirect inference to estimate a discrete choice model.

Let $u^m_{itj}(\beta)$ denote the latent utility that individual i attaches to alternative $j \in \{1,\ldots,J-1\}$ in period t of the mth simulated data set, given structural parameters β (recall that the utility of the Jth alternative is normalized to 0). Rather than use the simulated choice $y^m_{itj}(\beta)$ when computing $\overline{\theta}_n(\beta)$, we propose to replace it by the following smooth function of the latent utilities,

$$y^m_{itj}(\beta,\lambda) \coloneqq K_\lambda[u^m_{itj}(\beta) - u^m_{it1}(\beta),\ \ldots,\ u^m_{itj}(\beta) - u^m_{it,J-1}(\beta)],$$

where $K : \mathbb{R}^{J-1} \to \mathbb{R}$ is a smooth, mean-zero multivariate cdf, and $K_\lambda(v) \coloneqq K(\lambda^{-1}v)$. As the smoothing parameter λ goes to 0, the preceding converges to $y^m_{itj}(\beta,0) = y^m_{itj}(\beta)$. Defining $\overline{\theta}_n(\beta,\lambda) \coloneqq \frac{1}{M}\sum_{m=1}^M \hat{\theta}^m_n(\beta,\lambda)$, where

$$\hat{\theta}^m_n(\beta,\lambda) \coloneqq \operatorname*{argmax}_{\theta\in\Theta} \mathcal{L}_n(y^m(\beta,\lambda),x;\theta), \tag{3.3}$$

we may regard $\overline{\theta}_n(\beta,\lambda)$ as a smoothed estimate of $\theta(\beta)$, for which it is consistent so long as $\lambda = \lambda_n \to 0$ as $n \to \infty$. Accordingly, an indirect inference estimator based on $\overline{\theta}_n(\beta,\lambda_n)$, which we shall henceforth term the *generalized indirect inference* (GII) estimator, ought to be consistent for β_0.

Each of the three approaches to indirect inference can be generalized simply by replacing each simulated choice $y^m_{itj}(\beta)$ with its smoothed counterpart $y^m_{itj}(\beta,\lambda_n)$. For the Wald and LR estimators, this entails using the smoothed sample binding function $\overline{\theta}_n(\beta,\lambda_n)$ in place of the unsmoothed estimate $\overline{\theta}_n(\beta)$. (See Section 4.2 below for the exact forms of the criterion functions.) The remainder of this paper is devoted to studying the properties of the resulting estimators, both analytically (Section 5) and through a series of simulation exercises (Section 6).

The GII approach was first suggested in an unpublished manuscript by Keane and Smith (2003), but they did not derive the asymptotic properties of the estimator. Despite this, GII has proven to be popular in practice, and has already been applied in a number of papers, such as Gan and Gong (2007), Cassidy (2012), Altonji, Smith, and Vidangos (2013), Morten (2013), Ypma (2013), Lopez-Mayan (2014) and Lopez Garcia (2015). Given the growing popularity of the method, a careful analysis of its asymptotic properties is obviously needed.

3.4 Related literature

Our approach to smoothing in a discrete choice model bears a superficial resemblance to that used by Horowitz (1992) to develop a smoothed version of Manski's (1985) maximum score estimator for a binary response model. As here, the smooth version of maximum score is constructed by replacing discontinuous indicators with smooth cdfs in the sample criterion function.

However, there is a fundamental difference in the statistical properties of the minimization problems solved by Manski's estimator, and the (unsmoothed) indirect inference estimator. Specifically, $n^{-1/2}$-consistent estimators are available for the *unsmoothed* problem considered in this paper (see Theorem 5.1 below, or Pakes and Pollard, 1989); whereas, in the case of Manski's (1985) maximum score estimator, only $n^{-1/3}$-consistency is obtained without smoothing (see Kim and Pollard, 1990), and smoothing yields an estimator with an improved rate of convergence.

A potentially more relevant analogue for the present paper is smoothed quantile regression. This originates with Horowitz's (1998) work on the smoothed least absolute deviation estimator, extended to more general quantile regression and quantile-IV models by Whang (2006), Otsu (2008) and Kaplan and Sun (2012). The latter papers do not smooth the criterion function, but rather the estimating equations (approximate first-order conditions) that equivalently define the estimator. These first-order conditions involve indicator-type discontinuities like those in our problem, smoothed in the same way. Insofar as the problem of solving the estimating equations is analogous to the minimum-distance problem solved by the II estimator, the effects of smoothing are similar: in each case smoothing (if done appropriately) affects neither the rate of convergence nor the limiting distribution of the estimator, relative to its unsmoothed counterpart.

The motivation for smoothing in the quantile regression case involves the potential for higher-order asymptotic improvements.[4] In contrast, in the present setting, which involves structural models of possibly great complexity, the potential for higher-order improvements is limited.[5] The key motivation for smoothing in our case is computational.

Accordingly, much of this paper is devoted to a formal analysis of the potential computational gains from smoothing. In particular, Sections 5.4–5.6 are devoted to providing a theoretical foundation for our claim that smoothing facilitates the convergence of standard derivative-based optimization that are widely used to solve (smooth) optimization problems in practice.

For the class of models considered in this paper, two leading alternative estimation methods that might be conisidered are simulated maximum likelihood (SML) in conjunction with the Geweke, Hajivassiliou and Keane (GHK) smooth probability simulator (see Section 4 in Geweke and Keane, 2001), and the nonparametric simulated maximum likelihood (NPSML) estimator (Diggle and Gratton, 1984; Fermanian and Salanié, 2004; Kristensen and Shin, 2012). However, the GHK simulator can only be computed in models possessing a special structure – which is true for Models 1, 2 and 4 above, but *not* for Model 3 – while in models that involve a mixture of discrete and continuous outcomes, NPSML may require the calculation of rather high-dimensional kernel density estimates in order to construct the likelihood, the accuracy of which may require simulating the model a prohibitively large number of times.

Finally, an alternative approach to smoothing the II estimator is importance sampling, as in Keane and Sauer (2010) and Sauer and Taber (2013). The basic idea is to simulate data from the structural model only once (at the initial estimate of β). One holds these simulated data fixed as one iterates. Given an updated estimate of β, one re-weights the original simulated data points, so those initial simulations that are more (less) likely under the new β (than under the initial β) get more (less) weight in forming the updated objective function.

In our view the GII and importance sampling approaches both have virtues. The main limitation of the importance sampling approach is that in many models the importance sample weights may themselves be computationally difficult to construct. Keane and Sauer (2010),

[4]While potential computational benefits have been noted in passing, we are not aware of any attempt to demonstrate these formally, in the manner of Theorems 5.3–5.5 below.

[5]This is particularly evident when the auxiliary model consists of a system of regression equations, as per Section 4.4 below. For while smoothing does indeed reduce the variability of the simulated (discrete) outcomes $y_{it}^m(\beta, \lambda)$, this may *increase* the variance with which some parameters of the auxiliary model are estimated, if y_{it} appears as a regressor in that model: as will be the case for Models 2 and 3 (see Sections 6.2 and 6.3 below). (Note that any such increase, while certainly possible, is of only second-order importance, and disappears as $\lambda_n \to 0$.)

when working with models similar to those in Section 2, assume that all variables are measured with error, which gives a structure that impies very simple weights. In many contexts such a measurement error assumption may be perfectly sensible. But the GII method can be applied directly to the models of Section 2 without adding any auxillary assumptions (or parameters).

4 Further refinements and the choice of auxiliary model

4.1 Smoothing in dynamic models

For models in which latent utilities depend on past choices (as distinct from past *utilities*, which are already smooth), such as Models 2 and 3 above, the performance of GII may be improved by making a further adjustment to the smoothing proposed in Section 3.3. The nature of this adjustment is best illustrated in terms of the example provided by Model 2. In this case, it is clear that setting

$$y_{it}^m(\beta,\lambda) \coloneqq K_\lambda[b_1 x_{it} + b_2 y_{i,t-1}^m(\beta) + \epsilon_{it}^m],$$

where $y_{i,t-1}^m(\beta)$ denotes the *unsmoothed* choice made at date $t-1$, will yield unsatisfactory results, insofar as the $y_{it}^m(\beta,\lambda)$ so constructed will remain discontinuous in β. To some extent, this may be remedied by modifying the preceding to

$$y_{it}^m(\beta,\lambda) \coloneqq K_\lambda[b_1 x_{it} + b_2 y_{i,t-1}^m(\beta,\lambda) + \epsilon_{it}^m], \tag{4.1}$$

with $y_{i0}^m(\beta,\lambda) \coloneqq 0$, as per the specification of the model. However, while the $y_{it}^m(\beta,\lambda)$'s generated through this recursion will indeed be smooth (i.e., twice continuously differentiable), the nesting of successive approximations entailed by (4.1) implies that for large t, the derivatives of $y_{it}^m(\beta,\lambda)$ may be highly irregular unless a relatively large value of λ is employed.

This problem may be avoided by instead computing $y_{it}^m(\beta,\lambda)$ as follows. Defining $v_{itk}^m(\beta) \coloneqq b_1 x_{it} + b_2 \mathbf{1}\{k=1\} + \epsilon_{it}^m$, we see that the *unsmoothed* choices satisfy

$$y_{it}(\beta) = \mathbf{1}\{v_{it0}^m(\beta) \geq 0\} \cdot [1 - y_{i,t-1}(\beta)] + \mathbf{1}\{v_{it1}^m(\beta) \geq 0\} \cdot y_{i,t-1}(\beta),$$

which suggests using the following recursion for the smoothed choices,

$$y_{it}^m(\beta,\lambda) \coloneqq K_\lambda[v_{it0}^m(\beta)] \cdot [1 - y_{i,t-1}^m(\beta,\lambda)] + K_\lambda[v_{it1}^m(\beta)] \cdot y_{i,t-1}^m(\beta,\lambda), \tag{4.2}$$

with $y_{i0}^m(\beta,\lambda) \coloneqq 0$. This indeed yields a valid approximation to $y_{it}(\beta)$, as $\lambda \to 0$. The smoothed choices computed using (4.2) involve no nested approximations, but merely sums of products involving terms of the form $K_\lambda[v_{isk}^m(\beta)]$. The derivatives of these are well-behaved with respect to λ, even for large t, and are amenable to the theoretical analysis of Section 5.

Nonetheless, we find that even if smoothing is done by simply using (4.1), GII appears to work well in practice. This will be shown in the simulation exercises reported in Section 6.

4.2 Bias reduction via jackknifing

As we noted in Section 3.3, GII inherits the consistency of the II estimator, provided that $\lambda_n \to 0$ as $n \to \infty$. However, as smoothing necessarily imparts a bias to the sample binding function $\overline{\theta}_n(\beta, \lambda_n)$, and thence to the GII estimator, we need λ_n to shrink to zero at a sufficiently fast rate if GII is to enjoy the same limiting distribution as the unsmoothed estimator. On the other hand, if $\lambda_n \to 0$ too rapidly, derivatives of the GII criterion function will become highly irregular, impeding the ability of derivative-based optimization routines to locate the minimum.

Except for certain special cases, the smoothing bias is of the order $\|\theta(\beta_0, \lambda) - \theta(\beta_0, 0)\| = O(\lambda)$ and no smaller. Thus, it is only dominated by the estimator variance if $n^{1/2}\lambda_n = o_p(1)$. On the other hand, it follows from Proposition 5.1 below that $n^{1-1/p_0}\lambda_n^{2l-1} \to \infty$ is necessary to ensure that the lth order derivatives of the GII criterion function converge, uniformly in probability, to their population counterparts. Here $p_0 \in (1, \infty]$ depends largely on the order of moments possessed by the exogenous covariates x (see Assumption L below). Thus, even in the most favorable case of $p_0 = \infty$, one can only ensure asymptotic negligibility of the bias (relative to the variance) at the cost of preventing second derivatives of the sample criterion function from converging to their population counterparts, a convergence that is necessary to ensure the good performance of at least some derivative-based optimization routines (see Section 5.6 below).

Fortunately, these difficulties can easily be overcome by applying Richardson extrapolation – commonly referred to as "jackknifing" in the statistics literature – to the smoothed sample binding function. Provided that the *population* binding function is sufficiently smooth, a Taylor series expansion gives $\theta_l(\beta, \lambda) = \theta_l(\beta, 0) + \sum_{r=1}^{s} \alpha_{rl}(\beta)\lambda^r + o(\lambda^s)$ as $\lambda \to 0$, for $l \in \{1, \ldots, d_\theta\}$. Then, for a fixed choice of $\delta \in (0, 1)$, we have the first-order extrapolation,

$$\theta_l^1(\beta, \lambda) := \frac{\theta_l(\beta, \delta\lambda) - \delta\theta_l(\beta, \lambda)}{1 - \delta} = \theta_l(\beta, 0) + \delta \sum_{r=2}^{s} (\delta^{r-1} - 1)\alpha_{rl}(\beta)\lambda^r + o(\lambda^s),$$

for every $l \in \{1, \ldots, d_\theta\}$. By an iterative process, for $k \leq s - 1$ we can construct a kth order extrapolation of the binding function, which satisfies

$$\theta^k(\beta, \lambda) := \sum_{r=0}^{k} \gamma_{rk}\theta(\beta, \delta^r\lambda) = \theta(\beta, 0) + O(\lambda^{k+1}), \tag{4.3}$$

where the weights $\{\gamma_{rk}\}_{r=0}^k$ (which can be negative) satisfy $\sum_{r=0}^k \gamma_{rk} = 1$, and may be calculated using Algorithm 1.3.1 in Sidi (2003). It is immediately apparent that the kth order jackknifed sample binding function,

$$\overline{\theta}_n^k(\beta, \lambda_n) := \sum_{r=0}^{k} \gamma_{rk}\overline{\theta}_n(\beta, \delta^r\lambda_n) \tag{4.4}$$

will enjoy an asymptotic bias of order $O_p(\lambda_n^{k+1})$, whence only $n^{1/2}\lambda_n^{k+1} = o_p(1)$ is necessary for the bias to be asymptotically negligible.

In the case where $\hat{\theta}_n^m(\beta, \lambda) = g(T_n^m(\beta, \lambda))$, for some differentiable transformation g of a vector T_n^m of sufficient statistics (as in Section 5 below), jackknifing could be applied directly to these statistics. Thus, if we were to set $\hat{\theta}_n^{mk}(\beta, \lambda_n) := g(\sum_{r=0}^k \gamma_{rk} T_n^m(\beta, \lambda_n))$, then $\frac{1}{M}\sum_{m=1}^M \hat{\theta}_n^{mk}(\beta, \lambda_n)$ would also have an asymptotic bias of order $O_p(\lambda_n^{k+1})$. This approach may have computational

advantages if the transformation g is relatively costly to compute (e.g. when it involves matrix inversion). Note that since T_n^m will generally involve averages of nonlinear transformations of kernel smoothers, it will not generally be possible to achieve the same bias reduction through the use of higher-order kernels; whereas if only linear transformations were involved, both jackknifing and higher-order kernels would yield identical estimators (see, e.g., Jones and Foster, 1993).

Jackknifed GII estimators of order $k \in \mathbb{N}_0$ may now be defined as the minimizers of:

$$Q_{nk}^e(\beta, \lambda_n) \coloneqq \begin{cases} \|\overline{\theta}_n^k(\beta, \lambda_n) - \hat{\theta}_n\|_{W_n}^2 & \text{if } e = \text{W} \\ -\mathcal{L}_n(y; x, \overline{\theta}_n^k(\beta, \lambda_n)) & \text{if } e = \text{LR} \\ \|\frac{1}{M}\sum_{m=1}^M \dot{\mathcal{L}}_n^{mk}(\beta, \lambda_n; \hat{\theta}_n)\|_{V_n}^2 & \text{if } e = \text{LM} \end{cases} \tag{4.5}$$

where $\dot{\mathcal{L}}_n^{mk}(\beta, \lambda; \hat{\theta}_n) \coloneqq \sum_{r=0}^k \gamma_{rk} \dot{\mathcal{L}}_n(y^m(\beta, \lambda), x; \hat{\theta}_n)$ denotes the jackknifed score function; the un-jackknifed estimators may be recovered by taking $k = 0$. Let $Q_k^e(\beta, \lambda)$ denote the large-sample limit of $Q_{nk}^e(\beta, \lambda)$; note that $\beta \mapsto Q_k^e(\beta, 0)$ is smooth and does not depend on k.

4.3 Bias reduction via a Newton-Raphson step

By allowing the number of simulations M to increase with the sample size, we can accelerate the rate at which $\overline{\theta}_n^k$ converges to the binding function. The convergence of the smoothed derivatives of $\overline{\theta}_n^k$ should then follow under less restrictive conditions on λ_n. That is, it may be possible for the derivatives to converge, while still ensuring that the bias is $o(n^{-1/2})$, even with $k = 0$. Since the evaluation of Q_{nk} is potentially costly when M is very large, one possible approach would be to minimize Q_{nk} using a very small initial value of M (e.g. $M = 1$). One could then increase M to an appropriately large value, and then compute a new estimate by taking at least one Newton-Raphson step (applied to the new criterion).

A rigorous analysis of this estimator is beyond the scope of this paper; we assume that M is *fixed* throughout Section 5. Heuristically, since $\overline{\theta}_n^k$ is computed using nM observations, it should be possible to show that if $M = M_n \to \infty$, then the conditions specified in Proposition 5.1 below would remain the same, except with nM_n replacing every appearance of n in (5.7).

4.4 Choosing an auxiliary model

Efficiency is a key consideration when choosing an auxiliary model. As discussed in Section 3.1, indirect inference (generalized or not) has the same asymptotic efficiency as maximum likelihood when the auxiliary model is correctly specified in the sense that it provides a correct statistical description of the observed data (Gallant and Tauchen, 1996). Thus, from the perspective of efficiency, it is important to choose an auxiliary model (or a class of auxiliary models) that is flexible enough to provide a good description of the data.

Another important consideration is computation time. For the Wald and LR approaches to indirect inference, the auxiliary parameters must be estimated repeatedly using different simulated data sets. For this reason, it is critical to use an auxiliary model that can be estimated quickly and efficiently. This consideration is less important for the LM approach, as it does not work directly with the estimated auxiliary parameters, but instead uses the first-order conditions (the score vector) that defines these estimates.

To meet the twin criteria of statistical and computational efficiency, in Section 6 we use linear probability models (or, more accurately, sets of linear probability models) as the auxillary model. This class of models is flexible in the sense that an individual's current choice can be allowed to depend on polynomial functions of lagged choices and of current and lagged exogenous variables. These models can also be very quickly and easily estimated using ordinary least squares. Section 6 describes in detail how we specify the linear probability models for each of Models 1–4. For Model 5, the Heckman selection model, the auxillary model would be a set of OLS regressions with mixed discrete/continuous dependent variables.

5 Asymptotic and computational properties

While GII could in principle be applied to any model of the form (2.1) – and others besides – in order to keep this paper to a manageable length, the theoretical results of this section will require that some further restrictions be placed on the structure of the model. Nonetheless, these restrictions are sufficiently weak to be consistent with each of Models 1–5 from Section 2. We shall only provide results for the Wald and LR estimators, when these are jackknifed as per (4.4) above; but it would be possible to extend our arguments so as to cover the LM estimator, and the alternative jackknifing procedure (in which the statistics T_n^m are jackknifed) outlined in Section 4.2.

5.1 A general framework

Individual i is described by vectors $x_i \in \mathbb{R}^{d_x}$ and $\eta_i \in \mathbb{R}^{d_\eta}$ of observable and unobservable characteristics; x_i includes *all* the covariates appearing in either the structural model or the auxiliary model (or both). η_i is a vector of independent variates that are also independent of x_i, and normalized to have unit variance. Their marginal distributions are fully specified by the model, allowing these to be simulated. Collect $z_i \coloneqq (x_i^\mathsf{T}, \eta_i^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{d_z}$, and define the projections $[x(\cdot), \eta(\cdot)]$ so that $(x_i, \eta_i) = [x(z_i), \eta(z_i)]$. Individual i has a vector $y(z_i; \beta, \lambda) \in \mathbb{R}^{d_y}$ of smoothed outcomes, parametrized by $(\beta, \lambda) \in \mathrm{B} \times \Lambda$, with $\lambda = 0$ corresponding to true, unsmoothed outcomes under β. At this level of abstraction, we need not make any notational distinction between choices made by an individual at the same date (over competing alternatives), vs. choices made at distinct dates; we note simply that each corresponds to some element of $y(\cdot)$. With this notation, the mth simulated choices may be written as $y(z_i^m; \beta, \lambda)$; since the same x_i's are used across all simulations, we have $x(z_i^m) = x(z_i^{m'})$ but $\eta(z_i^m) \neq \eta(z_i^{m'})$ for $m' \neq m$.

In line with the discussion in Section 4.4, we shall assume that the auxiliary model takes the form of a system of seemingly unrelated regressions (SUR; see e.g. Section 10.2 in Greene, 2008)

$$y_r(z_i; \beta, \lambda) = \alpha_{xr}^\mathsf{T} \Pi_{xr} x(z_i) + \alpha_{yr}^\mathsf{T} \Pi_{yr} y(z_i; \beta, \lambda) + \xi_{ri}, \tag{5.1}$$

where $\xi_i \coloneqq (\xi_{1i}, \ldots, \xi_{d_y i})^\mathsf{T} \sim_{\text{i.i.d.}} N[0, \Sigma_\xi]$, and Π_{xr} and Π_{yr} are selection matrices (i.e. matrices that take at most one unit value along each row, and have zeros everywhere else); let $\alpha_r \coloneqq (\alpha_{xr}^\mathsf{T}, \alpha_{yr}^\mathsf{T})^\mathsf{T}$. Typically, Σ_ξ will be assumed block diagonal: for example, we may only allow those ξ_{ri}'s pertaining to alternatives from the same period to be correlated. The auxiliary parameter vector θ collects a subset (or possibly all) of the elements of $(\alpha_1^\mathsf{T}, \ldots, \alpha_{d_y}^\mathsf{T})^\mathsf{T}$ and the (unrestricted)

elements of Σ_ξ^{-1}. (For the calculations involving the score vector in Appendix C, it shall be more convenient to treat the model as being parametrized in terms of Σ_ξ^{-1}.)

