Math. Program., Ser. B 108, 207–234 (2006)

Digital Object Identifier (DOI) 10.1007/s10107-006-0708-6

Fabian Bastin* · Cinzia Cirillo · Philippe L. Toint

Convergence theory for nonconvex stochastic programming with an application to mixed logit

Received: November 21, 2003 / Accepted: October 13, 2005
Published online: April 25, 2006 –

Abstract. Monte Carlo methods have extensively been used and studied in the area of stochastic programming. Their convergence properties typically consider global minimizers or first-order critical points of the sample average approximation (SAA) problems and minimizers of the true problem, and show that the former converge to the latter for increasing sample size. However, the assumption of global minimization essentially restricts the scope of these results to convex problems. We review and extend these results in two directions: we allow for local SAA minimizers of possibly nonconvex problems and prove, under suitable conditions, almost sure convergence of local second-order solutions of the SAA problem to second-order critical points of the true problem. We also apply this new theory to the estimation of mixed logit models for discrete choice analysis. New useful convergence properties are derived in this context, both for the constrained and unconstrained cases, and associated estimates of the simulation bias and variance are proposed.

1. Introduction

Stochastic programming, that is mathematical programming where uncertainty is introduced in the problem by the use of random variables, is today recognized as an important area of operations research (see the books by Birge and Louveaux [9] and by Kall and Wallace [28], for instance). Amongst the methods of stochastic programming, Monte Carlo techniques are well-known tools for the case where the random variables are either discrete with a large number of possible realizations, or continuous. However, to our knowledge, the convergence theory for these methods has so far been limited to the case where the minimization of the approximating subproblems is assumed to produce a global minimum in all the feasible set (see for instance Shapiro [37]), a first-order critical point (Gürkan, Özge and Robinson [22, 23], Shapiro [38]) or a solution in a complete local minimizing set with respect to some nonempty open bounded set (Robinson [33]). As a result, they have mostly been applied to linear or convex problems, because then, such assumptions are not restrictive.

It is our purpose to extend the theoretical understanding of this class of methods to the case where the global minimization assumption no longer holds: the minimization of the subproblem is allowed to converge to a local minimizer, irrespectively of the set of minimizers of the true problem. This investigation is worthwhile, in particular because

F. Bastin: Department of Mathematics, University of Namur, Belgium and Cerfacs, France. e-mail: fabian.bastin@cerfacs.fr

C. Cirillo, Ph.L. Toint: Transportation Research Group, Department of Mathematics, University of Namur, Belgium.

* Research Fellow of the Belgian National Fund for Scientific Research

it opens the possibility of considering stochastic problems with nonconvex objective functions. We introduce our approach by reviewing consistency results when only first-order critical points are considered. We next study second-order criticality issues and show that, when the sample size tends to infinity, approximating local solutions may have limit points that are not (local) solutions of the true problem. We then set conditions under which second-order properties are preserved for limit points. This is interesting because there may be more than one solution in the nonconvex case, and they often do not share the same constraint qualification properties.

Nonconvex stochastic problems do occur in practice, and we will apply our convergence results to the specific and important case of (possibly) constrained parameter estimation in mixed logit models. Mixed logit modelling is one of the most powerful tools currently available to estimate individual demand from discrete choice responses. In spite of their inherent complexity, they are becoming very popular among researchers and practitioners in economics and transportation (see, for instance, Montmarquette, Cannings, and Mahseredjian [29], Bhat and Castelar [7], Cirillo and Axhausen [11], Hensher and Greene [25]). Their advantages include the possibility to estimate taste variations, to account for state dependence across observations and to avoid the problem of restricted substitution patterns in the standard logit model. However, the complexity of the likelihood function, the loss of an easy behavioural interpretation of the results and a heavier computational burden mitigate these advantages. In particular, mixed logit model estimation involves the evaluation of multidimensional integrals describing the choice probabilities, which are typically calculated, in real applications, by the following sampling (simulation) technique. For each individual in the considered population, pseudo-random sequences are drawn from a given density and, for each draw, observed parts of the alternatives utilities are calculated conditionally to this realization and inserted in the logit formula. The integral giving the probability choice for this individual is then approximated by the mean of the conditional probabilities. Gouriéroux and Monfort [19], as well as Hajivassiliou and McFadden [24], have shown that the computed estimators are, under reasonable assumptions, asymptotically consistent and efficient (in the statistical sense). But, even in this form, evaluation costs can be prohibitive. The current research approach has thus shifted, in order to reduce computational time and simulation error, to quasi-Monte Carlo approaches instead of pure Monte Carlo methods. Bhat [5] and Train [40] have advocated Halton sequences for mixed logit models estimation and have reported that they perform much better than random draws. However, Bhat [6] has pointed out that Halton sequences rapidly deteriorate in the coverage of the integration domain for high integration dimensions and has proposed the use of scrambled Halton sequences. He also randomized these sequences in order to allow the computation of the simulation variance of the model parameters. Hess, Polak, and Daly [26] have shown that scrambled Halton sequences can be very sensitive to the number of draws, and can behave poorly when this number increases. Various other alternatives have then been proposed: Hess, Train, and Polak [27] have considered the use of modified latin hypercube sequences, but their results have been mitigated in Sivakumar, Bhat, and Ökten [39] and Bastin, Cirillo, and Hess [2], while Sándor and Train [35] have examined the use of (t, m, s)-nets, and Garrido [16] has explored the application of Sobol sequences.

The second purpose of this paper is nevertheless to provide additional insight in the process of estimating mixed logit models that can be derived from considering the question in the framework of pure Monte Carlo methods. The main reason for returning to the pure Monte Carlo framework is that it completely avoids the problems of sample correlations and loss of uniform coverage in the estimation of high-dimensional integrals. For such problems, practitioners report that Monte Carlo methods are again competitive compared to quasi-Monte Carlo approaches (Deak [13], Hess, Train, and Polak [27]). We apply our convergence results for stochastic programs to develop almost sure convergence of the approximating solutions to the true maximum likelihood estimators, when the population size is fixed, covering both constrained and unconstrained problems as well as nonlinear utilities. These results are theoretically interesting since they complete the classical ones in mixed logit theory (see Train [41]), exploring convergence in probability and in distribution when both sample and population sizes grow. The asymptotic behaviour of the approximating solutions, when the population size increases, is also briefly discussed in this paper. A second reason for our interest in Monte Carlo techniques is that statistical inference can be easily used to provide computable estimates of the simulation bias and variance.

The paper is organized as follows. We introduce the general stochastic problem and its application to mixed logit models in Section 2. Sections 3 and 4 discuss our convergence theory for the general problem, while Section 5 applies them to the mixed logit case and explores bias and variance estimates. Some conclusions and perspectives are outlined in Section 6.

2. Stochastic programming and mixed logit parameter estimation

2.1. The stochastic problem

A classical problem in stochastic programming is the minimization of the expectation of some function depending on a random variable (see Birge and Louveaux [9] or Kall and Wallace [28] for a more complete exposition):

$$\min_{z \in S} g(z) = E_P \left[G(z, \boldsymbol{\xi}) \right], \tag{2.1}$$

where $z \in \mathbb{R}^m$ is a vector of decision variables, S is a compact subset of $\mathbb{R}^m$ representing feasible solutions of the above problem, $\boldsymbol{\xi}$ is a real random vector defined on the probability space $(\Xi, \mathcal{F}, P)$ and taking values in $\left(\mathbb{R}^k, \mathcal{B}^k\right)$, $G: \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}$ is a real valued function, and $E_P[\cdot]$ is the expectation w.r.t. the measure P. We assume that for every $z \in S$ the expected value function $g(z)$ is well defined, i.e. that the function $G(z, \cdot)$ is $\mathcal{F}$-measurable and P-integrable. For simplicity, we restrict ourselves in a first step to the case where the set S is deterministic.

If the distribution function of $\boldsymbol{\xi}$ is continuous or discrete with a large number of possible realizations, $g(z)$ is usually very hard to evaluate. Solving the problem (2.1) thus becomes difficult and we have to turn to approximations such as Monte Carlo methods (see Shapiro [37, 38] for a review). In these methods, the original problem (2.1) is

replaced by successive approximations obtained by generating samples $\xi_1, \ldots, \xi_N$. The approximation for a sample of size N is

$$\min_{z \in S} \hat{g}_N(z) = \frac{1}{N} \sum_{i=1}^{N} G(z, \xi_i). \tag{2.2}$$

We refer to (2.1) and (2.2) as the true (or expected value) and the sample average approximation (SAA) problems, respectively.