Several estimators of θ are available, most notably OLS, feasible GLS, and maximum likelihood, all of which agree only under certain conditions.[6] For concreteness, we shall assume that both the data-based and simulation-based estimates of θ are produced by maximum likelihood. However, the results of this paper could be easily extended to cover the case where either (or both) of these estimates are computed using OLS or feasible GLS. (In those cases, the auxiliary estimator can be still be written as a function of a vector of sufficient statistics, a property that greatly facilitates the proofs of our results.)

We shall also need to restrict the manner in which $y(\cdot)$ is parametrized. To that end, we introduce the following collections of linear indices

$$\nu_r(z;\beta) \coloneqq z^\mathsf{T}\Pi_{\nu r}\gamma(\beta) \qquad r \in \{1,\dots,d_\nu\} \tag{5.2a}$$

$$\omega_r(z;\beta) \coloneqq z^\mathsf{T}\Pi_{\omega r}\gamma(\beta) \qquad r \in \{1,\dots,d_\omega\}, \tag{5.2b}$$

where $\gamma : \mathrm{B} \to \Gamma$, and Π_l^ν and Π_l^ω are selection matrices. We shall generally suppress the z argument from ν and ω, and other quantities constructed from them, throughout the sequel. Our principal restriction on $y(\cdot)$ is that it should be constructed from (ν,ω) as follows. Let $d_c \geq d_\omega$; for each $r \in \{1,\dots,d_c\}$, let $\mathcal{S}_r \subseteq \{1,\dots,d_v\}$ and define

$$\tilde{y}_r(\beta,\lambda) \coloneqq \omega_r(\beta) \cdot \prod_{s\in\mathcal{S}_r} K_\lambda[\nu_s(\beta)] \tag{5.3}$$

collecting these in the vector $\tilde{y}(\beta,\lambda)$; where now $K : \mathbb{R} \to [0,1]$ is a smooth *univariate* cdf, and $K_\lambda(v) \coloneqq K(\lambda^{-1}v)$.[7] Note that $d_c \geq d_\omega$, and that we have defined

$$\omega_r(z;\beta) \coloneqq 1 \qquad r \in \{d_\omega+1,\dots,d_c\}. \tag{5.4}$$

Let $\eta_\omega \coloneqq \Pi_{\eta\omega}\eta$ select the elements of η upon which ω actually depends (as determined by the $\Pi_{\omega r}$ matrices), and let $W_r \geq 1$ denote an envelope for ω_r, in the sense that $|\omega_r(z;\beta)| \leq W_r(z)$ for all $\beta \in \mathrm{B}$. Let $\varrho_{\min}(A)$ denote the smallest eigenvalue of a symmetric matrix A.

Our results rely on the following low-level assumptions on the structural model:

Assumption L (low-level conditions)**.**

L1 *(y_i, x_i) is i.i.d. over i, and $\eta_i^m = \eta(z_i^m)$ is independent of x_i and i.i.d. over i and m;*

L2 *$y(\beta,\lambda) = D\tilde{y}(\beta,\lambda)$ for some $D \in \mathbb{R}^{d_y \times d_c}$, for $\tilde{y}$ as in (5.3);*

L3 *$\gamma : \mathrm{B} \to \Gamma$ in (5.2) is twice continuously differentiable;*

[6] In Section 6, exact numerical agreement between these estimators is ensured by requiring the auxiliary model equations referring to alternatives from the same period to have the same set of regressors.

[7] Keane and Smith (2003) suggested using the multivariate logistic cdf, $L(v) \coloneqq 1/(1+\sum_{j=1}^{J-1} \mathrm{e}^{-v_j})$, and this is used in the simulation exercises presented in Section 6. But L has no particular advantages over other choices of K, and, for the theoretical results work we shall in fact assume that the smoothing is implemented using suitable products of univariate cdfs. This assumption eases some of our arguments (but it is unlikely that it is necessary for our results).

L4 *for each* $k \in \{1, \dots, d_\eta\}$, $\operatorname{var}(\eta_{ki}) = 1$, *and* η_{ki} *has a density* f_k *with*

$$\sup_{u \in \mathbb{R}} (1 + |u|^4) f_k(u) < \infty;$$

L5 *there exists an* $\epsilon > 0$ *such that, for every for every* $r \in \{1, \dots, d_\nu\}$ *and* $\beta \in \mathrm{B}$,

$$\operatorname{var}(\nu_r(z_i; \beta) \mid \eta_{\omega i}, x_i) \geq \epsilon;$$

L6 *there exists a* $p_0 \geq 2$ *such that for each* $r \in \{1, \dots, d_c\}$, $\mathbb{E}(W_r^4 + \|z_i\|^4) < \infty$, $\mathbb{E}|W_r \|z_i\|^3|^{p_0} < \infty$ *and* $\mathbb{E}|W_r^2 \|z_i\|^2|^{p_0} < \infty$;

L7 $\inf_{(\beta,\lambda) \in \mathrm{B} \times \Lambda} \varrho_{\min}[\mathbb{E}\overline{y}(z_i; \beta, \lambda)\overline{y}(z_i; \beta, \lambda)^\mathsf{T}] > 0$, *where* $\overline{y}(\beta, \lambda) \coloneqq [y(\beta, \lambda)^\mathsf{T}, x^\mathsf{T}]^\mathsf{T}$; *and*

L8 *the auxiliary model is a Gaussian SUR, as in (5.1).*

Remark 5.1. (5.2) entails that the estimator criterion function Q_n depends on β only through $\gamma(\beta)$, i.e. $Q_n(\beta) = \tilde{Q}_n(\gamma(\beta))$ for some $\tilde{Q}_n$. Since the derivatives of $\tilde{Q}_n$ with respect to γ take a reasonably simple form, we shall establish the convergence of $\partial_\beta^l Q_n$ to $\partial_\beta^l Q$, for $l \in \{1, 2\}$, by first proving the corresponding result for $\partial_\gamma^l \tilde{Q}_n$ and then applying the chain rule. Here, as elsewhere in the paper, $\partial_\beta f$ denotes the gradient of $f : \mathrm{B} \to \mathbb{R}^d$ (the transpose of the Jacobian), and $\partial_\beta^2 f$ the Hessian; see Section 6.3 of Magnus and Neudecker, 2007, for a definition of the latter when $k \geq 2$.

Remark 5.2. Assumption L is least restrictive in models with purely discrete outcomes, for which we may take $d_\omega = 0$. In particular, L6 reduces to the requirement that $\mathbb{E}\|z_i\|^{3p_0} < \infty$.

Remark 5.3. As the examples discussed in Section 5.2 illustrate, except in the case where current (discrete) choices depend on past choices, it is generally possible to take $D = I_{d_y}$ in L2, so that $y(\beta, \lambda) = \tilde{y}(\beta, \lambda)$.

Consistent with the notation adopted in the previous sections of this paper, let η^m denote the mth set of simulated unobservables, and $y^m(\beta, \lambda)$ the associated smoothed outcomes, for $m \in \{1, \dots, M\}$. We may set $\Lambda = [0, 1]$ below without loss of generality. Let $\mathcal{F}$ denote a σ-field with respect to which the observed data and simulated variables are measurable, for all n, and recall the definition of ℓ given in (3.1) above. We then have the following:

Assumption R (regularity conditions)**.**

R1 *The structural model is correctly specified:* $y_i = y(z_i^0; \beta_0, 0)$ *for some* $\beta_0 \in \operatorname{int} \mathrm{B}$;

R2 $\theta_0 \coloneqq \theta(\beta_0, 0) \in \operatorname{int} \Theta$;

R3 *the binding function* $\theta(\beta, \lambda)$ *is single-valued, and is* $(k_0 + 1)$*-times differentiable in* β *for all* $(\beta, \lambda) \in (\operatorname{int} \mathrm{B}) \times \Lambda$;

R4 $\beta \mapsto \theta(\beta, 0)$ *is injective;*

R5 $\{\lambda_n\}$ *is an* $\mathcal{F}$*-measurable sequence with* $\lambda_n \xrightarrow{p} 0$;

R6 *the order* $k \in \{1, \dots, k_0\}$ *of the jackknifing is chosen such that* $n^{1/2} \lambda_n^{k+1} = o_p(1)$;

R7 *K in (5.3) is a twice continuously differentiable cdf, for a distribution having integer moments of all orders, and density $\dot{K}$ symmetric about the origin; and*

R8 *$W_n \xrightarrow{p} W$, for some positive definite W.*

Remark 5.4. R4 formalizes the requirement that the auxiliary model be "sufficiently rich" to identify the parameters of the structural model; $d_\theta \geq d_\beta$ evidently is necessary for R4 to be satisfied.

Remark 5.5. R5 permits the bandwidth to be sample-dependent, as distinct from assuming it to be a "given" deterministic sequence. This means our results hold *uniformly* in smoothing parameter sequences satisfying certain growth rate conditions: see Remark 5.7 below for details. R6 ensures that, in conjunction with the choice of λ_n, the jackknifing is such as to ensure that the bias introduced by the smoothing is asymptotically negligible. R7 will be satisfied for many standard choices of K, such as the Gaussian cdf, and many smooth, compactly supported kernels.

Assumptions L and R are sufficient for all of our main results. But to allow these to be stated at a higher level of generality – and thus permitting their application to a broader class of structural and auxiliary models than are consistent with Assumption L – we shall find it useful to phrase our results as holding under Assumption R and the following high-level conditions. To state these, define $\mathcal{L}_n(\theta) \coloneqq \mathcal{L}_n(y, x; \theta)$, $\mathcal{L}(\theta) \coloneqq \mathbb{E}\mathcal{L}_n(\theta)$ and $\ell_i^m(\beta, \lambda; \theta) \coloneqq \ell(y_i^m(\beta, \lambda), x_i; \theta)$. $\dot{\ell}_i^m$ and $\ddot{\mathcal{L}}_n$ respectively denote the gradient of ℓ_i^m and the Hessian of $\mathcal{L}_n$ with respect to θ, while for a metric space (Q, d), $\ell^\infty(Q)$ denotes the space of bounded and measurable real-valued functions on Q.

Assumption H (high-level conditions)**.**

H1 *$\mathcal{L}_n$ is twice continuously differentiable on* $\operatorname{int}\Theta$*;*

H2 *for $l \in \{0, 1, 2\}$, $\partial_\theta^l \mathcal{L}_n(\theta) \xrightarrow{p} \partial_\theta^l \mathcal{L}(\theta)$, and*

$$\frac{1}{n}\sum_{i=1}^n \dot{\ell}_i^{m_1}(\beta_1, \lambda_1; \theta_1)\dot{\ell}_i^{m_2}(\beta_2, \lambda_2; \theta_2)^\mathsf{T} \xrightarrow{p} \mathbb{E}\dot{\ell}_i^{m_1}(\beta_1, \lambda_1; \theta_1)\dot{\ell}_i^{m_2}(\beta_2, \lambda_2; \theta_2)^\mathsf{T}$$

uniformly on $\mathrm{B} \times \Lambda$ *and compact subsets of* $\operatorname{int}\Theta$*, for every $m_1, m_2 \in \{0, 1, \ldots, M\}$;*

H3 *ψ^m is a mean-zero, continuous Gaussian process on* $\mathrm{B} \times \Lambda$ *such that*

$$\psi_n^m(\beta, \lambda) \coloneqq n^{1/2}[\hat{\theta}_n^m(\beta, \lambda) - \theta(\beta, \lambda)] \rightsquigarrow \psi^m(\beta, \lambda)$$

in $\ell^\infty(\mathrm{B} \times \Lambda)$, jointly in $m \in \{0, 1, \ldots, M\}$;

H4 *for any (possibly) random sequence $\beta_n = \beta_0 + o_p(1)$ and λ_n as in* R5*,*

$$\psi_n^m(\beta_n, \lambda_n) = -H^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^n \dot{\ell}_i^m(\beta_0, 0; \theta_0) + o_p(1) =: -H^{-1}\phi_n^m + o_p(1) \rightsquigarrow -H^{-1}\phi^m, \quad (5.5)$$

jointly in $m \in \{0, 1, \ldots, M\}$,[8] *where $H \coloneqq \mathbb{E}\ddot{\mathcal{L}}_n(\theta) = \ddot{\mathcal{L}}(\theta)$ and $\{\phi^m\}_{m=0}^M$ is jointly Gaussian*

[8]The first equality in (5.5) is only relevant for $m \geq 1$.

with

$$\Sigma := \mathbb{E}\phi^{m_1}\phi^{m_1\mathsf{T}} = \mathbb{E}\phi_n^{m_1}\phi_n^{m_1\mathsf{T}} \qquad \mathrm{R} := \mathbb{E}\phi^{m_1}\phi^{m_2\mathsf{T}} = \mathbb{E}\phi_n^{m_1}\phi_n^{m_2\mathsf{T}} \tag{5.6}$$

for every $m_1, m_2 \in \{0, 1, \ldots, M\}$*; and*

H5 $\{\lambda_n\}$ *is such that for some* $l \in \{0, 1, 2\}$*,*

$$\sup_{\beta\in\mathrm{B}} \|\partial_\beta^l \hat{\theta}_n^m(\beta, \lambda_n) - \partial_\beta^l \theta(\beta, 0)\| = o_p(1).$$

The sufficiency of our low-level conditions for the preceding may be stated formally follows.

Proposition 5.1. *Suppose Assumptions L and R hold. Then Assumption H holds with* $l = 0$ *in* H5*. Further, if* $\lambda_n > 0$ *for all* n*, with*

$$n^{1-1/p_0}\lambda_n^{2l'-1}/\log(\lambda_n^{-1} \vee n) \xrightarrow{p} \infty \tag{5.7}$$

for some $l' \in \{1, 2\}$*, then* H5 *holds with* $l = l'$*.*

Remark 5.6. It is evident from (5.7) that – as noted in Section 4.2 above – the convergence of the higher-order derivatives of the sample binding function requires more stringent conditions on the smoothing sequence $\{\lambda_n\}$. We shall be accordingly careful, in stating our results below, to identify the weakest form of H5 (and correspondingly, of (5.7)) that is required for each of these.

Remark 5.7. Let $\{\underline{\lambda}_n\}$ and $\{\overline{\lambda}_n\}$ be deterministic sequences satisfying (5.7) and $\overline{\lambda}_n = o(1)$ respectively, and set $\Lambda_n := [\underline{\lambda}_n, \overline{\lambda}_n]$. Then, as indicated in Remark 5.5 above, $\mathcal{F}$-measurability of $\{\lambda_n\}$ entails that the convergence in H5 holds uniformly over $\lambda \in \Lambda_n$, in the sense that

$$\sup_{(\beta,\lambda)\in\mathrm{B}\times\Lambda_n} \|\partial_\beta^l \hat{\theta}_n^m(\beta, \lambda) - \partial_\beta^l \theta(\beta, \lambda)\| = o_p(1).$$

A similar interpretation applies to Theorems 5.3–5.5 below.

Proposition 5.1 is proved in Appendix C.

5.2 Application to examples

We may verify that each of the models from Section 2 satisfy L2–L6. In all cases, x_i collects all the (unique) elements of $\{x_{it}\}_{t=1}^T$, together with any additional exogenous covariates used to estimate the auxiliary model; while η_i collects the elements of $\{\eta_{it}\}_{t=1}^T$. Note that for the discrete choice Models 1–4, since the η_i are Gaussian L6 will be satisfied if $\mathbb{E}\|x_i\|^{3p_0} < \infty$. L7 is a standard non-degeneracy condition.

Model 1. $u_{it} = bx_{it} + \sum_{s=1}^t r^{t-s}\eta_{is}$ by backward substitution. So we set $(d_\nu, d_\omega) = (T, 0)$, with

$$\nu_t(z_i; \beta) = x_t(z_i)b(\beta) + \sum_{s=1}^t \eta_s(z_i)d_{ts}(\beta),$$

where $\beta = (b, r)$, $b(\beta) = b$ and $d_{ts}(\beta) = r^{t-s}$; while $x_t(z_i)$ and $\eta_s(z_i)$ select the appropriate elements of z_i, which collects $\{x_{it}\}$, $\{\eta_{it}\}$, and any other exogenous covariates used in the auxiliary

model. Thus L2 and L3 hold (formally, take $\gamma(\beta) = (b(\beta), \{d_{ts}(\beta)\})$). L5 follows from the $\eta_t(z_i)$'s being standard Gaussian.

Model 2. As per the discussion in Section 4.1, and (4.2) in particular, we define

$$\nu_{tk}(z_i;\beta) \coloneqq x_t(z_i)b_1(\beta) + b_2(\beta)\mathbf{1}\{k=1\} + \sum_{s=1}^{t} \eta_s(z_i)d_{ts}(\beta)$$

where the right-hand side quantities are defined by analogy with the preceding example. Setting

$$y_t(\beta,\lambda) \coloneqq K_\lambda[\nu_{t0}(\beta)] \cdot [1 - y_{t-1}(\beta,\lambda)] + K_\lambda[\nu_{t1}(\beta)] \cdot y_{t-1}(\beta,\lambda) \tag{5.8}$$

with $y_0(\beta,\lambda) \coloneqq 0$ thus yields smoothed choices having the form required by L2 and L3, as may be easily verified by backwards substitution. L5 again follows from Gaussianity of $\eta_t(z_i)$.

An identical recursion to (5.8) also works for Model 3. Model 4 may be handled in a similar way to Model 1, but it is in certain respects simpler, because the errors are not serially dependent. Finally, it remains to consider:

Model 5. From the preceding examples, it is clear that $\omega(z_i;\beta) = w_i$ and $\nu(z_i;\beta) = u_i$ can be written in the linear index form (5.2). The observable outcomes are the individual's decision to work, and also his wage if he decides to work. These may be smoothly approximated by:

$$y_1(\beta,\lambda) \coloneqq K_\lambda[\nu(\beta)] \qquad\qquad y_2(\beta,\lambda) \coloneqq \omega(\beta) \cdot K_\lambda[\nu(\beta)].$$

respectively. Thus L2–L5 hold just as in the other models. L6 holds, in this case, if $\mathbb{E}\|z_i\|^{4p_0} < \infty$.

5.3 Limiting distributions of GII estimators

We now present our asymptotic results. Note that Assumptions R and H are maintained throughout the following (even if not explicitly referenced), though in accordance with Remark 5.6 above, we shall always explicitly state the order of l in H5 that is required for each of our theorems.

Our first result concerns the limiting distributions of the minimizers of the Wald and LR criterion functions, as displayed in (4.5) above. For $e \in \{\mathrm{W}, \mathrm{LR}\}$, let $\hat{\beta}^e_{nk}$ be a near-minimizer of Q^e_{nk}, in the sense that

$$Q^e_{nk}(\hat{\beta}^e_{nk}, \lambda_n) \leq \inf_{\beta \in \mathrm{B}} Q^e_{nk}(\beta, \lambda_n) + o_p(n^{-1}). \tag{5.9}$$

The limiting variance of both estimators will have the familiar sandwich form. To allow the next result to be stated succinctly, define

$$\Omega(U,V) \coloneqq (G^\mathsf{T}UG)^{-1}G^\mathsf{T}UH^{-1}VH^{-1}UG(G^\mathsf{T}UG)^{-1} \tag{5.10}$$

where $G \coloneqq [\partial_\beta \theta(\beta_0, 0)]^\mathsf{T}$ denotes the Jacobian of the binding function at $(\beta_0, 0)$, $H = \mathbb{E}\ddot{\mathcal{L}}_n(\theta)$, and U and V are symmetric matrices.

Theorem 5.1 (limiting distributions)**.** *Suppose* H5 *holds with* $l = 0$*. Then*

$$n^{1/2}(\hat{\beta}^e_{nk} - \beta_0) \rightsquigarrow N[0, \Omega(U_e, V_e)],$$

where

$$U_e \coloneqq \begin{cases} W & \text{if } e = \mathrm{W} \\ H & \text{if } e = \mathrm{LR} \end{cases} \qquad V_e \coloneqq \left(1 + \frac{1}{M}\right)(\Sigma - \mathrm{R}) \tag{5.11}$$

Remark 5.8. In view of Proposition 5.1 and the remark that follows it, Theorem 5.1 does *not* restrict the rate at which $\lambda_n \xrightarrow{p} 0$ from *below*; indeed, it continues to hold even if $\lambda_n = 0$ for all n, in which case the estimation problem is closely related to that considered by Pakes and Pollard (1989). Thus, while the theorem provides the "desired" limiting distribution for our estimators, it fails to provide a justification (or motivation) for the smoothing proposed in this paper, and is in this sense unsatisfactory (or incomplete).

Remark 5.9. Note that the order of jackknifing does not affect the limiting distribution of the estimator: this has only a second-order effect, which vanishes as $\lambda_n \to 0$.

Remark 5.10. It is possible to define the LR estimator as the minimizer of

$$Q_n^{\mathrm{LR}}(\beta) \coloneqq -\tilde{\mathcal{L}}_n(y; x, \overline{\theta}_n^k(\beta, \lambda_n)),$$

where the average log-likelihood $\tilde{\mathcal{L}}_n$ need not correspond to that maximized by $\hat{\theta}_n^m$, provided that the maximizers of both $\mathcal{L}_n$ and $\tilde{\mathcal{L}}_n$ are consistent for the same parameters. For example, $\hat{\theta}_n^m$ might be OLS (and the associated residual covariance estimators), whereas $\tilde{\mathcal{L}}_n$ is the average log-likelihood for a SUR model. Suppose that the maximizer $\tilde{\theta}_n$ of $\tilde{\mathcal{L}}_n$ satisfies the following analogue of (5.5),

$$n^{1/2}(\tilde{\theta}_n - \theta_0) = -\tilde{H}^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^{n} \dot{\tilde{\ell}}_i^m(\beta_0, 0; \theta_0) + o_p(1) \rightsquigarrow -\tilde{H}^{-1}\tilde{\phi}^0,$$

and define $\tilde{\Sigma} \coloneqq \mathbb{E}\tilde{\phi}^0\tilde{\phi}^{0\mathsf{T}}$, and $\tilde{\mathrm{R}} \coloneqq \mathbb{E}\tilde{\phi}^0\phi^{m\mathsf{T}}$ for $m \geq 1$. Then the conclusions of Theorem 5.1 continue to hold, except that the $H^{-1}VH^{-1}$ appearing in (5.10) must be replaced by

$$\tilde{H}^{-1}\tilde{\Sigma}\tilde{H}^{-1} - (\tilde{H}^{-1}\tilde{\mathrm{R}}H^{-1} + H^{-1}\tilde{\mathrm{R}}^{\mathsf{T}}\tilde{H}^{-1}) + H^{-1}\left[\frac{1}{M}\Sigma + \left(1 - \frac{1}{M}\right)\mathrm{R}\right]H^{-1} \tag{5.12}$$

and $U_{\mathrm{LR}} = \tilde{H}$, where $\tilde{H} \coloneqq \mathbb{E}\ddot{\tilde{\mathcal{L}}}_n(y, x; \theta_0)$. Regarding the estimation of these quantities, see Remark 5.11 below. (Note that (5.12) reduces to $H^{-1}VH^{-1}$ when $\tilde{H} = H$, $\tilde{\Sigma} = \Sigma$ and $\tilde{\mathrm{R}} = \mathrm{R}$.)