2.2. Discrete choice models and mixed logit

The field of discrete choice modelling attempts to provide an operational description of how individuals perform a selection amongst a finite (discrete) set of alternatives. Choice between competing products in a marketing campaign (see for instance Anderson, De Palma and Thisse [1]) or between transportation modes for travel (see e.g. Ben-Akiva and Lerman [4]) are good examples of the many possible applications.

In this theory, the probability of an individual choosing a given alternative is modelled as a function of his/her socio-economic characteristics and of the relative attractiveness of the alternative. Let $\mathcal{A}$ the set of alternatives and I the population size. The set of alternatives available to individual i ($i = 1, \ldots, I$) is represented by $\mathcal{A}(i) \subseteq \mathcal{A}$. For each individual i, each available alternative $A_j \in \mathcal{A}(i)$ ($j = 1, \ldots, |\mathcal{A}(i)|$) has an associated utility U_{ij}, which is typically split into two components,

$$U_{ij} = V_{ij} + \boldsymbol{\epsilon}_{ij}.$$

In this description, $V_{ij} = V_{ij}(\beta_j, x_{ij})$ is a function of some model parameters β_j and of x_{ij}, the observed attributes of alternative A_j, while $\boldsymbol{\epsilon}_{ij}$ is a random term reflecting the unobserved part of the utility. Without loss of generality, it can be assumed that the residuals $\boldsymbol{\epsilon}_{ij}$ are random variables with zero mean and a certain probability distribution to be specified. The parameter vectors β_j ($j = 1, \ldots, |\mathcal{A}(i)|$) are assumed to be constant for all individuals but may vary across alternatives. The theory then assumes that individual i selects the alternative that maximizes his/her utility. In other terms, he/she chooses A_j if

$$U_{ij} \geq U_{il}, \quad \forall A_l \in \mathcal{A}(i).$$

Thus the probability of choosing alternative A_j is given by

$$P_{ij} = P\left[\epsilon_{il} \leq \epsilon_{ij} + \left(V_{ij} - V_{il}\right), \forall A_l \in \mathcal{A}(i)\right].$$

A model parameter is called generic if it is involved in all alternatives, and has the same value for all of them. Otherwise it is said to be (alternative) specific. Since we can decompose a specific parameter in several parameters taking the same value for a subset of alternatives, and associated with null observations for others, we may assume, without loss of generality, that all parameters are generic. In order to simplify the notation, we will hence omit the subscript j for parameter vectors.

A popular distribution in discrete choice models is the Gumbel distribution, also called the extreme value type I distribution. Its probability distribution function is

$$f(x) = \mu e^{-\mu(x-\eta)} e^{-e^{-\mu(x-\eta)}},$$

where η is a location parameter and $\mu > 0$ is a scale factor. Its mean is $\eta + \gamma_E/\mu$, where $\gamma_E \approx 0.57721$ is the Euler constant. The popularity of this distribution partly lies in the fact that it allows to express the choice probabilities in a very simple form. Assume indeed that the residuals $\boldsymbol{\epsilon}_{ij}$ are independently Gumbel distributed (with mean 0 and scale factor 1.0). The probability that the individual i chooses the alternative j is then expressed by the logit formula

$$L_{ij}(\beta) = \frac{e^{V_{ij}(\beta, x_{ij})}}{\sum_{m=1}^{|\mathcal{A}(i)|} e^{V_{im}(\beta, x_{im})}}. \tag{2.3}$$

This is the multinomial logit model, which has some serious drawbacks. In particular the assumption that the error terms are identically and independently distributed (IID) across alternatives induces the independence of irrelevant alternatives (IIA) property, which states that, if some alternatives are removed from a choice set, the relative choice probabilities in the reduced choice set remain unchanged. A formal description of this property and of the associated difficulties can be found for instance in Ben-Akiva and Lerman [4], who show that the IIA assumption validity depends on the choice set structure, and that it may be unrealistic if the alternatives are not distinct for the individual.

Several extensions of the multinomial logit model have been proposed and allow to partially avoid the IID assumption. These extensions include the mixed logit models (or error components models) (see Bhat and Koppelman [8] for a review of these developments). Mixed-logit models use non-identical, non-independent random components, so they fully relax the IID assumption and overcome the rigid inter-alternative substitution pattern of the multinomial logit models. Using the random coefficients formalism, we relax the assumption that the parameters β are the same for all individuals, but assume instead that each parameter vector $\beta(i)$ $(i = 1, \ldots, I)$ is a realization of a random vector $\boldsymbol{\beta}$. Furthermore, $\boldsymbol{\beta}$ is itself assumed to be derived from a random vector $\boldsymbol{\gamma}$, specifying the random nature of the model, and a parameter vector θ, quantifying the population characteristics, which we express as

$$\boldsymbol{\beta} = h(\boldsymbol{\gamma}, \theta). \tag{2.4}$$

Assume for instance that $\boldsymbol{\beta}$ is a K-dimensional vector of independent normal variables whose k-th component is $N(\mu_k, \sigma_k^2)$, where $N(\mu, \sigma^2)$ designs a normal distribution of mean μ and variance σ^2. We may then choose $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_K)$, with $\boldsymbol{\gamma}_k \sim N(0, 1)$ and let the vector θ specify the means and standard deviations of the $\boldsymbol{\beta}_k$, that is $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots, \mu_K, \sigma_K)$. Therefore, (2.4) can be written in this case as $\boldsymbol{\beta} = \left(\mu_1 + \sigma_1 \boldsymbol{\gamma}_1, \mu_2 + \sigma_2 \boldsymbol{\gamma}_2, \ldots, \mu_K + \sigma_K \boldsymbol{\gamma}_K\right)$.

If we knew the realization $\gamma(i)$, and thus the value $\beta(i) = h(\gamma(i), \theta)$, for some individual i, the conditional probability that he/she chooses alternative j would then be

given by the logit formula (2.3). However, since $\boldsymbol{\beta}$ is random, we need to calculate the associated unconditional probability, which is obtained by integrating (2.3) over $\boldsymbol{\gamma}$:

$$P_{ij}(\theta) = E_P\left[L_{ij}(\boldsymbol{\gamma}, \theta)\right] = \int L_{ij}(\gamma, \theta) P(d\gamma) = \int L_{ij}(\gamma, \theta) f(\gamma) d\gamma, \quad (2.5)$$

where P is the probability measure associated with $\boldsymbol{\gamma}$, and $f(\cdot)$ its distribution function.

The unknown values of θ are then estimated by maximizing the corresponding log-likelihood function, i.e. by solving the program

$$\max_{\theta} LL(\theta) = \max_{\theta} \frac{1}{I} \sum_{i=1}^{I} \ln P_{ij_i}(\theta), \quad (2.6)$$

where j_i is the alternative choice made by the individual i. This involves the computation of $P_{ij_i}(\theta)$ of (2.5) for each individual i ($i = 1, \ldots, I$), which is impractical since it requires the evaluation of one multidimensional integral per individual. We therefore use a Monte Carlo estimate of $P_{ij_i}(\theta)$ obtained by sampling over γ, and given by

$$SP^R_{ij_i}(\theta) = \frac{1}{R} \sum_{r=1}^{R} L_{ij_i}(\gamma_{i,r}, \theta),$$

where R is the number of random draws $\gamma_{i,r}$, taken from the distribution function of $\boldsymbol{\gamma}$. As a result, θ is now computed as the solution (if it exists) of the simulated log-likelihood problem

$$\max_{\theta} SLL^R(\theta) = \max_{\theta} \frac{1}{I} \sum_{i=1}^{I} \ln SP^R_{ij_i}(\theta).$$

However, since I can be large (typically in the thousands), the evaluation of $SLL^R(\theta)$ may remain very expensive, as pointed out by Hensher and Greene [25].