The proofs of Theorem 5.1 and all other theorems in this paper are given in Appendix B.

5.4 Convergence of smoothed derivatives and variance estimators

Theorem 5.1 fails to indicate the possible benefits of smoothing, because it simply posits the existence of a near-minimizer of Q_{nk}, and thus entirely ignores how such a minimizer might be computed in practice. Ideally, smoothing should be shown to facilitate the convergence of derivative-based optimization procedures, when these are applied to the problem of minimizing Q_{nk}, while still yielding an estimator having the same limit distribution as in Theorem 5.1.

For the analysis of these procedures, the large-sample behavior of the derivatives of Q_{nk}

will naturally play an important role.[9] The uniform convergence of the derivatives of the sample binding function – and hence those of Q_{nk} – follows immediately from H5, and sufficient conditions for this convergence are provided by Proposition 5.1 above. Notably, when $l' \in \{1, 2\}$, (5.7) imposes exactly the sort of lower bound on λ_n that is absent from Theorem 5.1.

H5, with $l = 1$, implies that the derivatives of the smoothed criterion function can be used to estimate the Jacobian matrix G that appears in the limiting variances in Theorem 5.1. The remaining components, H and V_e, can be respectively estimated using the data-based auxiliary log-likelihood Hessian, and an appropriate transformation of the joint sample variance of all the auxiliary log-likelihood scores (i.e. using both the data- and simulation-based models). Define

$$A^\mathsf{T} \coloneqq \begin{bmatrix} I_{d_\theta} & -\frac{1}{M} I_{d_\theta} & \cdots & -\frac{1}{M} I_{d_\theta} \end{bmatrix}$$
$$s_{ni}^\mathsf{T} \coloneqq \begin{bmatrix} \dot{\ell}_i^0(\hat{\theta}_n)^\mathsf{T} & \dot{\ell}_i^1(\hat{\beta}_{nk}^e, \lambda_n; \hat{\theta}_n^1)^\mathsf{T} & \cdots & \dot{\ell}_i^M(\hat{\beta}_{nk}^e, \lambda_n; \hat{\theta}_n^M)^\mathsf{T} \end{bmatrix},$$

where $\hat{\theta}_n^m \coloneqq \hat{\theta}_n^m(\hat{\beta}_{nk}^e, \lambda_n)$, and $\dot{\ell}_i^0(\theta)$ denotes the gradient of $\ell(y_i, x_i; \theta)$. Then we have

Theorem 5.2 (variance estimation)**.** *Suppose* H5 *holds with* $l = 0$*. Then*

(i) $\hat{H}_n \coloneqq \ddot{\mathcal{L}}_n(\hat{\theta}_n) \xrightarrow{p} H$*;*

(ii) $\hat{V}_n \coloneqq A^\mathsf{T} \left(\frac{1}{n} \sum_{i=1}^n s_{ni} s_{ni}^\mathsf{T} \right) A \xrightarrow{p} V$*; and*

if H5 *holds with* $l = 1$*, then*

(iii) $\hat{G}_n \coloneqq \partial_\beta \overline{\theta}_n(\hat{\beta}_{nk}^e, \lambda_n) \xrightarrow{p} G$*, for* $e \in \{\mathrm{W}, \mathrm{LR}\}$.

Remark 5.11. For the situation envisaged in Remark 5.10, so long as the auxiliary model corresponding to $\tilde{\mathcal{L}}_n$ satisfies Assumptions R and H, a consistent estimate of (5.12) can be produced in the manner of (ii) above, if we replace s_{ni} by

$$\tilde{s}_{ni}^\mathsf{T} \coloneqq \begin{bmatrix} \dot{\tilde{\ell}}_i^0(\tilde{\theta}_n)^\mathsf{T} \hat{H}_n^{-1} & \dot{\ell}_i^1(\hat{\beta}_{nk}^{\mathrm{LR}}, \lambda_n; \hat{\theta}_n^1)^\mathsf{T} \tilde{H}_n^{-1} & \cdots & \dot{\ell}_i^M(\hat{\beta}_{nk}^{\mathrm{LR}}, \lambda_n; \hat{\theta}_n^M)^\mathsf{T} \tilde{H}_n^{-1} \end{bmatrix},$$

where $\tilde{H}_n \coloneqq \ddot{\tilde{\mathcal{L}}}_n(\tilde{\theta}_n)$ is consistent for U_{LR}.

5.5 Performance of derivative-based optimization procedures

The potential gains from smoothing may be assessed by comparing the performance of derivative-based optimization procedures, as they are applied to each of the following:

P1 the smoothed sample problem, of minimizing $\beta \mapsto Q_{nk}(\beta, \lambda_n)$; and

P2 its population counterpart, of minimizing $\beta \mapsto Q_k(\beta, 0)$.

Since Q_k is automatically smooth (even when $\lambda = 0$, owing to the smoothing effected by the expectation operator), derivative-based methods ought to be particularly suited to solving P2, and we may regard their performance when applied to this problem as representing an upper bound for their performance when applied to P1.

[9] Here, as throughout the remainder of this paper, we are concerned exclusively with the limiting behavior of the *exact* derivatives of Q_{nk}, ignoring any errors that might be introduced by numerical differentiation.

In the following section, we shall discuss in detail the convergence properties of three popular optimization routines: Gauss-Newton; quasi-Newton with BFGS updating; and a trust-region method. But before coming to these, we first provide a result that is of relevance to a broader range of derivative-based optimization procedures. Since such procedures will typically be designed to terminate at (near) roots of the first-order conditions,

$$\partial_\beta Q^e_{nk}(\beta, \lambda_n) = 0 \text{ in P1} \qquad\qquad \partial_\beta Q^e(\beta, 0) = 0 \text{ in P2}$$

for $e \in \{\mathrm{W}, \mathrm{LR}\}$, we shall provide conditions on λ_n, under which, for some $c_n = o_p(1)$,

(i) the set $R^e_{nk} \coloneqq \{\beta \in \mathrm{B} \mid \|\partial_\beta Q^e_{nk}(\beta, \lambda_n)\| \leq c_n\}$ of near roots is "consistent" for subsets of $R^e \coloneqq \{\beta \in \mathrm{B} \mid \partial_\beta Q^e(\beta, 0) = 0\}$; and

(ii) if $c_n = o_p(n^{-1/2})$, then any $\tilde{\beta}_n \in R^e_{nk}$ with $\tilde{\beta}_n \xrightarrow{p} \beta_0$ has the limiting distribution given by Theorem 5.1.

We interpret (i) as saying that smoothing yields a sample problem P1 that is "no more difficult" than the population problem P2, in the sense that the set of points to which derivative-based optimizers may converge to in P1 approximates its counterpart in P2, as $n \to \infty$. This is the strongest consistency result we can hope to prove here: as Q may have multiple stationary points, only one of which coincides with its (assumed interior) global minimum, it cannot generally be true that the whole of R^e_{nk} will be consistent for β_0. On the other hand, if we can select a consistent sequence of (near) roots from R^e_{nk}, as in (ii), then we may reasonably hope that this estimator sequence will enjoy the same limiting distribution as a (near) minimizer of Q_{nk}.

For $A, B \subseteq \mathrm{B}$, let $d_L(A, B) \coloneqq \sup_{a \in A} d(a, B)$ denote the one-sided distance from A to B, which has the property that $d_L(A, B) = 0$ if and only if $A \subseteq B$. Recall the definition of $\hat{\beta}^e_{nk}$ given in (5.9) above. Properties (i) and (ii) above can be more formally expressed as follows.

Theorem 5.3 (near roots)**.** *Suppose* H5 *holds with* $l = 1$*. Then*

(i) R^e_{nk} *is nonempty w.p.a.1., and* $d_L(R^e_{nk}, R^e) \xrightarrow{p} 0$*;*

(ii) *if* $c_n = o_p(n^{-1/2})$*,* $\tilde{\beta}_n \in R^e_{nk}$ *and* $\tilde{\beta}_n \xrightarrow{p} \beta_0$*, then* $n^{1/2}(\tilde{\beta}_n - \hat{\beta}^e_{nk}) = o_p(1)$*, and so* $\tilde{\beta}_n$ *has the limiting distribution given by Theorem 5.1; and*

(iii) *any* $\tilde{\beta}_n \in R^e_{nk}$ *satisfying* $Q^e_{nk}(\tilde{\beta}_n) \leq \inf_{\beta \in R^e_{nk}} Q^e_{nk}(\beta) + o_p(1)$ *has* $\tilde{\beta}_n \xrightarrow{p} \beta_0$*.*

Remark 5.12. Of course, the requirement that $\tilde{\beta}_n \xrightarrow{p} \beta_0$ cannot be verified in practice; but one may hope to satisfy it by running the optimization routine from L different starting points located throughout B, obtaining a collection of terminal values $\{\beta_{nl}\}_{l=1}^L$, and then setting $\tilde{\beta}_n = \beta_{nl}$ such that $Q_{nk}(\tilde{\beta}_n) \leq Q_{nk}(\beta_{nl'})$ for all $l' \in \{1, \ldots, L\}$.

Some optimization routines, such as the trust-region method considered in the next section, may only be allowed to terminate when the second-order conditions for a minimum are also satisfied. Defining

$$S^e_{nk} \coloneqq \{\beta \in R^e_{nk} \mid \varrho_{\min}[\partial^2_\beta Q^e_{nk}(\beta, \lambda_n)] \geq 0\} \qquad S^e \coloneqq \{\beta \in R^e \mid \varrho_{\min}[\partial^2_\beta Q^e(\beta, 0)] \geq 0\}, \tag{5.13}$$

we have the following

Theorem 5.4 (near roots satisfying second-order conditions)**.** *Suppose* H5 *holds with* $l = 2$*. Then parts* (i) *and* (ii) *of Theorem 5.3 hold with* S^e_{nk} *and* S^e *in place of* R^e_{nk} *and* R^e *respectively.*

Remark 5.13. The utility of this result may be seen by considering a case in which Q has many stationary points, but only a single local minimum at β_0. Then while Theorem 5.3 only guarantees convergence to one of these stationary points, Theorem 5.4 ensures consistency for β_0 – at a cost of requiring that the routine also check the second-order conditions for a minimum. This is why stronger conditions must be imposed on λ_n in Theorem 5.4; we now need the *second* derivatives of Q_{nk} to provide reliable information about the curvature of Q_k in large samples.

5.6 Convergence results for specific procedures

Our final result concerns the question of whether certain optimization routines, if initialized from within an appropriate region of the parameter space and iterated to convergence, will yield the maximizer of Q_{nk}, and thus an estimator having the limiting distribution displayed in Theorem 5.1. In some respects, our work here is related to previous work on k-step estimators, which studies the limiting behavior of estimators computed as the outcome of a sequence of quasi-Newton iterations (see e.g. Robinson, 1988). However, we shall depart from that literature in an important respect, by *not* requiring that our optimization routines be initialized by a sequence of starting values $\beta_n^{(0)}$ that are assumed consistent for β_0 (often at some rate). Rather, we shall require only that $\beta_n^{(0)} \in \mathrm{B}_0 \subset \mathrm{B}$ for a *fixed* region B_0 satisfying the conditions noted below.

We consider two popular line-search optimization methods – Gauss-Newton, and quasi-Newton with BFGS updating – as well as a trust-region algorithm. When applied to the problem of minimizing an objective Q, each of these routines proceed as follows: given an iterate $\beta^{(s)}$, locally approximate Q by the following quadratic model,

$$f_{(s)}(\beta) \coloneqq Q(\beta^{(s)}) + \nabla_{(s)}^{\mathsf{T}}(\beta - \beta^{(s)}) + \tfrac{1}{2}(\beta - \beta^{(s)})^{\mathsf{T}}\Delta_{(s)}(\beta - \beta^{(s)}), \tag{5.14}$$

where $\nabla_{(s)} \coloneqq \partial_\beta Q(\beta^{(s)})$. A new iterate $\beta^{(s+1)}$ is then generated by approximately minimizing $f_{(s)}$ with respect to β. The main differences between these procedures concern the choice of approximate Hessian $\Delta_{(s)}$, and the manner in which $f_{(s)}$ is (approximately) minimized. A complete specification of each of the methods considered here is provided in Appendix A (see also Fletcher, 1987, and Nocedal and Wright, 2006); note that the Gauss-Newton method can only be applied to the Wald criterion function, since only this criterion has the least-squares form required by that method.

We shall impose the following conditions on the population criterion Q, which are sufficient to ensure that each of these procedures, once started from some $\beta^{(0)} \in \mathrm{B}_0$, will converge to the global minimizer of Q. As noted above, since Q may have many other stationary points, B_0 must be chosen so as to exclude these (except when the trust region method is used); hence our convergence results are of an essentially local character. (Were we to relax this condition on B_0, then the arguments yielding Theorem 5.5 below could be modified to establish that these procedures always converge to *some* stationary point of Q.) To state our conditions, let $\sigma_{\min}(D) \coloneqq \varrho_{\min}^{1/2}(D^{\mathsf{T}}D)$ denote the smallest singular value of a (possibly non-square) matrix D,

and recall $G(\beta) = [\partial_\beta \theta(\beta, 0)]^\mathsf{T}$, the Jacobian of the binding function.

Assumption O (optimization routines)**.** *Let* $Q \in \{Q_k^{\mathrm{W}}, Q_k^{\mathrm{LR}}\}$*. Then* $\mathrm{B}_0 = \mathrm{B}_0(Q)$ *may be chosen as any compact subset of* $\operatorname{int} \mathrm{B}$ *for which* $\beta_0 \in \operatorname{int} \mathrm{B}_0$ *and* $\mathrm{B}_0 = \{\beta \in \mathrm{B} \mid Q(\beta) \leq Q(\beta_1)\}$ *for some* $\beta_1 \in \mathrm{B}$*; and either*

GN $\|G(\beta)^\mathsf{T} W g(\beta)\| \neq 0$ *for all* $\beta \in \mathrm{B}_0 \backslash \{\beta_0\}$ *and* $\inf_{\beta \in \mathrm{B}_0} \sigma_{\min}[G(\beta)] > 0$*;*

QN Q *is strictly convex on* B_0*; or*

TR *for every* $\beta \in \mathrm{B}_0 \backslash \{\beta_0\}$*,* $\|\partial_\beta Q(\beta)\| = 0$ *implies* $\varrho_{\min}[\partial_\beta^2 Q(\beta)] < 0$*.*

Remark 5.14. Note that $\|G(\beta)^\mathsf{T} W g(\beta)\| \neq 0$ is equivalent to $\|\partial_\beta Q_k^{\mathrm{W}}(\beta)\| \neq 0$. Both GN and QN thus imply that Q has no stationary points in B_0, other than that which corresponds to the minimum at β_0. TR, on the other hand, permits such points to exist, provided that they are not local minima. In this respect, it places the weakest conditions on Q, and does so because the trust-region method utilizes second-derivative information in a manner that the other two methods do not.

Before analyzing the convergence properties of these optimization routines, we must first specify the conditions governing their termination. Let $\{\beta^{(s)}\}$ denote the sequence of iterates generated by a given routine r, from some starting point $\beta^{(0)}$. When $r \in \{\mathrm{GN}, \mathrm{QN}\}$, we shall allow the optimization to terminate at the first s – denoted s^* – for which a near root is located, in the sense that $\|\partial_\beta Q_{nk}^e(\beta^{(s)})\| \leq c_n$, where $c_n = o_p(n^{-1/2})$. That is, s^* is the smallest s for which $\beta^{(s)} \in R_{nk}^e$. This motivates the definition, for $r \in \{\mathrm{GN}, \mathrm{QN}\}$, of

$$\overline{\beta}_{nk}^e(\beta^{(0)}, r) \coloneqq \begin{cases} \beta^{(s^*)} & \text{if } \beta^{(s)} \in R_{nk}^e \text{ for some } s \in \mathbb{N} \\ \beta^{(0)} & \text{otherwise,} \end{cases} \tag{5.15}$$

which describes the terminal value of the optimization routine, with the convention that this is set to $\beta^{(0)}$ if a near root is never located. In the case that $r = \mathrm{TR}$, we shall allow the routine to terminate only at those near roots at which the second-order sufficient conditions for a local minimum are also satisfied. In this way, s^* now becomes the smallest s for which $\beta^{(s)} \in S_{nk}^{\mathrm{TR}}$, and $\overline{\beta}_{nk}^e(\beta^{(0)}, \mathrm{TR})$ may be defined exactly as in (5.15), except with S_{nk}^{TR} in place of R_{nk}^{TR}.[10]

For the purposes of the next result, let $\hat{\beta}_{nk}^e$ denote the exact minimizer of Q_{nk}^e.

Theorem 5.5 (derivative-based optimizers)**.** *Suppose* $r \in \{\mathrm{GN}, \mathrm{QN}, \mathrm{TR}\}$ *and* $e \in \{\mathrm{W}, \mathrm{LR}\}$*, and that the corresponding part of Assumption O holds for some* B_0*. Then*

$$\sup_{\beta^{(0)} \in \mathrm{B}} \|\overline{\beta}_{nk}^e(\beta^{(0)}, r) - \hat{\beta}_{nk}^e\| = o_p(n^{-1/2})$$

holds if either

[10] It may be asked why we do not also propose checking the second-order conditions upon termination when $r \in \{\mathrm{GN}, \mathrm{QN}\}$. Such a modification is certainly possible, but is perhaps of doubtful utility. Consider the problem of minimizing some (deterministic) criterion function that has multiple roots, only one of which corresponds to a local (and also global) minimum, a scenario envisaged in TR. In this case, the best we can hope to prove is that the Gauss-Newton and quasi-Newton routines will have *some* of those roots as points of accumulation, but they might *never* enter the vicinity of the local minimum (see Theorems 6.5 and 10.1 in Nocedal and Wright, 2006). On the other hand, the trust-region algorithm considered here is guaranteed to have the local minimum as a point of accumulation, under certain conditions (see Moré and Sorensen, 1983, Theorem 4.13).

(i) $(r, e) = (\mathrm{GN}, \mathrm{W})$ *and* H5 *holds with* $l = 1$*; or*

(ii) $r \in \{\mathrm{QN}, \mathrm{TR}\}$ *and* H5 *holds with* $l = 2$.

Remark 5.15. Convergence of the Gauss-Newton method requires the weakest conditions on λ_n of all three algorithms. This is because the Hessian approximation $\Delta_{n,(s)} \coloneqq G_n(\beta^{(s)})^\mathsf{T} W_n G_n(\beta^{(s)})$ used by Gauss-Newton is valid for criteria having the same minimum-distance structure as Q_n^{W}; here $G_n(\beta) \coloneqq \partial_\beta \overline{\theta}_n^k(\beta, \lambda_n)$. Thus the uniform convergence of G_n is sufficient to ensure that $\Delta_{n,(s)}$ behaves suitably in large samples, whence only H5 with $l = 1$ is required.

6 Monte Carlo results

This section conducts a set of Monte Carlo experiments to assess the performance of the GII estimator, in terms of bias, efficiency, and computation time. The parameters of Models 1–4 (see Section 2) are estimated a large number of times using "observed" data generated by the respective models. For each model, the Monte Carlo experiments are conducted for several sets of parameter configurations. For Models 1, 2, and 4, the parameters are estimated in each Monte Carlo replication using both GII and simulated maximum likelihood (SML) in conjunction with the GHK smooth probability simulator (cf. Lee, 1997). Model 3, which cannot easily be estimated via SML, is estimated using only GII. We omit Model 5, as Altonji, Smith, and Vidangos (2013) already present results showing that GII performs well for Heckman selection-type models.

In all cases, we use the LR approach to (generalized) indirect inference to construct our estimates. We do this for two reasons. First, unlike the Wald and LM approaches, the LR approach does not require the estimation of a weight matrix. In this respect, the LR approach is easier to implement than the other two approaches. Furthermore, because estimates of optimal weight matrices often do not perform well in finite samples (see e.g. Altonji and Segal, 1996), the LR approach is likely to perform better in small samples. Second, because the LR approach is asymptotically equivalent to the other two approaches when the auxiliary model is correctly specified, the relative inefficiency of the LR estimator is likely to be small when the auxiliary model is chosen judiciously.

To optimize the criterion functions, we use a version of the Davidon-Fletcher-Powell algorithm (as implemented in Chapter 10 of Press, Flannery, Teukolsky, and Vetterling, 1993), which is closely related to the quasi-Newton routine analyzed in Section 5.6. The initial parameter vector in the hillclimbing algorithm is the true parameter vector. Most of the computation time in generalized indirect inference lies in computing ordinary least squares (OLS) estimates. The main cost in computing OLS estimates lies, in turn, in computing the $X^\mathsf{T}X$ part of $(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$. We use blocking and loop unrolling techniques to speed up the computation of $X^\mathsf{T}X$ by a factor of 2 to 3 relative to a "naive" algorithm.[11]

6.1 Results for Model 1

Model 1 is a two-alternative panel probit model with serially correlated errors and one exogenous

[11] To avoid redundant calculations, we also precompute and store for later use those elements of $X^\mathsf{T}X$ that depend only on the exogenous variables. We are grateful to James MacKinnon for providing code that implements the blocking and loop unrolling techniques.

regressor. It has two unknown parameters: the regressor coefficient b, and the serial correlation parameter r. We set $b = 1$ and consider $r \in \{0, 0.40, 0.85\}$. In the Monte Carlo experiments, $n = 1000$ and $T = 5$. As in all of the simulation exercises carried out in this paper, we compute the GII estimator via the two-step approach described in Section 4.3, using $(\lambda, M) = (0.03, 10)$ in the first step, and $(\lambda, M) = (0.003, 300)$ in the second. The exogenous variables (the x_{it}'s) are i.i.d. draws from a $N[0, 1]$ distribution, drawn anew for each Monte Carlo replication.

The auxiliary model consists of T linear probability models of the form

$$y_{it} = z_{it}^{\mathsf{T}} \alpha_t + \xi_{it}$$

where $\xi_{it} \sim_{\text{i.i.d.}} N[0, \sigma_t^2]$, z_{it} denotes the vector of regressors for individual i in time period t, and α_t and σ_t^2 are parameters to be estimated. We include in z_{it} both lagged choices and polynomial functions of current and lagged exogenous variables; the included variables change over time, so as to allow the auxiliary model to incorporate the additional lagged information that is available in later time periods. (When estimating the model on simulated data, the simulated lagged choices are of course replaced by their smoothed counterparts, as per the discussion in Section 4.1 above.) The auxiliary model is thus characterized by the parameters $\theta = \{\alpha_t, \sigma_t^2\}_{t=1}^T$; these are estimated by maximum likelihood (which corresponds to OLS here, under the distributional assumptions on ξ_{it}).

It is worth emphasizing that we include lagged choices (and lagged x's) in the auxiliary model despite the fact that the structural model does not exhibit true state dependence. But in Model 1 it is well-know that lagged choices are predictive of current choices (termed "spurious state dependence" by Heckman). This is a good illustration of how a good auxiliary model should be designed to capture the correlation patterns in the data, as opposed to the true structure.

To examine how increasing the "richness" of the auxiliary model affects the efficiency of the structural parameter estimates, we conduct Monte Carlo experiments using four nested auxiliary models. In all four, we impose the restrictions $\alpha_t = \alpha_q$ and $\sigma_t^2 = \sigma_q^2$, $t = q + 1, \ldots, T$, for some $q < T$. This is because the time variation in the estimated coefficients of the linear probability models comes mostly from the non-stationarity of the errors in the structural model, and so it is negligible after the first few time periods (we do not assume that the initial error is drawn from the stationary distribution implied by the law of motion for the errors).

In auxiliary model #1, $q = 1$ and the regressors in the linear probability model are given by: $z_{it} = (1, x_{it}, y_{i,t-1})$, $t = 1, \ldots, T$, where the unobserved y_{i0} is set equal to 0. We use this very simple auxiliary model to illustrate how GII can produce very inefficient estimates if one uses a poor auxiliary model. In auxiliary model #2, $q = 2$ and the regressors are $z_{i1} = (1, x_{i1})$, and

$$z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}),\ t \in \{2, \ldots, T\},$$

giving a total of 18 parameters. Auxiliary model #3 has $q = 4$, regressors

$$z_{i1} = (1, x_{i1}, x_{i1}^3) \qquad z_{i3} = (1, x_{i3}, y_{i2}, x_{i2}, y_{i1}, x_{i1})$$
$$z_{i2} = (1, x_{i2}, y_{i1}, x_{i1}) \qquad z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}, y_{i,t-2}, x_{i,t-2}, y_{i,t-3}),\ t \in \{4, \ldots, T\},$$

and 24 parameters. Finally, auxiliary model #4 has the same regressors as #3, except that

$$z_{i4} = (1, x_{i4}, y_{i3}, x_{i3}, y_{i2}, x_{i2}, y_{i1}, x_{i1})$$
$$z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}, y_{i,t-2}, x_{i,t-2}, y_{i,t-3}, x_{i,t-3}, y_{i,t-4}),\ t \in \{5, \ldots, T\}$$

so $q = 5$ and there are 35 parameters.

Table 1 presents the results of six sets of Monte Carlo experiments, each with 2000 replications. The first two sets of experiments report the results for simulated maximum likelihood, based on GHK, using 25 draws (SML #1) and 50 draws (SML #2). The remaining four sets of experiments report the results for generalized indirect inference, where GII #i refers to generalized indirect inference using auxiliary model #i. In each case, we report the average and the standard deviation of the parameter estimates. We also report the efficiency loss of GII #i relative to SML #2 in the columns labelled $\sigma_{\text{GII}}/\sigma_{\text{SML}}$, where we divide the standard deviations of the GII estimates by the standard deviations of the estimates for SML #2. Finally, we report the average time (in seconds) required to compute estimates (we use the Intel Fortran Compiler Version 7.1 on a 2.2GHz Intel Xeon processor running Red Hat Linux).

Table 1 contains several key findings:

First, both SML and GII generate estimates with very little bias.

Second, GII is less efficient than SML, but the efficiency losses are small provided that the auxiliary model is sufficiently rich. For example, auxiliary model #1 leads to large efficiency losses, particularly for the case of high serial correlation in the errors ($r = 0.85$). For models with little serial correlation ($r = 0$), however, auxiliary model #2 is sufficiently rich to to make GII almost as efficient as SML. When there is more serial correlation in the errors, auxiliary model #2 leads to reasonably large efficiency losses (as high as 30% when $r = 0.85$), but auxiliary model #3, which contains more lagged information in the linear probability models than does auxiliary model #2, reduces the worst efficiency loss to 13%. Auxiliary model #4 provides almost no efficiency gains relative to auxiliary model #3.

Third, GII is faster than SML: computing a set of estimates using GII with auxiliary model #3 takes about 30% less time than computing a set of estimates using SML with 50 draws.

For generalized indirect inference, we also compute (but do not report in Table 1) estimated asymptotic standard errors, using the estimators described in Theorem 5.2. In all cases, the averages of the estimated standard errors across the Monte Carlo replications are very close to (within a few percent of) the actual standard deviations of the estimates, suggesting that the asymptotic results provide a good approximation to the behavior of the estimates in samples of the size that we use.

6.2 Results for Model 2

Model 2 is a panel probit model with serially correlated errors, a single exogenous regressor, and a lagged dependent variable. It has three unknown parameters: b_1, the coefficient on the exogenous regressor, b_2, the coefficient on the lagged dependent variable, and r, the serial correlation parameter. We set $b_1 = 1$, $b_2 = 0.2$, and consider $r \in \{0, 0.4, 0.85\}$; $n = 1000$ and $T = 10$.

Table 2 presents the results of six sets of Monte Carlo experiments, each with 1000 replica-

tions; the labels SML #i and GII #i are to be interpreted exactly as for Table 1. The results are similar to those for Model 1. Both SML and GII generate estimates with very little bias. SML is more efficient than GII, but the efficiency loss is small when the auxiliary model is sufficiently rich (i.e., 17% at most for model #3, 15% at most for model #4). However, auxiliary model #1 can lead to very large efficiency losses, as can auxiliary model #2 if there is strong serial correlation.

Again, average asymptotic standard errors are close to the standard deviations obtained across the simulations (not reported). Finally, GII using auxiliary model #3 is about 25% faster than SML using 50 draws.

6.3 Results for Model 3

Model 3 is identical to Model 2, except there is an "initial conditions" problem: the econometrician does not observe individuals' choices in the first s periods. This is an excellent example of the type of problem that motivates this paper: SML is extremely difficult to implement, due to the problem of integrating over the initial conditions. But II is appealing, as it is still trivial to simulate data from the model. However, we need GII to deal with the discrete outcomes.

To proceed, our Monte Carlo experiments are parametrized exactly as for Model 2, except that we set $T = 15$, with choices in the first $s = 5$ time periods being unobserved (but note that exogenous variables *are* observed in these time periods).

Auxiliary model #1 is as for Models 1 and 2: $q = 1$ and the regressors are $z_{it} = (1, x_{it}, y_{i,t-1})$, $t = s+1, \ldots, T$, where the unobserved y_{is} is set equal to 0. In auxiliary model #2, $q = 2$ and the regressors are:

$$z_{i,s+1} = (1, x_{i,s+1}, x_{is}) \qquad z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}),\ t \in \{s+2, \ldots, T\},$$

for a total of 19 parameters. In auxiliary model #3, $q = 4$ and there are 27 parameters:

$$\begin{aligned} z_{i,s+1} &= (1, x_{i,s+1}, x_{i,s+1}^3, x_{is}, x_{i,s-1}) \\ z_{i,s+2} &= (1, x_{i,s+2}, y_{i,s+1}, x_{i,s+1}, x_{is}) \\ z_{i,s+3} &= (1, x_{i,s+3}, y_{i,s+2}, x_{i,s+2}, y_{i,s+1}, x_{i,s+1}) \\ z_{it} &= (1, x_{it}, y_{i,t-1}, x_{i,t-1}, y_{i,t-2}, x_{i,t-2}, y_{i,t-3}),\ t \in \{s+4, \ldots, T\} \end{aligned}$$

Finally, in auxiliary model #4, $q = 5$ and there are 41 parameters: relative to #3, $z_{i,s+1}$, $z_{i,s+2}$ and $z_{i,s+3}$ are augmented by an additional lag of x_{is}, and

$$\begin{aligned} z_{i,s+4} &= (1, x_{i,s+4}, y_{i,s+3}, x_{i,s+3}, y_{i,s+2}, x_{i,s+2}, y_{i,s+1}, x_{i,s+1}) \\ z_{it} &= (1, x_{it}, y_{i,t-1}, x_{i,t-1}, y_{i,t-2}, x_{i,t-2}, y_{i,t-3}, x_{i,t-3}, y_{i,t-4}),\ t \in \{s+5, \ldots, T\}. \end{aligned}$$

Table 3 presents the results of four sets of Monte Carlo experiments, each with 1000 replications. There are two key findings: First, as with Models 1 and 2, GII generates estimates with very little bias. Second, increasing the "richness" of the auxiliary model leads to large efficiency gains relative to auxiliary model #1, particularly when the errors are persistent. However, auxiliary model #4 provides few efficiency gains relative to auxiliary model #3.

6.4 Results for Model 4

Model 4 is a (static) three-alternative probit model with eight unknown parameters: three coefficients in each of the two equations for the latent utilities ($\{b_{1i}\}_{i=0}^{2}$ and $\{b_{2i}\}_{i=0}^{2}$) and two parameters governing the covariance matrix of the disturbances in these equations (c_1 and c_2). We set $b_{10} = b_{20} = 0$, $b_{11} = b_{12} = b_{21} = b_{22} = 1$, $c_2 = 1$, and consider $c_1 \in \{0, 1.33\}$ (implying that the disturbances in the latent utilities are respectively independent, or have a correlation of 0.8). We set $n = 2000$.

The auxiliary model is a pair of linear probability models, one for each of the first two alternatives:

$$y_{i1} = z_i^{\mathsf{T}}\alpha_1 + \xi_{i1}$$
$$y_{i2} = z_i^{\mathsf{T}}\alpha_2 + \xi_{i2},$$

where z_i consists of polynomial functions of the exogenous variables $\{x_{ij}\}_{j=1}^{3}$, and $\xi_i \sim_{\text{i.i.d.}} N[0, \Sigma_\xi]$. The auxiliary model parameters $\theta = (\alpha_1, \alpha_2, \Sigma_\xi)$ are estimated by OLS; this corresponds to maximum likelihood – even though Σ_ξ is not diagonal – because the same regressors appear in both equations.

We conduct Monte Carlo experiments using four nested versions of the auxiliary model. In auxiliary model #1, $z_i = (1, x_{i1}, x_{i2}, x_{i3})$, giving a total of 11 parameters. Auxiliary model #2 adds all the second-order products of these variables, as well as one third-order product to z_i, i.e.

$$z_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i1}^2, x_{i2}^2, x_{i3}^2, x_{i1}x_{i2}, x_{i1}x_{i3}, x_{i2}x_{i3}, x_{i1}x_{i2}x_{i3}),$$

for a total of 25 parameters. In auxiliary model #3, z_i contains all third-order products (for a total of 43 parameters) and in auxiliary model #4, z_i contains all fourth-order products (for a total of 67 parameters).

Tables 4 and 5 present the results of six sets of Monte Carlo experiments, each with 1000 replications; the labels SML #i and GII #i are to be interpreted exactly as for Table 1. The key findings are qualitatively similar to those for Models 1, 2, and 3. First, both SML and GII generate estimates with very little bias. Second, auxiliary model #1, which contains only linear terms, leads to large efficiency losses relative to SML (as large as 50%). But auxiliary model #2, which contains terms up to second order, reduces the efficiency losses substantially (to no more than 15% when the errors are uncorrelated, and to no more than 26% when $c = 1.33$). Auxiliary model #3, which contains terms up to third order, provides additional small efficiency gains (the largest efficiency loss is reduced to 20%), while auxiliary model #4, which contains fourth-order terms, provides few, if any, efficiency gains relative to auxiliary model #3. Finally, computing estimates using GII with auxiliary model #3 takes about 30% less time than computing estimates using SML with 50 draws.

7 Conclusion

Discrete choice models play an important role in many fields of economics, from labor economics to industrial organization to macroeconomics. Unfortunately, these models are usually quite

challenging to estimate (except in special cases like MNL where choice probabilities have closed forms). Simulation-based methods like SML and MSM have been developed that can be used for more complex models like MNP. But in many important cases (models with initial conditions problems and Heckman selection models being leading cases) even these methods are very difficult to implement.

In this paper we develop and implement a new simulation-based method for estimating models with discrete or mixed discrete/continuous outcomes. The method is based on indirect inference. But the traditional II approach is not easily applicable to discrete choice models because one must deal with a non-smooth objective surface. The key innovation here is that we develop a generalized method of indirect inference (GII), in which the auxiliary models that are estimated on the actual and simulated data may differ (provided that the estimates from both models share a common probability limit). This allows us to chose an auxiliary model for the simulated data such that we obtain an objective function that is a smooth function of the structural parameters. This smoothness renders GII practical as a method for estimating discrete choice models.

Our theoretical analysis goes well beyond merely deriving the limiting distribution of the minimizer of the GII criterion function. Rather, in keeping with computational motivation of this paper, we show that the proposed smoothing facilitates the convergence of derivative-based optimizers, in the sense that the smoothing leads to a sample optimization problem that is no more difficult than the corresponding population problem, where the latter involves the minimization of a necessarily smooth criterion. This provides a rigorous justification for using standard derivative-based optimizers to compute the GII estimator, which is also shown to inherit the limiting distribution of the (unsmoothed) II estimator. Inferences based on the GII estimates may thus be drawn in the standard manner, via the usual Wald statistics. Our results on the convergence of derivative-based optimizers seem to be new to the literature.

We also provide a set of Monte Carlo experiments to illustrate the practical usefulness of GII. In addition to being robust and fast, GII yields estimates with good properties in small samples. In particular, the estimates display very little bias and are nearly as efficient as maximum likelihood (in those cases where simulated versions of maximum likelihood can be used) provided that the auxiliary model is chosen judiciously.

GII could potentially be applied to a wide range of discrete and discrete/continuous outcome models beyond those we consider in our Monte Carlo experiments. Indeed, GII is sufficiently flexible to accommodate almost any conceivable model of discrete choice, including, discrete choice dynamic programming models, discrete dynamic games, etc. We hope that applied economists from a variety of fields find GII a useful and easy-to-implement method for estimating discrete choice models.

8 References

Altonji, J. G., and L. M. Segal (1996): "Small-sample bias in GMM estimation of covariance structures," *Journal of Business and Economic Statistics*, 14(3), 353–66.

Altonji, J. G., A. A. Smith, and I. Vidangos (2013): "Modeling earnings dynamics," *Econometrica*, 81(4), 1395–1454.

AN, M. Y., AND M. LIU (2000): "Using indirect inference to solve the initial-conditions problem," *Review of Economics and Statistics*, 82(4), 656–67.

BILLINGSLEY, P. (1968): *Convergence of Probability Measures*. Wiley, New York (USA).

CASSIDY, H. (2012): "Skills, tasks, and occupational choice," University of Western Ontario.

CHERNOZHUKOV, V., AND H. HONG (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115(2), 293–346.

DIGGLE, P. J., AND R. J. GRATTON (1984): "Monte Carlo methods of inference for implicit statistical models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2), 193–227.

EINMAHL, U., AND D. M. MASON (2005): "Uniform in bandwidth consistency of kernel-type function estimators," *The Annals of Statistics*, 33(3), 1380–1403.

EISENHAUER, P., J. J. HECKMAN, AND S. MOSSO (2015): "Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments," *International Economic Review*, 56(2), 331–357.

ENGLE, R. F., AND D. L. MCFADDEN (eds.) (1994): *Handbook of Econometrics*, vol. IV. Elsevier.

FERMANIAN, J.-D., AND B. SALANIÉ (2004): "A nonparametric simulated maximum likelihood estimation method," *Econometric Theory*, 20(4), 701–34.

FLETCHER, R. (1987): *Practrical Methods of Optimization*. Wiley, Chichester (UK), 2nd edn.

GALLANT, A. R., AND G. TAUCHEN (1996): "Which moments to match?," *Econometric Theory*, 12(4), 657–81.

GAN, L., AND G. GONG (2007): "Estimating interdependence between health and education in a dynamic model," Working Paper 12830, National Bureau of Economic Research.

GENTON, M. G., AND E. RONCHETTI (2003): "Robust indirect inference," *Journal of the American Statistical Association*, 98(461), 67–76.

GEWEKE, J., AND M. P. KEANE (2001): "Computationally intensive methods for integration in econometrics," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. 5. Elsevier.

GOURIEROUX, C., A. MONFORT, AND E. RENAULT (1993): "Indirect inference," *Journal of Applied Econometrics*, 8(S1), S85–S118.

GREENE, W. H. (2008): *Econometric Analysis*. Pearson Prentice Hall, New Jersey (USA), 6th edn.

HECKMAN, J. J. (1981): "The incidental parameters problem and the problem of initial conditions in estimating a discrete time–discrete data stochastic process," in Manski and McFadden (1981), pp. 179–95.

HOROWITZ, J. L. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica*, 60(3), 505–31.

——— (1998): "Bootstrap methods for median regression models," *Econometrica*, 66(6), 1327–51.

Jones, M. C., and P. J. Foster (1993): "Generalized jackknifing and higher order kernels," *Journal of Nonparametric Statistics*, 3(1), 81–94.

Kaplan, D. M., and Y. Sun (2012): "Smoothed estimating equations for instrumental variables quantile regression," University of California, San Diego.

Keane, M., and A. A. Smith (2003): "Generalized indirect inference for discrete choice models," Yale University.

Keane, M. P. (1994): "A computationally practical simulation estimator for panel data," *Econometrica*, 62, 95–116.

Keane, M. P., and R. M. Sauer (2010): "A computationally practical simulation estimation algorithm for dynamic panel data models with unobserved endogenous state variables," *International Economic Review*, 51(4), 925–958.

Kim, J., and D. Pollard (1990): "Cube root asymptotics," *Annals of Statistics*, 18(1), 191–219.

Kormiltsina, A., and D. Nekipelov (2012): "Approximation properties of Laplace-type estimators," UC Berkeley.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Kristensen, D., and Y. Shin (2012): "Estimation of dynamic models with nonparametric simulated maximum likelihood," *Journal of Econometrics*, 167(1), 76–94.

Lee, L.-F. (1997): "Simulated maximum likelihood estimation of dynamic discrete choice statistical models: some Monte Carlo results," *Journal of Econometrics*, 82(1), 1–35.

Lerman, S., and C. F. Manski (1981): "On the use of simulated frequencies to approximate choice probabilities," in Manski and McFadden (1981), pp. 305–319.

Li, T., and B. Zhang (2015): "Affiliation and entry in first-price auctions with heterogeneous bidders: an analysis of merger effects," *American Economic Journal: Microeconomics*, 7(2), 188–214.

Lopez Garcia, I. (2015): "Human capital and labor informality in Chile: a life-cycle approach," Working Paper WR-1087, RAND Corporation.

Lopez-Mayan, C. (2014): "Microeconometric analysis of residential water demand," *Environmental and Resource Economics*, 59(1), 137–166.

Magnac, T., J.-M. Robin, and M. Visser (1995): "Analysing incomplete individual employment histories using indirect inference," *Journal of Applied Econometrics*, 10(1), S153–S169.

Magnus, J. R., and H. Neudecker (2007): *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Chichester (UK), 3rd edn.

Manski, C. F. (1985): "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27(3), 313–33.

Manski, C. F., and D. McFadden (eds.) (1981): *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press, Cambridge, MA.

McFadden, D. L. (1989): "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, 57, 995–1026.

MORÉ, J. J., AND D. C. SORENSEN (1983): "Computing a trust region step," *SIAM Journal on Scientific and Statistical Computing*, 4(3), 553–72.

MORTEN, M. (2013): "Temporary migration and endogenous risk sharing in village india," Stanford University.

NAGYPÁL, É. (2007): "Learning by doing vs. learning about match quality: Can we tell them apart?," *Review of Economic Studies*, 74(2), 537–66.

NEWEY, W. K., AND D. L. MCFADDEN (1994): "Large sample estimation and hypothesis testing," in Engle and McFadden (1994), pp. 2111–2245.

NOCEDAL, J., AND S. J. WRIGHT (2006): *Numerical Optimization*. Springer, 2nd edn.

NOLAN, D., AND D. POLLARD (1987): "U-processes: rates of convergence," *Annals of Statistics*, 15(2), 780–99.

OTSU, T. (2008): "Conditional empirical likelihood estimation and inference for quantile regression models," *Journal of Econometrics*, 142(1), 508–38.

PAKES, A., AND D. POLLARD (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, 57(5), 1027–57.

POLLARD, D. (1984): *Convergence of Stochastic Processes*. Springer, New York (USA).