We finally notice that the mixed logit problem (2.5)–(2.6) can be viewed as a generalized stochastic programming problem similar to (2.1), since we can write (2.5)–(2.6) as

$$\min_{\theta} g(\theta) = -\min_{\theta} LL(\theta) = -\frac{1}{I} \min_{\theta} \sum_{i=1}^{I} \ln E_P\left[L_{ij_i}(\gamma, \theta)\right]. \quad (2.7)$$

The associated sample average approximation problem is then written as

$$\min_{\theta} \hat{g}_N(\theta) = -\min_{\theta} SLL^R(\theta) = -\frac{1}{I} \min_{\theta} \sum_{i=1}^{I} \ln SP^R_{ij_i}(\theta), \quad (2.8)$$

where $N = R\,I$. We will denote by θ^* a solution of (2.7) and by θ^*_R a solution of (2.8). The generalization is minor since it only consists in optimizing a sum of logarithms of expectations, instead of a single expectation.

3. First-order convergence for stochastic programs

We now investigate the convergence of the solutions and optimal values of the sequence of SAA problems (2.2) to a solution and optimal value of (2.1) for $N \to \infty$. We introduce the basic concepts used in this paper by reviewing first-order convergence.

Let z_N^* be a first-order critical point for problem (2.2), that is the Karush-Kuhn-Tucker (KKT) conditions are satisfied at z_N^*. (The KKT conditions will be formally expressed in Section 3.2.) In order to stress the dependence of z_N^* on the successive draws $\xi_1, \ldots, \xi_N$, we will often use the notation $z_N^*(\xi_1, \ldots, \xi_N)$, or $z_N^*\left(\overline{\xi}\right)$, since $(\xi_1, \ldots, \xi_N)$ can be seen as the finite truncation of an infinite sequence $\overline{\xi} := \{\xi_k\}_{k=1}^{\infty}$. Since S is a compact set, the sequence of SAA solutions has a non-empty set of (finite) limit points. Our first aim is to study under which reasonable assumptions, such a limit point is a first-order critical point for the true problem (2.1).

Since the set of limit points depends on the sequence of realizations $\overline{\xi}$, which is not known a priori, we have to introduce a suitable probability space on which we can define random variables whose realizations are such (infinite) sequences. Consider the stochastic process

$$\overline{\boldsymbol{\xi}} = \{\boldsymbol{\xi}_k\}_{k=1}^{\infty},$$

later called the sampling process, where the random vectors $\boldsymbol{\xi}_k$, $k = 1, \ldots, \infty$, are assumed to be independent and identically distributed (IID). From the IID property and the Kolmogorov consistency theorem (see for instance Parthasarathy [31], Chapter V, Theorem 5.1), we can construct the infinite-dimensional probability space

$$(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi), \tag{3.1}$$

where the measure P_Π has the property that for any non-zero natural j,

$$P_\Pi[B] = \prod_{i=1}^{j} P[B_i],$$

for any set $B = \prod_{i=1}^{j} B_i \times \prod_{i=j+1}^{\infty} \Xi$, with $B_i \in \mathcal{F}$, $i = 1, \ldots, j$. In other terms, the marginal measures defined on $\prod_{i=1}^{j}(\Xi, \mathcal{F})$, with finite j ($j = 1, \ldots$), correspond to the product measures $\prod_{i=1}^{j} P$, as expected. We can therefore view $\overline{\boldsymbol{\xi}}$ as a random variable on (3.1), whose realizations are processes formed by the successive draws ξ_k, $k = 1, \ldots, \infty$, i.e. $\overline{\xi} = \{\xi_k\}_{k=1}^{\infty}$.

It is useful at this stage to introduce some notations which will be used throughout the paper. We use the symbols

- $\xrightarrow{a.s.}$ for almost sure convergence;
- $\xrightarrow{p}$ for convergence in probability;
- $\Rightarrow$ for convergence in distribution.

We refer the reader to Davidson [12] for the definitions of the various types of convergence, which are mentioned here in order of decreasing strength. In what follows, and unless explicitly stated, we will assume that the terms *almost every* and *almost surely* refer

to the infinite-dimensional space $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, which allows explicit consideration of the sets of realizations whose elements are of the form $\{\xi_N\}_{N=1}^{\infty}$. (In other words, results expressed in these terms hold for almost every sampling process.) Reference to another probability space will be denoted by prefixing the terms *almost every* and *almost surely* by the measure defined on this probability space, and the expression *almost every sampling process* will be prefixed by the probability measure associated with the probability space of each element of the process. As above, we continue to use bold symbols to denote random variables, while a realization of such a variable is represented in standard font. We also denote by $[z]_i$ the i-th component of the vector z.