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY, AND W. T. VETTERLING (1993): *Numerical Recipes: the art of scientific computing*. Cambridge University Press, Cambridge (UK), 2nd edn.

ROBINSON, P. M. (1988): "The stochastic difference between econometric statistics," *Econometrica*, 56(3), 531–548.

SAUER, R. M., AND C. TABER (2013): "Indirect inference with importance sampling," Royal Holloway, University of London.

SIDI, A. (2003): *Practical Extrapolation Methods: Theory and Applications*. Cambridge University Press, Cambridge (UK).

SKIRA, M. M. (2015): "Dynamic wage and employment effects of elder parent care," *International Economic Review*, 56(1), 63–93.

SMITH, JR., A. A. (1990): "Three Essays on the Solution and Estimation of Dynamic Macroeconomic Models," Ph.D. thesis, Duke University.

——— (1993): "Estimating nonlinear time-series models using simulated vector autoregressions," *Journal of Applied Econometrics*, 8(S1), S63–S84.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes: with applications to statistics*. Springer, New York (USA).

WHANG, Y.-J. (2006): "Smoothed empirical likelihood methods for quantile regression models," *Econometric Theory*, 22(2), 173–205.

YPMA, J. Y. (2013): "Dynamic models of continuous and discrete outcomes; methods and applications," Ph.D. thesis, University College London.

Table 1

Monte Carlo Results for Model 1

	Mean		Std. dev.		$\sigma_{\text{GII}}/\sigma_{\text{SML}}$		Time
	b	r	b	r	b	r	(sec.)
			$b = 1,\ r = 0$				
SML #1	1.000	−0.002	0.0387	0.0454	—	—	0.76
SML #2	1.001	−0.000	0.0373	0.0468	—	—	1.53
GII #1	0.998	0.002	0.0390	0.0645	1.05	1.37	0.67
GII #2	0.993	0.001	0.0386	0.0490	1.03	1.05	0.72
GII #3	0.992	0.001	0.0393	0.0490	1.05	1.05	0.91
GII #4	0.988	0.001	0.0390	0.0485	1.05	1.04	0.99
			$b = 1,\ r = 0.4$				
SML #1	0.995	0.385	0.0400	0.0413	—	—	0.78
SML #2	0.999	0.392	0.0390	0.0410	—	—	1.54
GII #1	0.998	0.399	0.0454	0.0616	1.16	1.50	0.70
GII #2	0.993	0.396	0.0410	0.0456	1.05	1.11	0.72
GII #3	0.991	0.395	0.0417	0.0432	1.07	1.05	0.91
GII #4	0.987	0.392	0.0416	0.0432	1.07	1.05	0.97
			$b = 1,\ r = 0.85$				
SML #1	0.984	0.833	0.0452	0.0333	—	—	0.74
SML #2	0.993	0.842	0.0432	0.0316	—	—	1.47
GII #1	0.994	0.846	0.0791	0.0672	1.83	2.13	0.71
GII #2	0.991	0.845	0.0511	0.0412	1.18	1.30	0.74
GII #3	0.992	0.846	0.0492	0.0357	1.14	1.13	0.93
GII #4	0.988	0.841	0.0490	0.0357	1.13	1.13	1.00

Table 2

Monte Carlo Results for Model 2

	Mean			Std. dev.			$\sigma_{\text{GII}}/\sigma_{\text{SML}}$			Time
	b_1	r	b_2	b_1	r	b_2	b_1	r	b_2	(sec.)
$b_1 = 1$, $r = 0$, $b_2 = 0.2$										
SML #1	1.000	0.001	0.200	0.0274	0.0357	0.0355	—	—	—	2.47
SML #2	1.002	0.002	0.199	0.0273	0.0362	0.0365	—	—	—	4.89
GII #1	0.999	0.001	0.199	0.0267	0.0571	0.0437	0.98	1.58	1.20	2.72
GII #2	0.996	0.000	0.199	0.0267	0.0379	0.0379	0.98	1.05	1.04	2.80
GII #3	0.995	0.001	0.199	0.0269	0.0377	0.0376	0.99	1.04	1.03	3.66
GII #4	0.993	0.000	0.198	0.0270	0.0377	0.0375	0.99	1.04	1.03	4.06
$b_1 = 1$, $r = 0.4$, $b_2 = 0.2$										
SML #1	0.994	0.379	0.214	0.0278	0.0314	0.0397	—	—	—	2.42
SML #2	0.999	0.389	0.206	0.0287	0.0316	0.0397	—	—	—	4.82
GII #1	0.997	0.397	0.198	0.0339	0.0587	0.0544	1.18	1.86	1.37	2.73
GII #2	0.994	0.396	0.198	0.0293	0.0386	0.0462	1.02	1.22	1.16	2.82
GII #3	0.993	0.396	0.197	0.0289	0.0343	0.0431	1.01	1.09	1.09	3.64
GII #4	0.991	0.395	0.196	0.0289	0.0348	0.0434	1.01	1.10	1.09	4.02
$b_1 = 1$, $r = 0.85$, $b_2 = 0.2$										
SML #1	0.974	0.831	0.220	0.0321	0.0174	0.0505	—	—	—	2.78
SML #2	0.987	0.840	0.208	0.0327	0.0159	0.0507	—	—	—	5.47
GII #1	1.000	0.854	0.183	0.0952	0.0633	0.1185	2.91	3.98	2.34	3.01
GII #2	0.992	0.852	0.190	0.0417	0.0266	0.0721	1.28	1.67	1.42	2.92
GII #3	0.992	0.851	0.191	0.0383	0.0179	0.0547	1.17	1.13	1.08	3.68
GII #4	0.990	0.850	0.188	0.0379	0.0175	0.0548	1.15	1.10	1.09	4.06

Table 3

Monte Carlo Results for Model 3

	Mean			Std. dev.			Time
	b_1	r	b_2	b_1	r	b_2	(sec.)
			$b_1 = 1$, $r = 0$, $b_2 = 0.2$				
GII #1	0.997	−0.000	0.200	0.0272	0.0532	0.0387	3.91
GII #2	0.994	−0.001	0.200	0.0271	0.0387	0.0347	4.01
GII #3	0.993	−0.001	0.199	0.0272	0.0385	0.0345	4.81
GII #4	0.991	−0.001	0.199	0.0275	0.0389	0.0347	5.38
			$b_1 = 1$, $r = 0.4$, $b_2 = 0.2$				
GII #1	0.994	0.397	0.198	0.0361	0.0518	0.0493	3.99
GII #2	0.991	0.397	0.197	0.0309	0.0363	0.0430	4.00
GII #3	0.990	0.396	0.196	0.0306	0.0317	0.0399	4.80
GII #4	0.987	0.395	0.196	0.0302	0.0318	0.0400	5.35
			$b_1 = 1$, $r = 0.85$, $b_2 = 0.2$				
GII #1	0.993	0.851	0.184	0.0936	0.0403	0.1289	4.41
GII #2	0.986	0.851	0.191	0.0546	0.0249	0.0905	4.37
GII #3	0.987	0.850	0.189	0.0430	0.0140	0.0598	4.93
GII #4	0.984	0.849	0.185	0.0411	0.0136	0.0597	5.56

Table 4

Monte Carlo Results for Model 4

($b_{10} = 0$, $b_{11} = 1$, $b_{12} = 1$, $b_{20} = 0$, $b_{21} = 1$, $b_{22} = 1$, $c_1 = 0$, $c_2 = 1$)

	SML		GII				$\sigma_{\text{GII}}/\sigma_{\text{SML}}$			
	#1	#2	#1	#2	#3	#4	#1	#2	#3	#4
					Mean					
b_{10}	0.007	0.005	0.003	0.002	0.002	0.002	—	—	—	—
b_{11}	1.000	1.001	0.995	0.994	0.992	0.990	—	—	—	—
b_{12}	1.000	1.003	0.998	0.997	0.995	0.992	—	—	—	—
b_{20}	−0.001	−0.003	−0.006	−0.004	−0.004	0.004	—	—	—	—
b_{21}	1.006	1.007	1.001	0.999	0.997	0.996	—	—	—	—
b_{22}	1.005	1.007	1.004	1.000	0.998	0.996	—	—	—	—
c_1	0.020	0.010	0.007	0.005	0.005	0.006	—	—	—	—
c_2	1.004	1.003	1.006	1.001	1.001	1.002	—	—	—	—
					Std. dev.					
b_{10}	0.0630	0.0628	0.0720	0.0666	0.0656	0.0665	1.15	1.06	1.04	1.06
b_{11}	0.0686	0.0686	0.0872	0.0764	0.0741	0.0743	1.27	1.11	1.08	1.08
b_{12}	0.0572	0.0574	0.0719	0.0667	0.0632	0.0646	1.25	1.16	1.10	1.13
b_{20}	0.0663	0.0657	0.0745	0.0686	0.0677	0.0676	1.13	1.04	1.04	1.03
b_{21}	0.1065	0.1050	0.1395	0.1128	0.1095	0.1099	1.33	1.07	1.04	1.05
b_{22}	0.1190	0.1174	0.1593	0.1285	0.1249	0.1244	1.36	1.09	1.06	1.06
c_1	0.1091	0.1107	0.1303	0.1276	0.1224	0.1265	1.18	1.15	1.11	1.14
c_2	0.1352	0.1325	0.1991	0.1509	0.1439	0.1421	1.50	1.14	1.09	1.07
Time	11.5	23.1	7.1	10.4	16.4	34.1	—	—	—	—

Table 5

Monte Carlo Results for Model 4

($b_{10} = 0$, $b_{11} = 1$, $b_{12} = 1$, $b_{20} = 0$, $b_{21} = 1$, $b_{22} = 1$, $c_1 = 1.33$, $c_2 = 1$)

	SML		GII				$\sigma_{\text{GII}}/\sigma_{\text{SML}}$			
	#1	#2	#1	#2	#3	#4	#1	#2	#3	#4
					Mean					
b_{10}	−0.031	−0.017	0.000	−0.001	−0.000	−0.001	—	—	—	—
b_{11}	0.998	1.000	0.993	0.993	0.991	0.989	—	—	—	—
b_{12}	1.016	1.011	0.998	0.998	0.996	0.994	—	—	—	—
b_{20}	−0.011	−0.010	−0.011	−0.007	−0.007	−0.006	—	—	—	—
b_{21}	0.992	0.999	1.000	0.997	0.995	0.991	—	—	—	—
b_{22}	1.004	1.008	1.006	1.001	0.999	0.995	—	—	—	—
c_1	1.269	1.306	1.347	1.338	1.335	1.330	—	—	—	—
c_2	1.025	1.011	0.993	0.993	0.995	0.997	—	—	—	—
					Std. dev.					
b_{10}	0.0693	0.0698	0.0789	0.0776	0.0758	0.0757	1.13	1.11	1.09	1.08
b_{11}	0.0587	0.0588	0.0696	0.0658	0.0632	0.0636	1.18	1.12	1.07	1.08
b_{12}	0.0745	0.0737	0.0883	0.0801	0.0781	0.0782	1.20	1.09	1.06	1.06
b_{20}	0.0766	0.0764	0.0900	0.0801	0.0786	0.0780	1.18	1.05	1.03	1.02
b_{21}	0.0884	0.0886	0.1140	0.0969	0.0952	0.0943	1.29	1.09	1.07	1.06
b_{22}	0.1106	0.1103	0.1471	0.1204	0.1176	0.1153	1.34	1.09	1.07	1.05
c_1	0.1641	0.1707	0.2454	0.2152	0.2049	0.2041	1.44	1.26	1.20	1.20
c_2	0.1229	0.1206	0.1599	0.1387	0.1338	0.1311	1.33	1.15	1.11	1.09
Time	12.7	25.6	7.4	10.8	17.1	34.4	—	—	—	—

A Details of optimization routines

Both line-search methods (Gauss-Newton and quasi-Newton) involve the use of a positive definite Hessian $\Delta_{(s)}$ in the approximating model (5.14), and so the problem solved at step $s+1$ reduces to that of "approximately" solving

$$\min_{\alpha \in \mathbb{R}} Q(\beta^{(s)} + \alpha p_{(s)}), \tag{A.1}$$

where $p_{(s)} \coloneqq -\Delta_{(s)}^{-1} \nabla_{(s)}$. We do not require that $\alpha_{(s)}$ solve (A.1) exactly; we shall require only that it satisfy the strong Wolfe conditions,

$$\begin{aligned} Q(\beta^{(s)} + \alpha_{(s)} p_{(s)}) &\leq Q(\beta^{(s)}) + c_1 \alpha_{(s)} \nabla_{(s)}^{\mathsf{T}} p_{(s)} \\ |\dot{Q}(\beta^{(s)} + \alpha_{(s)} p_{(s)})^{\mathsf{T}} p_{(s)}| &\leq c_2 |\nabla_{(s)}^{\mathsf{T}} p_{(s)}| \end{aligned}$$

for $0 < c_1 < c_2 < 1$, where $\dot{Q} \coloneqq \partial_\beta Q$ (cf. (3.7) in Nocedal and Wright, 2006). For some such $\alpha_{(s)}$, we set $\beta^{(s+1)} = \beta^{(s)} + \alpha_{(s)} p_{(s)}$. For the Hessians $\Delta_{(s)}$, the Gauss-Newton method is only applicable to criteria of the form $Q(\beta) = \|g(\beta)\|_W^2$, and uses

$$\nabla^{(s)} \coloneqq -(G_{(s)}^{\mathsf{T}} W G_{(s)})^{\mathsf{T}} G_{(s)}^{\mathsf{T}} W g(\beta^{(s)}),$$

where $G_{(s)} \coloneqq [\partial_\beta g(\beta^{(s)})]^{\mathsf{T}}$. The Quasi-Newton method with BFGS updating starts with some initial positive definite $\Delta_{(0)}$, and updates it according to,

$$\Delta_{(s+1)} = \Delta_{(s)} - \frac{\Delta_{(s)} x_{(s)} x_{(s)}^{\mathsf{T}} \Delta_{(s)}}{x_{(s)}^{\mathsf{T}} \Delta_{(s)} x_{(s)}} + \frac{d_{(s)} d_{(s)}^{\mathsf{T}}}{d_{(s)}^{\mathsf{T}} x_{(s)}},$$

where $x_{(s)} \coloneqq \alpha_{(s)} p_{(s)}$ and $d_{(s)} = \nabla^{(s+1)} - \nabla^{(s)}$ (cf. (6.19) in Nocedal and Wright, 2006).

The trust region method considered here sets $\Delta_{(s)} = \partial_\beta^2 Q(\beta^{(s)})$, which need not be positive definite. The procedure then attempts to approximately minimize (5.14), subject to the constraint that $\|\beta\| \leq \delta_{(s)}$, where $\delta_{(s)}$ defines the size of the trust region, which is adjusted at each iteration depending on the value of

$$\rho_{(s)} \coloneqq \frac{Q(\beta^{(s)}) - Q(\beta^{(s+1)})}{f_{(s)}(0) - f_{(s)}(\beta^{(s+1)})},$$

which measures the proximity of the true reduction in Q at step s, with that predicted by the approximating model (5.14); the adjustment is made in accordance with Algorithm 4.2 in Moré and Sorensen (1983). Various algorithms are available for approximately solving (5.14) in this case, but we shall assume that Algorithm 3.14 from that paper is used.

B Proofs of theorems under high-level assumptions

Assumptions R and H are assumed to hold throughout this section, including H5 with $l = 0$. Whenever we require H5 to hold for some $l \in \{1, 2\}$, this will be explicitly noted. The relationships between the theorems and the auxiliary results (Propositions B.1–B.5) is illustrated in Figure B.1.

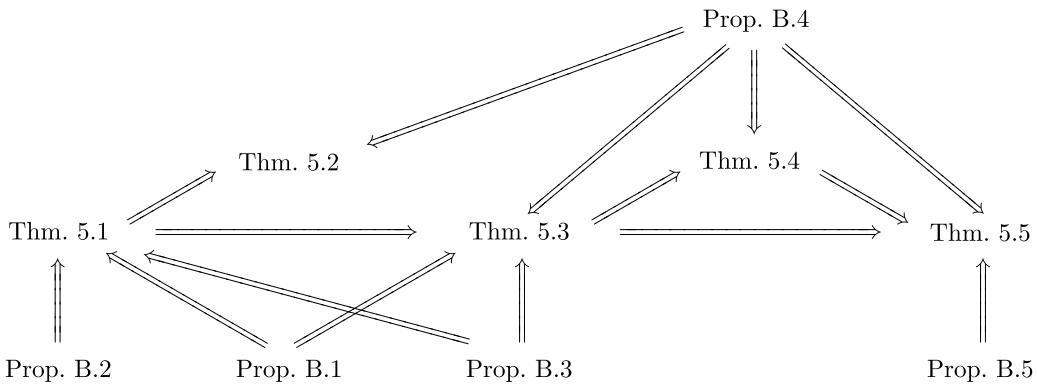

Figure B.1: Proofs of theorems

B.1 Preliminary results

Let $\beta_n \coloneqq \beta_0 + n^{-1/2}\delta_n$ for a (possibly) random $\delta_n = o_p(n^{1/2})$. Define

$$\Delta_n^k(\beta) \coloneqq n^{1/2}[\overline{\theta}_n^k(\beta, \lambda_n) - \overline{\theta}_n^k(\beta_0, \lambda_n)]$$

and recall that $G_n(\beta) \coloneqq \partial_\beta \overline{\theta}_n^k(\beta, \lambda_n)$ and $G \coloneqq [\partial_\beta \theta(\beta_0, 0)]^\mathsf{T}$. As in R5, $\lambda_n = o_p(1)$ is an $\mathcal{F}$-measurable sequence. As per R6, we fix the order of jackknifing $k \in \{0, \dots, k_0\}$ such that $n^{1/2}\lambda_n^{k+1} = o_p(1)$. Let $\mathcal{L}_n(\theta) \coloneqq \mathcal{L}_n(y, x; \theta)$ and $\mathcal{L}(\theta) \coloneqq \mathbb{E}\mathcal{L}_n(\theta)$. $\dot{\mathcal{L}}_n$ and $\ddot{\mathcal{L}}_n$ respectively denote the gradient and Hessian of $\mathcal{L}_n$, with $H \coloneqq \mathbb{E}\ddot{\mathcal{L}}_n(\theta) = \mathcal{L}(\theta)$; $N(\theta, \epsilon)$ denotes an open ball of radius ϵ, centered at θ.

Proposition B.1.

(i) $\sup_{\beta \in \mathrm{B}} \|\overline{\theta}_n^k(\beta, \lambda_n) - \theta^k(\beta, \lambda_n)\| \xrightarrow{p} 0$;

(ii) $\theta^k(\beta_0, \lambda_n) - \theta(\beta_0, 0) = O_p(\lambda_n^{k+1})$;

(iii) $\Delta_n^k(\beta_n) = G\delta_n + o_p(1 + \|\delta_n\|)$;

Proposition B.2. *For* $V = (1 + \frac{1}{M})(\Sigma - \mathrm{R})$,

$$Z_n \coloneqq n^{1/2}[\overline{\theta}_n^k(\beta_0, \lambda_n) - \theta^k(\beta_0, \lambda_n)] - n^{1/2}(\hat{\theta}_n - \theta_0) \rightsquigarrow N[0, H^{-1}VH^{-1}]. \tag{B.1}$$

Proposition B.3.

(i) $Q_{nk}^e(\beta, \lambda_n) \xrightarrow{p} Q_k^e(\beta, 0) =: Q^e(\beta)$ *uniformly on* B;

(ii) *for every* $\epsilon > 0$, $\inf_{\beta \in \mathrm{B} \setminus N(\beta_0, \epsilon)} Q^e(\beta) > Q(\beta_0)$; *and*

Proposition B.4. *If* H5 *holds for* $l = 1$, *then*

(i) $G_n(\beta_n) \xrightarrow{p} G$; *and*

if H5 *holds for* $l \in \{1, 2\}$ *then, uniformly on* B,

(ii) $\sup_{\beta \in \mathrm{B}} \|\partial_\beta^l \overline{\theta}_n^k(\beta, \lambda_n) - \partial_\beta^l \theta(\beta, 0)\| = o_p(1)$; *and*

(iii) $\partial_\beta^l Q_{nk}^e(\beta, \lambda_n) \xrightarrow{p} \partial_\beta^l Q_k^e(\beta, 0) = \partial_\beta^l Q(\beta)$.

For the next result, let $U : \Gamma \to \mathbb{R}$ be twice continuously differentiable with a global minimum at γ^*. Let $R_U \coloneqq \{\gamma \in \Gamma \mid \|\partial_\gamma U(\gamma)\| < \epsilon\}$ for some $\epsilon > 0$, and $S_U \coloneqq \{\gamma \in R_U \mid \varrho_{\min}[\partial_\gamma^2 U(\gamma)] \geq 0\}$. Applying a routine $r \in \{\text{GN}, \text{QN}, \text{TR}\}$ to U yields the iterates $\{\gamma^{(s)}\}$; let

$$\overline{\gamma}(\gamma^{(0)}, r) \coloneqq \begin{cases} \gamma^{(s^*)} & \text{if } \gamma^{(s)} \in R_U \text{ for some } s \in \mathbb{N} \\ \gamma^{(0)} & \text{otherwise,} \end{cases}$$

where s^* denotes the smallest s for which $\gamma^{(s)} \in R_U$. When $r = \text{TR}$, the definition of $\overline{\gamma}(\gamma^{(0)}, \text{TR})$ is analogous, but with S_U in place of R_U. In the statement of the next result, $\Gamma_0 \coloneqq \{\gamma \in \Gamma \mid U(\gamma) \leq U(\gamma_1)\}$ for some $\gamma_1 \in \Gamma$, and is a compact set with $\gamma^* \in \operatorname{int} \Gamma_0$. For a function $m : \Gamma \mapsto \mathbb{R}^{d_m}$, let $M(\gamma) \coloneqq [\partial_\gamma m(\gamma)]^\mathsf{T}$ denote its Jacobian.

Proposition B.5. *Let $r \in \{\text{QN}, \text{TR}\}$, and suppose that in addition to the preceding, either*

(i) *$r = \text{GN}$ and $U(\gamma) = \|m(\gamma)\|^2$, with $\inf_{\gamma \in \Gamma_0} \sigma_{\min}[M(\gamma)] > 0$; or*

(ii) *$r = \text{QN}$ and U is strictly convex on Γ_0;*

then $\overline{\gamma}(\gamma^{(0)}, r) \in R_U \cap \Gamma_0$ for all $\gamma^{(0)} \in \Gamma_0$. Alternatively, if $r = \text{TR}$, then $\overline{\gamma}(\gamma^{(0)}, r) \in S_U \cap \Gamma_0$ for all $\gamma^{(0)} \in \Gamma_0$.

B.2 Proofs of Theorems 5.1–5.5

Throughout this section, $\beta_n \coloneqq \beta_0 + n^{-1/2}\delta_n$ for a (possibly) random $\delta_n = o_p(n^{1/2})$. Let $Q_n^{\text{W}}(\beta) \coloneqq Q_{nk}^{\text{W}}(\beta, \lambda_n)$, $Q_n^{\text{LR}}(\beta) \coloneqq Q_{nk}^{\text{LR}}(\beta, \lambda_n)$, and $\overline{\theta}_n(\beta) \coloneqq \overline{\theta}_n^k(\beta, \lambda_n)$.

Proof of Theorem 5.1. We first consider the Wald estimator. We have

$$n[Q_n^{\text{W}}(\beta_n) - Q_n^{\text{W}}(\beta_0)] = 2n^{1/2}[\overline{\theta}_n^k(\beta_0) - \hat{\theta}_n]^\mathsf{T} W_n \Delta_n^k(\beta_n) + \Delta_n^k(\beta_n)^\mathsf{T} W_n \Delta_n^k(\beta_n).$$

For Z_n as defined in (B.1), we see that by Proposition B.1(ii) and R6

$$n^{1/2}[\overline{\theta}_n^k(\beta_0) - \hat{\theta}] = Z_n + n^{1/2}[\theta^k(\beta_0, \lambda_n) - \theta_0] = Z_n + o_p(1), \tag{B.2}$$

whence by Proposition B.1(iii),

$$n[Q_n^{\text{W}}(\beta_n) - Q_n^{\text{W}}(\beta_0)] = 2Z_n^\mathsf{T} W G \delta_n + \delta_n^\mathsf{T} G^\mathsf{T} W G \delta_n + o_p(1 + \|\delta_n\| + \|\delta_n\|^2). \tag{B.3}$$

Now consider the LR estimator. Twice continuous differentiability of the likelihood yields

$$\begin{aligned} n[Q_n^{\text{LR}}(\beta) - Q_n^{\text{LR}}(\beta_0)] &= -n[\mathcal{L}_n(\overline{\theta}_n^k(\beta_n)) - \mathcal{L}_n(\overline{\theta}_n^k(\beta_0))] \\ &= -n^{1/2}\dot{\mathcal{L}}_n(\overline{\theta}_n^k(\beta_0))^\mathsf{T} \Delta_n^k(\beta_n) - \frac{1}{2}\Delta_n^k(\beta_n)^\mathsf{T} \ddot{\mathcal{L}}_n(\overline{\theta}_n^k(\beta_0)) \Delta_n^k(\beta_n) \\ &\quad + o_p(\|\Delta_n^k(\beta_n)\|^2) \end{aligned}$$

where by Proposition B.1(ii) and H4,

$$\begin{aligned} n^{1/2}\dot{\mathcal{L}}_n[\overline{\theta}_n^k(\beta_0)] &= n^{1/2}\dot{\mathcal{L}}_n(\theta_0) + \ddot{\mathcal{L}}_n(\theta_0)n^{1/2}[\overline{\theta}_n^k(\beta_0) - \theta_0] + o_p(1) \\ &= H[Z_n + n^{1/2}(\theta^k(\beta_0, \lambda_n) - \theta_0)] \\ &= HZ_n + o_p(1) \end{aligned} \tag{B.4}$$

for Z_n as in (B.1). Thus by Proposition B.1(iii),

$$n[Q_n^{\mathrm{LR}}(\beta_n) - Q_n^{\mathrm{LR}}(\beta_n)] = -Z_n^{\mathsf{T}}HG\delta_n - \frac{1}{2}\delta_n^{\mathsf{T}}G^{\mathsf{T}}HG\delta_n + o_p(1 + \|\delta_n\| + \|\delta_n\|^2). \tag{B.5}$$

Consistency of $\hat{\beta}_{nk}^e$ follows from parts (i) and (ii) of Proposition B.3 and Corollary 3.2.3 in van der Vaart and Wellner (1996). Thus by applying Theorem 3.2.16 in van der Vaart and Wellner (1996) – or more precisely, the arguments following their (3.2.17) – to (B.3) and (B.5), we have

$$n^{1/2}(\hat{\beta}_{nk}^e - \beta_0) = -(G^{\mathsf{T}}U_eG)^{-1}G^{\mathsf{T}}U_eZ_n + o_p(1) \tag{B.6}$$

for U_e as in (5.11); the result now follows by Proposition B.2. □

Proof of Theorem 5.2. We first note that, in consequence of H3 and Theorem 5.1, $\hat{\beta}_{nk}^e \xrightarrow{p} \beta_0$, $\hat{\theta}_n \xrightarrow{p} \theta_0$, and $\hat{\theta}_n^m \coloneqq \hat{\theta}_n^m(\beta_{nk}^e, \lambda_n) \xrightarrow{p} \theta_0$. Part (i) then follows from R2, H2, and Lemma 2.4 in Newey and McFadden (1994). Defining $\dot{\ell}_i^m(\theta_0) \coloneqq \dot{\ell}_i^m(\beta_0, 0; \theta_0)$ for $m \in \{1, \ldots, M\}$ and

$$\varsigma_i^{\mathsf{T}} \coloneqq \begin{bmatrix} \dot{\ell}_i^0(\theta_0)^{\mathsf{T}} & \dot{\ell}_i^1(\beta_0, 0; \theta_0)^{\mathsf{T}} & \cdots & \dot{\ell}_i^M(\beta_0, 0; \theta_0)^{\mathsf{T}} \end{bmatrix},$$

H2 and H4 further imply that

$$A^{\mathsf{T}}\left(\frac{1}{n}\sum_{i=1}^n s_{ni}s_{ni}^{\mathsf{T}}\right)A \xrightarrow{p} A^{\mathsf{T}}(\mathbb{E}\varsigma_i\varsigma_i^{\mathsf{T}})A = A^{\mathsf{T}}\begin{bmatrix} \Sigma & \mathrm{R} & \cdots & \mathrm{R} \\ \mathrm{R} & \Sigma & \cdots & \mathrm{R} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{R} & \mathrm{R} & \cdots & \Sigma \end{bmatrix} A = V.$$

Part (iii) is an immediate consequence of Proposition B.4(i). □

Proof of Theorem 5.3. We first prove part (i). Let $\dot{Q}_n^e(\beta) \coloneqq \partial_\beta Q_n^e(\beta)$ and $\dot{Q}^e(\beta) \coloneqq \partial_\beta Q^e(\beta, 0)$. Since $\beta_0 \in \operatorname{int}\mathrm{B}$ and $Q_n^e(\beta) \xrightarrow{p} Q(\beta)$ uniformly on B, the global minimum of Q_n^e is interior to B, w.p.a.1., whence R_{nk}^e is non-empty w.p.a.1. Letting $\{\tilde{\beta}_n\}$ denote a (random) sequence with $\tilde{\beta}_n \in R_{nk}^e$ for all n sufficiently large, we have by Proposition B.4(iii) that

$$\dot{Q}^e(\tilde{\beta}_n) = \dot{Q}_n^e(\tilde{\beta}_n) + o_p(1) = o_p(1 + c_n) = o_p(1). \tag{B.7}$$

Since $\dot{Q}^e$ is continuous and B compact, it follows that $d(\tilde{\beta}_n, R^e) \xrightarrow{p} 0$, whence $d_L(R_{nk}^e, R^e) \xrightarrow{p} 0$.

We now turn to part (ii). Recall $\beta_n = \beta_0 + n^{1/2}\delta_n$ for some $\delta_n = o_p(n^{1/2})$. For the Wald criterion, taking δ_n such that $\beta_n \in R_{nk}^{\mathrm{W}}$ gives

$$o_p(1) = n^{1/2}\dot{Q}_n^{\mathrm{W}}(\beta_n)^{\mathsf{T}} = 2[n^{1/2}(\overline{\theta}_n^k(\beta_n) - \hat{\theta}_n)]^{\mathsf{T}}WG_n(\beta_n)$$

where, for Z_n as in (B.1),

$$n^{1/2}(\overline{\theta}_n^k(\beta_n) - \hat{\theta}_n) = n^{1/2}(\overline{\theta}_n^k(\beta_0) - \hat{\theta}_n) + \Delta_n^k(\beta_n) = Z_n + G\delta_n + o_p(1 + \|\delta_n\|)$$

by (B.2), R6, and parts (ii) and (iii) of Proposition B.1. Hence, using Proposition B.4(i),

$$o_p(1) = 2[\delta_n^\mathsf{T} G^\mathsf{T} W G + Z_n^\mathsf{T} W G] + o_p(1 + \|\delta_n\|). \tag{B.8}$$

Similarly, for the LR criterion, taking $\beta_n \in R_{nk}^{\mathrm{LR}}$ in this case gives

$$o_p(1) = n^{1/2}\partial_\beta Q_n^{\mathrm{LR}}(\beta_n)^\mathsf{T} = n^{1/2}\dot{\mathcal{L}}_n[\overline{\theta}_n^k(\beta_n)]^\mathsf{T} G_n(\beta_n)$$

where by the twice continuous differentiability of the likelihood, Proposition B.1(iii) and (B.4),

$$\begin{aligned} n^{1/2}\dot{\mathcal{L}}_n[\overline{\theta}_n^k(\beta_n)] &= n^{1/2}\dot{\mathcal{L}}_n[\overline{\theta}_n^k(\beta_0)] + \ddot{\mathcal{L}}_n(\overline{\theta}_n^k(\beta_0))\Delta_n^k(\beta_n) + o_p(\|\Delta_n^k(\beta_n)\|) \\ &= HZ_n + HG\delta_n + o_p(1 + \|\delta_n\|). \end{aligned}$$

Thus by Proposition B.4(i),

$$o_p(1) = \delta_n^\mathsf{T} G^\mathsf{T} H G + Z_n^\mathsf{T} H G + o_p(1 + \|\delta_n\|). \tag{B.9}$$

By specializing (B.8) and (B.9) to the case where $\delta_n = n^{1/2}(\tilde{\beta}_{nk}^e - \beta_0)$, for $\tilde{\beta}_{nk}^e$ satisfying the requirements of part (ii) of the theorem, we see that for U_e as in (5.11),

$$n^{1/2}(\tilde{\beta}_{nk}^e - \beta_0) = -(G^\mathsf{T} U_e G)^{-1} G^\mathsf{T} U_e Z_n + o_p(1) = n^{1/2}(\hat{\beta}_{nk}^e - \beta_0) + o_p(1)$$

for $e \in \{\mathrm{W}, \mathrm{LR}\}$, in consequence of (B.6).

Finally, we turn to part (iii). Let $\hat{\beta}_n$ denote the minimizer of $Q_n^e(\beta)$, which lies in R_{nk} w.p.a.1., by part (i), and $\tilde{\beta}_n$ another (random) sequence satisfying the requirements of part (iii). By Proposition B.3 and the consistency of $\hat{\beta}_n$ (Theorem 5.1),

$$Q^e(\beta_0) + o_p(1) = Q_n^e(\hat{\beta}_n) + o_p(1) \geq Q_n^e(\tilde{\beta}_n) \geq Q_n^e(\hat{\beta}_n) = Q^e(\beta_0) + o_p(1).$$

Thus $Q^e(\tilde{\beta}_n) = Q_n^e(\tilde{\beta}_n) + o_p(1) \xrightarrow{p} Q^e(\beta_0)$, also by Proposition B.3; whence $\tilde{\beta}_n \xrightarrow{p} \beta_0$, since Q^e has a well-separated minimum at β_0. □

Proof of Theorem 5.4. Let $\ddot{Q}_n^e(\beta) \coloneqq \partial_\beta^2 Q_n^e(\beta)$, $\ddot{Q}^e(\beta) \coloneqq \partial_\beta^2 Q^e(\beta, 0)$, and $\{\tilde{\beta}_n\}$ be a (random) sequence with $\tilde{\beta}_n \in S_{nk}^e$ for all n sufficiently large. Then by Proposition B.4(iii),

$$\mathbb{P}\{\varrho_{\min}[\ddot{Q}^e(\tilde{\beta}_n)] < -\epsilon\} = \mathbb{P}\{\varrho_{\min}[\ddot{Q}_n^e(\tilde{\beta}_n)] + o_p(1) < -\epsilon\} \leq \mathbb{P}\{o_p(1) < -\epsilon\} \to 0$$

for any $\epsilon > 0$. Hence by Theorem 5.3(i), the continuity of $\dot{Q}^e$ and $\ddot{Q}^e$, $d(\tilde{\beta}_n, S^e) \xrightarrow{p} 0$. Part (ii) follows immediately from the corresponding part of Theorem 5.3 and the fact that $S_{nk}^e \subseteq R_{nk}^e$. □

Proof of Theorem 5.5. For each $r \in \{\mathrm{GN}, \mathrm{QN}, \mathrm{TR}\}$, suppose that there exists a $\mathrm{B}_0 \subseteq \mathrm{B}$ such that

$U = Q_n^e(\beta) \coloneqq Q_{nk}^e(\beta, \lambda_n)$ satisfies the corresponding part of Proposition B.5, w.p.a.1. Then

$$\mathbb{P}\{\overline{\beta}_{nk}^e(\beta^{(0)}, r) \in R_{nk}^e \cap \mathrm{B}_0,\ \forall \beta^{(0)} \in \mathrm{B}_0\} \xrightarrow{p} 0$$

for $r \in \{\mathrm{GN}, \mathrm{QN}\}$, and also for $r = \mathrm{TR}$ with S_{nk}^e in place of R_{nk}^e; we may take $c_n = o_p(n^{-1/2})$ in the definition R_{nk}^e. Further, $R^e \cap \mathrm{B}_0 = \{\beta_0\}$ under GN and QN, while $S^e \cap \mathrm{B}_0 = \{\beta_0\}$ under TR. Thus, when $r \in \{\mathrm{GN}, \mathrm{QN}\}$ we have w.p.a.1,

$$\sup_{\beta^{(0)} \in \mathrm{B}_0} d(\overline{\beta}_{nk}^e(\beta^{(0)}, r), \beta_0) \leq d_L(R_{nk}^e \cap \mathrm{B}_0, \{\beta_0\}) = o_p(n^{-1/2})$$

with the final estimate following by Theorem 5.3. When $r = \mathrm{TR}$, the preceding holds with S_{nk}^e in place of R_{nk}^e, in this case via Theorem 5.4.

It thus remains to verify that the requirements of Proposition B.5 hold w.p.a.1. When $r = \mathrm{GN}$, it follows from Proposition B.4(i), the continuity of $\sigma_{\min}(\cdot)$ and GN that

$$0 < \inf_{\beta \in \mathrm{B}_0} \sigma_{\min}[G(\beta)] = \inf_{\beta \in \mathrm{B}_0} \sigma_{\min}[G_n(\beta)] + o_p(1),$$

whence $\inf_{\beta \in \mathrm{B}_0} \sigma_{\min}[G_n(\beta)] > 0$ w.p.a.1. When $r = \mathrm{QN}$, it follows from Proposition B.4(iii) and QN that

$$0 < \inf_{\beta \in \mathrm{B}_0} \varrho_{\min}[\partial_\beta^2 Q^e(\beta)] = \inf_{\beta \in \mathrm{B}_0} \varrho_{\min}[\partial_\beta^2 Q_n^e(\beta)] + o_p(1)$$

whence Q_n^e is strictly convex on B_0 w.p.a.1. When $r = \mathrm{TR}$, there are no additional conditions to verify. □

B.3 Proofs of Propositions B.1–B.5

Proof of Proposition B.1. Part (i) follows by H3 and the continuous mapping theorem. Part (ii) is immediate from (4.3). For part (iii), we note that for $\beta_n = \beta_0 + n^{1/2}\delta_n$ with $\delta_n = o_p(n^{1/2})$ as above,

$$\begin{aligned}\Delta_n^k(\beta_n) = n^{1/2}[\overline{\theta}_n^k(\beta_n, \lambda_n) - \theta^k(\beta_n, \lambda_n)] \\ - n^{1/2}[\overline{\theta}_n^k(\beta_0, \lambda_n) - \theta^k(\beta_0, \lambda_n)] + n^{1/2}[\theta^k(\beta_n, \lambda_n) - \theta^k(\beta_0, \lambda_n)].\end{aligned}$$

Since $\overline{\theta}_n^k$ is a linear combination of the $\hat{\theta}_n^m$'s, it is clear from H3 that the first two terms converge jointly in distribution to identical limits (since $\beta_n \xrightarrow{p} \beta_0$). For the final term, continuous differentiability of θ^k (R3 above) entails that

$$\begin{aligned}n^{1/2}[\theta^k(\beta_n, \lambda_n) - \theta^k(\beta_0, \lambda_n)] &= [\partial_\beta \theta^k(\beta_0, \lambda_n)]^{\mathsf{T}}(\beta_n - \beta_0) + o_p(\|\beta_n - \beta_0\|) \\ &= G\delta_n + o_p(1 + \|\delta_n\|).\end{aligned}$$

□

Proof of Proposition B.2. Note first that

$$n^{1/2}[\overline{\theta}_n^k(\beta_0,\lambda_n)-\theta^k(\beta_0,\lambda_n)]=\sum_{r=0}^{k}\gamma_{rk}\cdot n^{1/2}[\overline{\theta}_n(\beta_0,\delta^r\lambda_n)-\theta(\beta_0,\delta^r\lambda_n)]$$
$$=\frac{1}{M}\sum_{m=1}^{M}\sum_{r=0}^{k}\gamma_{rk}\psi_n^m(\beta_0,\delta^r\lambda_n)\rightsquigarrow\frac{1}{M}\sum_{m=1}^{M}\psi^m(\beta_0,0),$$

by (4.3), (4.4), H3 and $\sum_{r=0}^{k}\gamma_{rk}=1$. By H3, this holds jointly with

$$n^{1/2}(\hat{\theta}_n-\theta_0)\rightsquigarrow\psi^0(\beta_0,0).$$

Since H4 implies that $\psi^m(\beta_0,0)=H^{-1}\phi^m$, the limiting variance of Z_n is equal to

$$\operatorname{var}\left[\psi^0(\beta_0,0)-\frac{1}{M}\sum_{m=1}^{M}\psi^m(\beta_0,0)\right]=H^{-1}\operatorname{var}\left[\phi^0-\frac{1}{M}\sum_{m=1}^{M}\phi^m\right]H^{-1}=H^{-1}VH^{-1}$$

where the final equality follows from H4 and straightforward calculations. □

Proof of Proposition B.3. We first prove part (i). For the Wald estimator, this is immediate from Proposition B.1(i). For the LR estimator, it follows from Proposition B.1(i), H2 and the continuous mapping theorem (arguing as on pp. 144f. of Billingsley, 1968), that

$$Q_{nk}^{\mathrm{LR}}(\beta)=(\mathcal{L}_n\circ\overline{\theta}_n^k)(\beta,\lambda_n)\xrightarrow{p}(\mathcal{L}\circ\theta^k)(\beta,0)=Q^{\mathrm{LR}}(\beta),$$

uniformly on B.

For part (ii), we note that $\beta\mapsto\theta^k(\beta,0)$ is continuous by R3, while the continuity of $\mathcal{L}$ is implied by H2, since $\mathcal{L}_n$ is continuous. Thus Q^e is continuous for $e\in\{\mathrm{W},\mathrm{LR}\}$, and by R4 is uniquely minimized at β_0. Hence $\beta\mapsto Q^e(\beta)$ has a well-separated minimum, which by R1 is interior to B. □

Proof of Proposition B.4. Part (ii) is immediate from H5, (4.4) and the continuous mapping theorem; it further implies part (i). For part (iii), recall $\dot{Q}_n^e(\beta)=\partial_\beta Q_n^e(\beta)$, and $G_n(\beta)=[\partial_\beta\overline{\theta}_n^k(\beta)]^{\mathsf{T}}$. Then we have

$$\dot{Q}_n^{\mathrm{W}}(\beta)=G_n(\beta)^{\mathsf{T}}W_n[\overline{\theta}_n(\beta)-\hat{\theta}_n]\qquad\qquad\dot{Q}_n^{\mathrm{LR}}(\beta)=G_n(\beta)^{\mathsf{T}}\dot{\mathcal{L}}_n[\overline{\theta}_n^k(\beta)].$$

Part (i), and similar arguments as were used are used in the proof of part (i) of Proposition B.3, yield that $\dot{Q}_n^e(\beta)\xrightarrow{p}\partial_\beta Q^e(\beta,0)=:\dot{Q}^e(\beta)$ uniformly on B. The proof that the second derivatives converge uniformly is analogous. □

Proof of Proposition B.5. For $r=\mathrm{GN}$, the result follows by Theorem 10.1 in Nocedal and Wright (2006); for $r=\mathrm{QN}$, by their Theorem 6.5; and for $r=\mathrm{TR}$, by Theorem 4.13 in Moré and Sorensen (1983). □

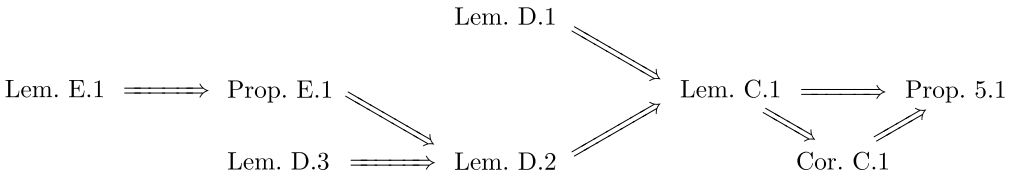

Figure C.1: Proof of Proposition 5.1

C Sufficiency of the low-level assumptions

We shall henceforth maintain both Assumptions L and R, and address the question of whether these are sufficient for Assumption H; that is, we shall prove Proposition 5.1. The main steps leading to the proof are displayed in Figure C.1.

Recall that, as per L8, the auxiliary model is the Gaussian SUR displayed in (5.1) above. For simplicity, we shall consider only the case where Σ_ξ is unrestricted, but our arguments extend straightforwardly to the case where Σ_ξ is block diagonal (as would typically be imposed when $T > 1$). Recall that θ collects the elements of α and Σ_ξ^{-1}. Fix an $m \in \{0, 1, \dots M\}$, and define

$$\xi_{ri}(\alpha) \coloneqq y_r(z_i; \beta, \lambda) - \alpha_{xr}^\mathsf{T} \Pi_{xr} x(z_i) - \alpha_{yr}^\mathsf{T} \Pi_{yr} y(z_i; \beta, \lambda),$$

temporarily suppressing the dependence of y (and hence ξ_{ri}) on m. Collecting $\xi_i \coloneqq (\xi_{1i}, \dots, \xi_{d_y i})^\mathsf{T}$, the average log-likelihood of the auxiliary model can be written as

$$\mathcal{L}_n(y, x; \theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, x_i; \theta) = -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \det \Sigma_\xi - \frac{1}{2} \operatorname{tr} \left[\Sigma_\xi^{-1} \frac{1}{n} \sum_{i=1}^n \xi_i(\alpha) \xi_i(\alpha)^\mathsf{T} \right].$$

Deduce that there are functions L and l, which are three times continuously differentiable in both arguments (at least on $\operatorname{int} \Theta$), such that

$$\mathcal{L}_n(y, x; \theta) = L(T_n; \theta) \qquad \ell(y_i, x_i; \theta) = l(t_i; \theta) \tag{C.1}$$

where

$$t_i^m(\beta, \lambda) = \begin{bmatrix} y(z_i^m; \beta, \lambda) \\ x(z_i^m) \end{bmatrix} \begin{bmatrix} y(z_i^m; \beta, \lambda)^\mathsf{T} & x(z_i^m)^\mathsf{T} \end{bmatrix}$$

and $T_n^m \coloneqq \operatorname{vech}(\mathcal{T}_n^m)$, for

$$\mathcal{T}_n^m(\beta, \lambda) \coloneqq \frac{1}{n} \sum_{i=1}^n t_i^m(\beta, \lambda) t_i^m(\beta, \lambda)^\mathsf{T}. \tag{C.2}$$

Further, direct calculation gives

$$\partial_{\alpha_{xr}} \ell_i(\theta) = \sum_{s=1}^{d_y} \sigma^{rs} \xi_{si}(\alpha) \Pi_{xr} x(z_i) \qquad \partial_{\alpha_{yr}} \ell_i(\theta) = \sum_{s=1}^{d_y} \sigma^{rs} \xi_{si}(\alpha) \Pi_{yr} y(z_i; \beta, \lambda) \tag{C.3}$$

and

$$\partial_{\sigma^{rs}} \ell_i(\theta) = \frac{1}{2} \sigma_{rs} - \frac{1}{2} \xi_{ri}(\alpha) \xi_{si}(\alpha). \tag{C.4}$$

Since the elements of the score vector $\dot{\ell}_i(\theta) = \partial_\theta \ell_i(\theta)$ necessarily take one of the forms displayed in (C.3) or (C.4), we may conclude that, for any compact subset $A \subset \Theta$, there exists a C_A such that

$$\mathbb{E} \sup_{\theta \in A} \|\dot{\ell}_i(\theta)\|^2 \leq C_A \mathbb{E}\|z_i\|^4 < \infty \tag{C.5}$$

with the second inequality following from L6.

Regarding the maximum likelihood estimator (MLE), we note that the concentrated average log-likelihood is given by

$$\mathcal{L}_n(y, x; \alpha) = -\frac{d_y}{2}(\log 2\pi + 1) - \frac{1}{2} \log \det \left[\frac{1}{n} \sum_{i=1}^{n} \xi_i(\alpha) \xi_i(\alpha)^\mathsf{T} \right] = L_c(T_n; \alpha)$$

which is three times continuously differentiable in α and T_n, so long as $\mathcal{T}_n$ is non-singular. By the implicit function theorem, it follows that $\hat{\alpha}_n$ may be regarded as a smooth function of T_n. Noting the usual formula for the ML estimates of Σ_ξ, this holds also for the components of θ referring to Σ_ξ^{-1}, whence

$$\hat{\theta}_n^m(\beta, \lambda) = h[T_n^m(\beta, \lambda)] \tag{C.6}$$

for some h that is twice continuously differentiable on the set where $\mathcal{T}_n^m$ has full rank. Under L7, this occurs uniformly on $\mathrm{B} \times \Lambda$ w.p.a.1., and so to avoid tiresome circumlocution, we shall simply treat h as if it were everywhere twice continuously differentiable throughout the sequel. Letting $T(\beta, \lambda) \coloneqq \mathbb{E} T_n^0(\beta, \lambda)$, we note that the population binding function is given by

$$\theta(\beta, \lambda) = h[T(\beta, \lambda)]. \tag{C.7}$$

Define $\varphi_n^m(\beta, \lambda) \coloneqq n^{1/2}[T_n^m(\beta, \lambda) - T(\beta, \lambda)]$, and let $[\varphi^m(\beta, \lambda)]_{m=0}^M$ denote a vector-valued continuous Gaussian process on $\mathrm{B} \times \Lambda$ with covariance kernel

$$\operatorname{cov}(\varphi^{m_1}(\beta_1, \lambda_1), \varphi^{m_2}(\beta_2, \lambda_2)) = \operatorname{cov}(T_n^{m_1}(\beta_1, \lambda_1), T_n^{m_2}(\beta_2, \lambda_2)).$$

Note that L6, in particular the requirement that $\mathbb{E}\|z_i\|^4 < \infty$, ensures that this covariance exists and is finite.

Lemma C.1.

(i) $\varphi_n^m(\beta, \lambda) \rightsquigarrow \varphi^m(\beta, \lambda)$ *in* $\ell^\infty(\mathrm{B} \times \Lambda)$, *jointly for* $m \in \{0, \ldots, M\}$; *and*

(ii) *if (5.7) holds for* $l' = l \in \{1, 2\}$, *then*

$$\sup_{\beta \in \mathrm{B}} \|\partial_\beta^l T_n^m(\beta, \lambda_n) - \partial_\beta^l T(\beta, 0)\| = o_p(1) \tag{C.8}$$

By an application of the delta method, we thus have

Corollary C.1. *For* $\dot{h}(\beta, \lambda) \coloneqq \partial_\beta h[T(\beta, \lambda)]$,

$$\psi_n^m(\beta, \lambda) \coloneqq n^{1/2}[\hat{\theta}_n^m(\beta, \lambda) - \theta(\beta, \lambda)] \rightsquigarrow \dot{h}(\beta, \lambda) \varphi^m(\beta, \lambda) =: \psi^m(\beta, \lambda) \tag{C.9}$$

in $\ell^\infty(\mathrm{B} \times \Lambda)$, *jointly for* $m \in \{0, \ldots, M\}$.

The proof of Lemma C.1 appears in Appendix D.

Proof of Proposition 5.1. H1 follows from the twice continuous differentiability of L in (C.1). The first part of H2 is an immediate consequence of Lemma C.1(i) and the smoothness of L; the second part is implied by (C.5) and Lemma 2.4 in Newey and McFadden (1994). H3 follows from Corollary C.1, and immediately entails that, for $\beta_n = \beta_0 + o_p(1)$ and $m \in \{1, \dots, M\}$, $\psi_n^m(\beta_n, \lambda_n) = \psi_n^m(\beta_0, 0) + o_p(1)$, where

$$\psi_n^m(\beta_0, 0) = n^{1/2}[\hat{\theta}_n^m(\beta_0, 0) - \theta(\beta_0, 0)] = -H^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^{n} \dot{\ell}_i^m(\beta_0, 0; \theta_0) + o_p(1)$$

for $m \in \{0, 1, \dots, M\}$; the final equality follows from the consistency of $\hat{\theta}_n$ (as implied by Corollary C.1) and the arguments used to prove Theorem 3.1 in Newey and McFadden (1994). By definition, $\phi_n^m \coloneqq n^{-1/2}\sum_{i=1}^n \dot{\ell}_i^m(\beta_0, 0; \theta_0)$ whereupon the rest of H4 follows by the central limit theorem, in view of L1 and (C.5). Finally, H5 follows from (C.6), (C.7), Lemma C.1(ii) and the chain rule. □

D Proof of Lemma C.1

For the purposes of the proofs undertaken in this section, we may suppose without loss of generality that $\tilde{D} = I_{d_y}$ in L2, $\gamma(\beta) = \beta$ in L3, and $\|K\|_\infty \leq 1$. Recalling (5.3) above, we have

$$y_r(\beta, \lambda) = \omega_r(\beta) \cdot \prod_{s \in \mathcal{S}_r} K_\lambda[\nu_s(\beta)] =: \omega_r(\beta) \cdot \mathbb{K}(\mathcal{S}_r; \beta, \lambda). \tag{D.1}$$

Let $\dot{K}$ and $\ddot{K}$ respectively denote the first and second derivatives of K. For future reference, we here note that

$$\begin{aligned} \partial_\beta y_r(\beta, \lambda) &= z_{wr} \cdot \mathbb{K}(\mathcal{S}_r; \beta, \lambda) + \lambda^{-1} w_r(\beta) \sum_{s \in \mathcal{S}_r} z_{vs} \cdot \mathbb{K}_s(\mathcal{S}_r; \beta, \lambda) \\ &=: D_{r1}(\beta, \lambda) + \lambda^{-1} D_{r2}(\beta, \lambda) \end{aligned} \tag{D.2}$$

where $z_{vr} \coloneqq \Pi_{vr}^\mathsf{T} z$, $z_{wr} \coloneqq \Pi_{wr}^\mathsf{T} z$ and $\mathbb{K}_s(\mathcal{S}; \beta, \lambda) \coloneqq \dot{K}_\lambda[v_s(\beta)] \cdot \mathbb{K}(\mathcal{S}\backslash\{s\}; \beta, \lambda)$; and

$$\begin{aligned} \partial_\beta^2 y_r(\beta, \lambda) &= \lambda^{-1} \sum_{s \in \mathcal{S}_r} [z_{wr} z_{vs}^\mathsf{T} + z_{vs} z_{wr}^\mathsf{T}] \cdot \mathbb{K}_s(\mathcal{S}_r; \beta, \lambda) \\ &\qquad + \lambda^{-2} w_r(\beta) \sum_{s \in \mathcal{S}_r} \sum_{t \in \mathcal{S}_r} z_{vs} z_{vt}^\mathsf{T} \cdot \mathbb{K}_{st}(\mathcal{S}_r; \beta, \lambda) \\ &=: \lambda^{-1} H_{r1}(\beta, \lambda) + \lambda^{-2} H_{r2}(\beta, \lambda) \end{aligned} \tag{D.3}$$

for

$$\mathbb{K}_{st}(\mathcal{S}; \beta, \lambda) \coloneqq \begin{cases} \ddot{K}_\lambda[v_s(\beta)] \cdot \mathbb{K}(\mathcal{S}\backslash\{s\}; \beta, \lambda) & \text{if } s = t, \\ \dot{K}_\lambda[v_s(\beta)] \cdot \dot{K}_\lambda[v_t(\beta)] \cdot \mathbb{K}(\mathcal{S}\backslash\{s, t\}; \beta, \lambda) & \text{if } s \neq t. \end{cases}$$

D.1 Proof of part (ii)

In view of (C.2), the scalar elements of $T_n(\beta, \lambda)$ that depend on (β, λ) take either of the following forms:

$$\tau_{n1}(\beta, \lambda) \coloneqq \mathbb{E}_n[y_r(\beta, \lambda) y_s(\beta, \lambda)] \qquad \tau_{n2}(\beta, \lambda) \coloneqq \mathbb{E}_n[y_r(\beta, \lambda) x_t] \tag{D.4}$$

for some $r, s \in \{1, \ldots, d_y\}$, or $t \in \{1, \ldots, d_x\}$, where $\mathbb{E}_n f(\beta, \lambda) \coloneqq \frac{1}{n} \sum_{i=1}^n f(z_i; \beta, \lambda)$. (Throughout the following, all statements involving r, s and t should be interpreted as holding for all possible values of these indices.) For $k \in \{1, 2\}$ and $l \in \{0, 1, 2\}$, define $\tau_k(\beta, \lambda) \coloneqq \mathbb{E}\tau_{nk}(\beta, \lambda)$ – a typical scalar element of $T(\beta, \lambda)$ – and $\tau_k^{[l]}(\beta, \lambda) \coloneqq \mathbb{E}\partial_\beta^l \tau_{nk}(\beta, \lambda)$. Thus part (ii) of Lemma C.1 will follow once we have shown that

$$\partial_\beta^l \tau_{nk}(\beta, \lambda_n) = \tau_k^{[l]}(\beta, \lambda_n) + o_p(1) = \partial_\beta^l \tau_k(\beta, 0) + o_p(1) \tag{D.5}$$

uniformly in $\beta \in \mathrm{B}$. The second equality in (D.5) is implied by

Lemma D.1. $\tau_k^{[l]}(\beta, \lambda_n) \xrightarrow{p} \partial_\beta^l \tau_k(\beta, 0)$, *uniformly on* B, *for* $k \in \{1, 2\}$ *and* $l \in \{0, 1, 2\}$.

The proof appears at the end of this section. We turn next to the first equality in (D.5). We require the following definitions. A function $F : \mathscr{Z} \mapsto \mathbb{R}$ is an *envelope* for the class $\mathcal{F}$ if $\sup_{f \in \mathcal{F}} |f(z)| \leq F(z)$. For a probability measure $\mathbb{Q}$ and a $p \in (1, \infty)$, let $\|f\|_{p,\mathbb{Q}} \coloneqq (\mathbb{E}_\mathbb{Q} |f(z_i)|^p)^{1/p}$. $\mathcal{F}$ is *Euclidean* for the envelope F if

$$\sup_{\mathbb{Q}} N(\epsilon \|F\|_{1,\mathbb{Q}}, \mathcal{F}, L_{1,\mathbb{Q}}) \leq C_1 \epsilon^{-C_2}$$

for some C_1 and C_2 (depending on $\mathcal{F}$), where $N(\epsilon, \mathcal{F}, L_{1,\mathbb{Q}})$ denotes the minimum number of $L_{1,\mathbb{Q}}$-balls of diameter ϵ needed to cover $\mathcal{F}$. For a parametrized family of functions $g(\beta, \lambda) = g(z; \beta, \lambda) : \mathscr{Z} \mapsto \mathbb{R}^{d_1 \times d_2}$, let $\mathcal{F}(g) \coloneqq \{g(\beta, \lambda) \mid (\beta, \lambda) \in \mathrm{B} \times \Lambda\}$. Since B is compact, we may suppose without loss of generality that $\mathrm{B} \subseteq \{\beta \in \mathbb{R}^{d_\beta} \mid \|\beta\| \leq 1\}$, whence recalling (5.2) and (5.4) above,

$$|w_r(z; \beta)| \leq W_r \leq \begin{cases} \|z\| & \text{if } r \in \{1, \ldots d_w\} \\ 1 & \text{if } r \in \{d_w + 1, \ldots d_y\}. \end{cases}$$

Thus by Lemma 22 in Nolan and Pollard (1987)

D1 for $\mathbb{L} \in \{\mathbb{K}, \mathbb{K}_s, \mathbb{K}_{st}\}$, $s, t \in \{1, \ldots, d_y\}$ and $\mathcal{S} \subseteq \{1, \ldots, d_v\}$, the class

$$\mathcal{F}(\mathbb{L}, \mathcal{S}) \coloneqq \{\mathbb{L}(\mathcal{S}; \beta, \lambda) \mid (\beta, \lambda) \in \mathrm{B} \times \Lambda\}$$

is Euclidean with constant envelope; and

D2 for $r \in \{1, \ldots, d_y\}$, $\mathcal{F}(w_r)$ is Euclidean for W_r.

It therefore follows by a slight adaptation of the proof of Theorem 9.15 in Kosorok (2008) that

D3 $\mathcal{F}(y_r)$ is Euclidean for W_r;

D4 $\mathcal{F}(y_r D_{s1})$ and $\mathcal{F}(y_r D_{s2})$ are Euclidean for $W_r W_s \|z\|$

D5 $\mathcal{F}(x_t D_{s1})$ and $\mathcal{F}(x_t D_{s2})$ are Euclidean for $W_s\|z\|^2$;

D6 $\mathcal{F}(D_{s1}D_{r1}^\mathsf{T})$, $\mathcal{F}(D_{s1}D_{r2}^\mathsf{T})$, $\mathcal{F}(D_{s2}D_{r1}^\mathsf{T})$ and $\mathcal{F}(D_{s2}D_{r2}^\mathsf{T})$ are Euclidean for $W_rW_s\|z\|^2$;

D7 $\mathcal{F}(y_s H_{r1})$ and $\mathcal{F}(y_s H_{r2})$ are Euclidean for $W_rW_s\|z\|^2$; and

D8 $\mathcal{F}(x_t H_{r1})$ and $\mathcal{F}(x_t H_{r2})$ are Euclidean for $W_s\|z\|^3$.

Let $\mu_n f \coloneqq \frac{1}{n}\sum_{i=1}^n [f(z_i) - \mathbb{E}f(z_i)]$. Using the preceding facts, and the uniform law of large numbers given as Proposition E.1 below, we may prove

Lemma D.2. *The convergence*

$$\sup_{\beta\in\mathrm{B}} \mu_n|\partial_\beta^l[y_s(\beta,\lambda_n)y_r(\beta,\lambda_n)]| + \sup_{\beta\in\mathrm{B}} \mu_n|x_t\partial_\beta^l y_r(\beta,\lambda_n)| = o_p(1). \tag{D.6}$$

holds for $l = 0$, and also for $l \in \{1,2\}$ if (5.7) holds with $l' = l$.

The first equality in (C.8) now follows, and thus part (ii) of Lemma C.1 is proved.

Proof of Lemma D.1. Suppose $l = 2$; the proof when $l = 1$ is analogous (and is trivial when $l = 0$). Noting that

$$\partial_\beta^2(y_r y_s) = y_s\partial_\beta^2 y_r + (\partial_\beta y_r)(\partial_\beta y_s)^\mathsf{T} + (\partial_\beta y_s)(\partial_\beta y_r)^\mathsf{T} + y_r\partial_\beta^2 y_s, \tag{D.7}$$

it follows from (D.2), (D.3), D6 and D7 that for every $\lambda \in (0,1]$,

$$\|\partial_\beta^2(y_r y_s)\| \lesssim \lambda^{-2} W_r W_s(\|z\|^2 \vee 1),$$

which does not depend on β, and is integrable by L6. (Here $a \lesssim b$ denotes that $a \le Cb$ for some constant C not depending on b.) Thus by the dominated derivatives theorem, the second equality in

$$\tau_1^{[2]}(\beta,\lambda) = \mathbb{E}\partial_\beta^2\tau_{n1}(\beta,\lambda) = \partial_\beta^2\mathbb{E}\tau_{n1}(\beta,\lambda) = \partial_\beta^2\tau_1(\beta,\lambda)$$

holds for every $\lambda \in (0,1]$; the other equalities follow from the definitions of $\tau_k^{[l]}$ and τ_k. Deduce that, so long as $\lambda_n > 0$ (as per the requirements of Proposition 5.1 above),

$$\tau_1^{[2]}(\beta,\lambda_n) = \partial_\beta^2\tau_1(\beta,\lambda_n) \xrightarrow{p} \partial_\beta^2\tau_1(\beta,0)$$

by the uniform continuity of $\partial_\beta^2\tau_1$ on $\mathrm{B}\times\Lambda$. A similar reasoning – but now using D8 – gives the same result for $\tau_2^{[2]}$. □

The proof of Lemma D.2 requires the following result. Let $\mathcal{G}_{\omega,x}$ denote the σ-field generated by $\eta_\omega(z_i)$ and $x(z_i)$, and let η_ν denote those elements of η that are not present in η_ω. Recall that $\eta_\nu \perp\!\!\!\perp \mathcal{G}_{\omega,x}$.

Lemma D.3. *For every $p \in \{0,1,2\}$, $s,t \in \{1,\dots,d_v\}$, $\mathcal{S} \subseteq \{1,\dots,d_v\}$ and $\mathbb{L} \in \{\mathbb{K}_s, \mathbb{K}_{st}\}$*

$$\mathbb{E}[\|z_{\nu s}\|^p\|z_{\nu t}\|^p\mathbb{L}(\mathcal{S};\beta,\lambda)^2 \mid \mathcal{G}_{\omega,x}] \lesssim \lambda\mathbb{E}[\|z_{\nu s}\|^p\|z_{\nu t}\|^p \mid \mathcal{G}_{\omega,x}]. \tag{D.8}$$

Proof. Note that for any $\mathbb{L} \in \{\mathbb{K}_s, \mathbb{K}_{st}\}$,

$$\mathbb{L}(\mathcal{S}; \beta, \lambda) \lesssim L_\lambda[\nu_s(\beta)]$$

where $L(x) = \max\{|\dot{K}(x)|, |\ddot{K}(x)|\}$. Let d denote the dimensionality of η_ν, and fix a $\beta \in \mathrm{B}$. By L4 and L5, there is a $k \in \{1, \dots d\}$, possibly depending on β, and an $\epsilon > 0$ which does not, such that

$$\nu_s(\beta) = \nu_s^*(\beta) + \beta_k^* \eta_{\nu k}$$

with $|\beta_k^*| \geq \epsilon$ and $\nu_s^*(\beta) \perp\!\!\!\perp \eta_{\nu k}$. Let $\mathcal{G}_{\omega,x}^* \coloneqq \mathcal{G}_{\omega,x} \vee \sigma(\{\eta_{\nu l}\}_{l \neq k})$, so that $\nu_s^*(\beta)$ is $\mathcal{G}_{\omega,x}^*$-measurable, and let f_k denote the density of $\eta_{\nu k}$. Then for any $q \in \{0, \dots, 4\}$,

$$\begin{aligned}
\mathbb{E}\left[|\eta_{\nu k}|^q \mathbb{L}(\mathcal{S}; \beta, \lambda)^2 \mid \mathcal{G}_{\omega,x}^*\right] &\lesssim \mathbb{E}\left[|\eta_{\nu k}|^q L_\lambda^2(\nu_s^*(\beta) + \beta_k^* \eta_{\nu k}) \mid \mathcal{G}_{\omega,x}^*\right] \\
&= \int_{\mathbb{R}} |u|^q L_\lambda^2(\nu_s^*(\beta) + \beta_k^* u) f_k(u) \, \mathrm{d}u \\
&\lesssim (\beta_k^*)^{-1} \lambda \int_{\mathbb{R}} L^2(u) \, \mathrm{d}u \cdot \sup_{u \in \mathbb{R}} |u|^q f_k(u) \\
&\lesssim \epsilon^{-1} \lambda, \qquad \text{(D.9)}
\end{aligned}$$

since $\sup_{u \in \mathbb{R}} |u|^q f_k(u) < \infty$ under L4. Finally, we may partition $z_{\nu s} = (z_{\nu s}^{*\mathsf{T}}, \eta_{\nu k})^\mathsf{T}$ and $z_{\nu t} = (z_{\nu t}^{*\mathsf{T}}, \eta_{\nu k})^\mathsf{T}$, with the possibility that $z_{\nu s} = z_{\nu s}^*$ and $z_{\nu t} = z_{\nu t}^*$. Then by (D.9),

$$\mathbb{E}\left[\|z_{\nu s}\|^p \|z_{\nu t}\|^p \mathbb{L}(\mathcal{S}; \beta, \lambda)^2 \mid \mathcal{G}_{\omega,x}^*\right] \lesssim \lambda \|z_{\nu s}^*\|^p \|z_{\nu t}^*\|^p \leq \lambda \|z_{\nu s}\|^p \|z_{\nu t}\|^p.$$

The result now follows by the law of iterated expectations. □

Proof of Lemma D.2. We shall only provide the proof for first term on the left side of (D.6), when $l = 2$; the proof in all other cases are analogous, requiring appeal only to Proposition E.1 (or Theorem 2.4.3 in van der Vaart and Wellner, 1996, when $l = 0$) and the appropriate parts of D3–D8.

Recalling the decomposition of $\partial_\beta^2(y_r y_s)$ given in (D.7) above, we are led to consider

$$(\partial_\beta y_r)(\partial_\beta y_s)^\mathsf{T} = D_{s1} D_{r1}^\mathsf{T} + \lambda^{-1} D_{s2} D_{r1}^\mathsf{T} + \lambda^{-1} D_{s1} D_{r2}^\mathsf{T} + \lambda^{-2} D_{s2} D_{r2}^\mathsf{T} \qquad \text{(D.10)}$$

and

$$y_s \partial_\beta^2 y_r = \lambda^{-1} y_s H_{r1} + \lambda^{-2} y_s H_{r2}. \qquad \text{(D.11)}$$

Note that by Lemma D.3, and L6

$$\begin{aligned}
\mathbb{E}\|y_s H_{r2}\|^2 &\lesssim \mathbb{E}\left[|\omega_s(\beta)|^2 |\omega_r(\beta)|^2 \sum_{s \in \mathcal{S}_r} \sum_{t \in \mathcal{S}_r} \mathbb{E}\left[\|z_{vs}\|^2 \|z_{vt}\|^2 |\mathbb{K}_{st}(\mathcal{S}_r; \beta, \lambda)|^2 \mid \mathcal{G}_{\omega,x}\right]\right] \\
&\lesssim \lambda \mathbb{E}\left[W_s^2 W_r^2 \sum_{s \in \mathcal{S}_r} \sum_{t \in \mathcal{S}_r} \mathbb{E}\|z_{vs}\|^2 \|z_{vt}\|^2\right] \\
&\lesssim \lambda
\end{aligned}$$

and analogously for each of H_{r1}, $D_{s1} D_{r1}^\mathsf{T}$, $D_{s2} D_{r1}^\mathsf{T}$, $D_{s1} D_{r2}^\mathsf{T}$ and $D_{s2} D_{r2}^\mathsf{T}$. By D6 and D7, the classes

formed from these parametrized functions are Euclidean, with envelopes that are p_0-integrable under L6 ($p_0 \geq 2$).

Application of Proposition E.1 to each of the terms in D6 and D7, with λ playing the role of δ^{-1} there, thus yields the result. Negligibility of the final terms in (D.10) and (D.11) entail the most stringent conditions on the rate at which λ_n may shrink to zero, due to the multiplication of these by λ^{-2}. □

D.2 Proof of part (i)

The typical scalar elements of T_n are as displayed in (D.4) above, i.e. they are averages of random functions of the form $\zeta_1(\beta, \lambda) := y_r(\beta, \lambda) y_s(\beta, \lambda)$ or $\zeta_2(\beta, \lambda) := x_t y_r(\beta, \lambda)$, for $r, s \in \{1, \ldots, d_y\}$ and $t \in \{1, \ldots, d_x\}$. It follows from D3 that $\mathcal{F}(\zeta_1)$ and $\mathcal{F}(\zeta_2)$ are Euclidean, with envelopes $F_1 := W_r W_s$ and $F_2 := \|z\| W_r$ respectively. Since both envelopes are square integrable under L6, we have

$$\sup_{\mathbb{Q}} N(\epsilon \|F_k\|_{2,\mathbb{Q}}, \mathcal{F}(\zeta_k), L_{2,\mathbb{Q}}) \leq C_1' \epsilon^{-C_2'}$$

for $k \in \{1, 2\}$. Hence (C.9) follows by Theorem 2.5.2 in van der Vaart and Wellner (1996).

E A uniform-in-bandwidth law of large numbers

This section provides a uniform law of large numbers (ULLN) for certain classes of parametrized functions, broad enough to cover products involving $K_\lambda[\nu_s(\beta)]$, and such generalizations as appear in Lemma D.3 above. Our ULLN holds *uniformly* in the inverse 'bandwith' parameter $\delta = \lambda^{-1}$; in this respect, it is related to some of the results proved in Einmahl and Mason (2005). However, while their arguments could be adapted to our problem, these would lead to stronger conditions on the bandwidth: in particular, p would have to be replaced by $2p$ in Proposition E.1 below. (On the other hand, their results yield explicit rates of uniform convergence, which are not of concern here.)

Consider the (pointwise measurable) function class

$$\mathcal{F}_\Delta := \{z \mapsto f_{(\gamma,\delta)}(z) \mid (\gamma, \delta) \in \Gamma \times \Delta\},$$

and put $\mathcal{F} := \mathcal{F}_{[1,\infty)}$. The functions $f_{(\gamma,\delta)} : \mathscr{Z} \to \mathbb{R}^d$ satisfy:

E1 $\sup_{\gamma \in \Gamma} \mathbb{E}\|f_{(\gamma,\delta)}(z_0)\|^2 \lesssim \delta^{-1}$ for every $\delta > 0$.

Let $F : \mathscr{Z} \to \mathbb{R}$ denote an envelope for $\mathcal{F}$, in the sense that

$$\sup_{(\gamma,\delta) \in \Gamma \times [1,\infty)} \|f_{(\gamma,\delta)}(z)\| \leq F(z)$$

for all $z \in \mathscr{Z}$. We will suppose that F may be chosen such that, additionally,

E2 $\mathbb{E}|F(z_0)|^p < \infty$; and

E3 $\sup_{\mathbb{Q}} N(\epsilon \|F\|_{1,\mathbb{Q}}, \mathcal{F}, L_{1,\mathbb{Q}}) \leq C\epsilon^{-d}$ for some $d \in (0, \infty)$.

Let $\{\overline{\delta}_n\}$ denote a real sequence with $\overline{\delta}_n \geq 1$, and $\Delta_n := [1, \overline{\delta}_n]$.

Proposition E.1. *Under* E1–E3*, if* $n^{1-1/p}/\overline{\delta}_n^{2m-1}\log(\overline{\delta}_n \vee n) \to \infty$ *for some* $m \geq 1$*, then*

$$\sup_{(\gamma,\delta)\in\Gamma\times\Delta_n} \delta^m \|\mu_n f_{(\gamma,\delta)}\| = o_p(1). \tag{E.1}$$

Remark E.1. Suppose δ_n is an $\mathcal{F}$-measurable sequence for which $n^{1-1/p}/\delta_n^{2m-1}\log(\delta_n \vee n) \xrightarrow{p} \infty$. Then for every $\epsilon > 0$, there exists a deterministic sequence $\{\overline{\delta}_n\}$ satisfying the requirements of Proposition E.1, and for which $\limsup_{n\to\infty} \mathbb{P}\{\delta_n \leq \overline{\delta}_n\} > 1 - \epsilon$. Deduce that

$$\sup_{\gamma\in\Gamma} \delta_n^m \|\mu_n f_{(\gamma,\delta_n)}\| = o_p(1).$$

The proof requires the following

Lemma E.1. *Suppose* $\mathcal{F}$ *is a (pointwise measurable) class with envelope* F*, satisfying*

(i) $\|F\|_\infty \leq \tau$*;*

(ii) $\sup_{f\in\mathcal{F}} \|f\|_{2,\mathbb{P}} \leq \sigma$*; and*

(iii) $\sup_{\mathbb{Q}} N(\epsilon\|F\|_{1,\mathbb{Q}}, \mathcal{F}, L_{1,\mathbb{Q}}) \leq C\epsilon^{-d}$.

Let $\theta \coloneqq \tau^{-1/2}\sigma$*,* $m \in \mathbb{N}$ *and* $x > 0$*. Then there exist* $C_1, C_2 \in (0,\infty)$*, not depending on* τ*,* σ *or* x*, such that*

$$\mathbb{P}\left\{\sigma^{-2}\sup_{f\in\mathcal{F}}|\mu_n f| > x\right\} \leq C_1 \exp[-C_2 n\theta^2(1+x^2) + d\log(\theta^{-2}x^{-1})] \tag{E.2}$$

for all $n \geq \frac{1}{8}x^{-2}\theta^{-2}$.

Proof of Proposition E.1. We first note that, by E2,

$$\max_{i\leq n}|F(z_i)| = o_p(n^{-1/p})$$

and so, letting $f^n_{(\gamma,\delta)}(z) \coloneqq f_{(\gamma,\delta)}(z)\mathbf{1}\{F(z) \leq n^{1/p}\}$, we have

$$\mathbb{P}\left\{\sup_{(\gamma,\delta)\in\Gamma\times\Delta_n} \delta^m|\mu_n[f_{(\gamma,\delta)} - f^n_{(\gamma,\delta)}]| = 0\right\} \leq \mathbb{P}\left\{\max_{i\leq n}|F(z_i)| > n^{1/p}\right\} = o(1).$$

It thus suffices to show that (E.1) holds when $f_{(\gamma,\delta)}$ is replaced by $f^n_{(\gamma,\delta)}$. Since E1 and E3 continue to hold after this replacement, it suffices to prove (E.1) when E2 is replaced by the condition that $\|F\|_\infty \leq n^{1/p}$, which shall be maintained throughout the sequel. (The dependence of f and F upon n will be suppressed for notational convenience.)

Letting $\delta_k \coloneqq \mathrm{e}^k$, define $\Delta_{nk} \coloneqq [\delta_k, \delta_{k+1} \wedge \overline{\delta}_n]$ for $k \in \{0,\ldots,K_n\}$, where $K_n \coloneqq \log\overline{\delta}_n$; observe that $\Delta_n = \bigcup_{k=0}^{K_n}\Delta_{nk}$. Set

$$\mathcal{F}_{nk} \coloneqq \{z \mapsto f_{(\gamma,\delta)}(z) \mid (\gamma,\delta) \in \Gamma\times\Delta_{nk}\}$$

and note that $\|F\|_\infty \leq n^{1/p}$ and $\sup_{f\in\mathcal{F}_{nk}}\|f\|_{2,\mathbb{P}} \leq \delta_k^{-1/2}$. Under E3, we may apply apply Lemma E.1 to each $\mathcal{F}_{nk}$, with $(\tau,\sigma) = (n^{1/p}, \delta_k^{-1/2})$ and $x = \delta_k^{1-m}\epsilon$, for some $\epsilon > 0$. There thus

exist $C_1, C_2 \in (0,\infty)$ depending on ϵ such that

$$\mathbb{P}\left\{\sup_{(\gamma,\delta)\in\Gamma\times\Delta_n}\delta^m|\mu_n f_{(\gamma,\delta)}|>\epsilon\right\}\leq\sum_{k=0}^{K_n}\mathbb{P}\left\{\delta_k^m\sup_{(\gamma,\delta)\in\Gamma\times\Delta_{nk}}|\mu_n f_{(\gamma,\delta)}|>\mathrm{e}^{-1}\epsilon\right\}$$
$$\leq C_1\sum_{k=0}^{K_n}\exp[-C_2 n\theta_{nk}^2\delta_k^{2(1-m)}+d\log(\theta_{nk}^{-2}\delta_k^{m-1})] \tag{E.3}$$

where $\theta_{nk}\coloneqq n^{-1/2p}\delta_k^{-1/2}$, provided

$$n\geq\tfrac{1}{8}\delta_k^{2(m-1)}\theta_{nk}^{-2}\epsilon^{-2},\ \forall k\in\{0,\ldots,K_n\}\iff n^{1-1/p}/\overline{\delta}_n^{2m-1}\geq\tfrac{1}{8}\epsilon^{-2}, \tag{E.4}$$

which holds for all n sufficiently large. In obtaining (E.4) we have used $\delta_k\leq\overline{\delta}_n$ and $\theta_{nk}\geq n^{-1/2p}\overline{\delta}_n^{-1/2}$, and these further imply that (E.3) may be bounded by

$$C_1(\log\overline{\delta}_n)\exp[-C_2 n^{1-1/p}\overline{\delta}_n^{-2m-1}(1+\epsilon^2)+d\log(\overline{\delta}_n^m n^{1/p})]\to 0$$

as $n\to\infty$. Thus (E.1) holds. □

Proof of Lemma E.1. Suppose (iii) holds. Define $\mathcal{G}\coloneqq\{\tau^{-1}f\mid f\in\mathcal{F}\}$, and $G\coloneqq\tau^{-1}F$. Then

$$\sup_{g\in\mathcal{G}}\|g\|_{2,\mathbb{P}}\leq\tau^{-1}\sup_{f\in\mathcal{F}}\|f\|_{2,\mathbb{P}}\leq\tau^{-1/2}\sigma=:\theta;$$

$\|g\|_\infty\leq 1$ for all $g\in\mathcal{G}$; and since $\|G_n\|_{1,\mathbb{Q}}\leq 1$, $N(\epsilon,\mathcal{G},L_{1,\mathbb{Q}})\leq C\epsilon^{-d}$. Hence, by arguments given in the proof of Theorem II.37 in Pollard (1984), there exist $C_1, C_2>0$, depending on x, such that

$$\mathbb{P}\left\{\sigma^{-2}\sup_{f\in\mathcal{F}}|\mu_n f|>x\right\}=\mathbb{P}\left\{\sup_{g\in\mathcal{G}}|\mu_n g|>\theta^2 x\right\}\leq C_1\exp[-C_2 n\theta^2(1+x^2)+d\log(\theta^{-2}x^{-1})]$$

for all $n\geq\frac{1}{8}x^{-2}\theta^{-2}$. □

F Index of key notation

Greek and Roman symbols

Listed in (Roman) alphabetical order. Greek symbols are listed according to their English names: thus Ω, as 'omega', appears before θ, as 'theta'.

β, β_0, B	structural model parameters, true value, parameter space	Sec. 2
$\hat{\beta}^e_{nk}$	GII estimator; near-minimizer of Q^e_{nk}	Sec. 5.3
$\overline{\beta}^e_{nk}(\beta^{(0)}, r)$	terminal value for routine r started at $\beta^{(0)}$	(5.15)
c_n	tuning sequence in the definition of R^e_{nk}	Sec. 5.5
d_β, d_θ, ...	dimensionality of β, θ, etc.	Sec. 3.1
$\mathbb{E}_n f$	sample average, $\frac{1}{n}\sum_{i=1}^n f(z_i)$	App. D.1
η_{it}	stochastic components of the structural model	Sec. 2
$\mathcal{F}$	σ-field supporting all observed and simulated variates	Sec. 5.1
G	Jacobian of the population binding function	Sec. 5.3
$G_n(\beta)$	Jacobian of the smoothed sample binding function	Rem. 5.15
$\gamma(\beta)$	(re-)parametrizes the structural model	(5.2)
γ_{rk}	jackknifing weights	(4.3)
H	auxiliary model (population) log-likelihood Hessian	Sec. 5.3
J	total number of alternatives	Sec. 2
k	order of jackknifing (unless otherwise defined)	R6
k_0	maximum order (less 1) of differentiability of $\beta \mapsto \theta(\beta, \lambda)$	R3
K, K_λ	smoothing kernel, $K_\lambda(x) \coloneqq K(\lambda^{-1}x)$	(5.3)
$\mathbb{K}$, $\mathbb{K}_s$, $\mathbb{K}_{st}$	product of kernel-type functions	App. E
$\ell(y_i, x_i; \theta)$	ith contribution to auxiliary model log-likelihood	(3.1)
$\ell(\beta, \lambda; \theta)$	abbreviates $\ell(y_i(\beta, \lambda), x_i; \theta)$	Sec. 5.1
$\ell^\infty(D)$	space of bounded functions on the set D	H3
$\mathcal{L}_n(y, x; \theta)$	auxiliary model average log-likelihood	(3.1)
λ, Λ	smoothing parameter, set of allowable values	Sec. 3.3
m	indexes the simulated dataset; $m = 0$ denotes the data	Sec. 3.1
M	total number of simulations	Sec. 3.1
$\mu_n f$	centered sample average, $\frac{1}{n}\sum_{i=1}^n [f(z_i) - \mathbb{E}f(z_i)]$	App. D.1
n	total number of individuals	Sec. 2
$N(\theta, \epsilon)$	open ball of radius ϵ centered at θ	App. B.1
$\nu_r(z; \beta)$	linear index in structural model	(5.2a)
$\omega_r(z; \beta)$	linear index in structural model	(5.2a)
$\Omega(U, V)$	variance matrix function	(5.10)
p_0	order of moments possessed by model variates	L6
ϕ^m_n, ϕ^m	standardized auxiliary sample score and its weak limit	(5.5)

ψ_n^m, ψ^m	centered auxiliary estimator process and its weak limit	H3
Q_{nk}^e	sample criterion for estimator e (jackknifed)	Sec. 4.2
Q_k^e	large-sample (unsmoothed) limit of Q_{nk}^e; note $Q_k^e = Q^e$	Sec. 4.2
R	auxiliary model score covariance, $\mathbb{E}\dot{\ell}_i^m(\theta_0)\dot{\ell}_i^{m'}(\theta_0)^\mathsf{T}$ for $m' \neq m$	(5.6)
R_{nk}^e, R^e	set of near-roots of Q_{nk}^e, exact roots of Q^e	Sec. 5.5
$\varrho_{\min}(A)$	smallest eigenvalue of symmetric matrix A	Sec. 5.1
S_{nk}^e, S^e	subset of R_{nk}^e, R^e satisfying second-order conditions	(5.13)
$\sigma_{\min}(B)$	smallest singular value of matrix B	Sec. 5.6
Σ	auxiliary model score variance, $\mathbb{E}\dot{\ell}_i^m(\theta_0)\dot{\ell}_i^m(\theta_0)^\mathsf{T}$	(5.6)
T	total number of time periods	Sec. 2
θ, Θ	auxiliary model parameters, parameter space	Sec. 3.1
θ_0	pseudo-true parameters implied by β_0	Sec. 3.1
$\hat{\theta}_n$	data-based estimate of θ	Sec. 3.1
$\hat{\theta}_n^m(\beta, \lambda)$	simulation-based estimate of θ	(3.3)
$\theta^k(\beta, \lambda)$	population binding function (smoothed, jackknifed)	(4.3)
$\overline{\theta}_n^k(\beta, \lambda)$	sample binding function (smoothed, jackknifed)	(4.4)
u_{itj}	utility of individual i from alternative j in period t	Sec. 2
$u_{itj}^m(\beta)$	simulated utilities at β	Sec. 3.3
U_e	"Hessian" component of limiting variance	(5.11)
V_e	"score" component of limiting variance	(5.11)
w.p.a.1	with probability approaching one	Thm. 5.3
$W_r(z)$	envelope for $\omega_r(z; \beta)$	Sec. 5.1
W_n, W	Wald weighting matrix and its probability limit	Sec. 3.1
x_{it}	exogenous covariates for individual i in period t	Sec. 2
y_{itj}	set = 1 if individual i chooses j in period t	Sec. 2
$y_{itj}^m(\beta, \lambda)$	smoothed simulated choice indicators at β	Sec. 3.3
z_i^m	collects x_i and η_i^m	Sec. 5.1

Symbols not connected to Greek or Roman letters

Ordered alphabetically by their description.

$\rightsquigarrow$	weak convergence (van der Vaart and Wellner, 1996)	H3
$\xrightarrow{p}$	convergence in probability	Sec. 5
$\|x\|$, $\|x\|_A$	Euclidean norm, A-weighted norm of x	Sec. 3.1
$\dot{f}$, $\ddot{f}$	gradient, hessian of f	Sec. 5.3
$\partial_\beta f$, $\partial_\beta^2 f$	gradient, hessian of f w.r.t. β	Rem. 5.1
$\lesssim$	left side bounded by the right side times a constant	App. D.1
$\|f\|_{p,\mathbb{Q}}$	$L^p(\mathbb{Q})$ norm of f, i.e. $(\mathbb{E}_\mathbb{Q}\|f(z_i)\|^p)^{1/p}$	App. D.1