We now state our assumptions.

A.0 The random draws $\{\xi_k\}_{k=1}^{\infty}$ are independently and identically distributed.
A.1 For P-almost every ξ, the function $G(\cdot, \xi)$ is continuously differentiable on S.
A.2 The family $G(z, \xi)$, $z \in S$, is dominated by a P-integrable function $K(\xi)$, i.e. $E_P[K]$ is finite and $|G(z, \xi)| \leq K(\xi)$ for all $z \in S$ and P-almost every ξ.

A.1 obviously implies that $G(\cdot, \xi)$ is continuous P-almost surely. This and **A.2** are typical assumptions of stochastic programming theory (see for instance Rubinstein and Shapiro [34]). The stronger form of **A.1** is justified by our interest in first-order optimality conditions, which require the objective function's gradient.

It is important to note (see [34] again) that **A.0**–**A.2** together imply that there exists a uniform law of large numbers (ULLN) on S, for the approximation $\hat{g}_N(z)$ of $g(z)$:

$$\sup_{z \in S} \left| \hat{g}_N(z) - g(z) \right| \xrightarrow{a.s.} 0 \quad \text{as } N \to \infty.$$

The ULLN property corresponds to the stochastic version of the uniform convergence of a sequence of functions, and **A.0**–**A.2** also imply that $g(z)$ is then continuous on S.

Our motivation to study first-order conditions leads also us to add a further assumption on the gradient.

A.3 Each gradient component $\frac{\partial}{\partial [z]_l} G(z, \xi)$ $(l = 1, \ldots, m)$, $z \in S$, is dominated by a P-integrable function.

This new assumption allows us to apply the results of Rubinstein and Shapiro [34], page 71, and deduce that the expected value function $g(z)$ is continuously differentiable over S, and that the expectation and gradient operator can be interchanged in the expression of the gradient, giving $\nabla_z g(z) = E_P \left[\nabla_z G(z, \boldsymbol{\xi}) \right]$. This also implies that $\nabla \hat{g}_N(z^*)$ is an unbiased estimator of $\nabla g(z^*)$.

First-order convergence can be derived from stochastic variational inequalities, as presented in Shapiro [38]. Consider a mapping $\Phi\colon \mathbb{R}^m \times \prod_{i=1}^{\infty} \mathbb{R}^k \to \mathbb{R}^m$ and a multifunction $\Gamma\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^m$. Suppose that the expectation $\phi(z) := E_{P_\Pi}[\Phi(z, \boldsymbol{\xi})]$ is well defined. We refer now to

$$\phi(z) \in \Gamma(z) \tag{3.2}$$

as the true, or expected value, generalized equation and say that a point $z^* \in \mathbb{R}^m$ is a solution of (3.2) if $\phi(z^*) \in \Gamma(z^*)$. If $\{\xi_1, \ldots, \xi_N\}$ is a random sample, we refer to

$$\hat{\phi}_N(z) \in \Gamma(z) \tag{3.3}$$

as the SAA generalized equation, where $\hat{\phi}_N(z) = N^{-1}\sum_{i=1}^{N}\Phi(z,\xi_i)$. We denote by S^* and S_N^* the sets of (all) solutions of the true (3.2) and SAA (3.3) generalized equations, respectively.

Let us denote by $d(x, A) := \inf_{x'\in A}\|x - x'\|$, the distance from $x \in \mathbb{R}^m$ to A, and $D(A, B) := \sup_{x\in A} D(x, B)$, the deviation of the set A from the set B. We then have the following result (Shapiro [38]):

Theorem 3.1. *Let $\mathcal{S}$ be a compact subset of $\mathbb{R}^m$ such that $S^* \subseteq \mathcal{S}$. Assume that*

(a) the multifunction $\Gamma(z)$ is closed, that is if $z_k \to z$, $y_k \in \Gamma(z_k)$ and $y_k \to y$, then $y \in \Gamma(z)$,
(b) the mapping $\phi(z)$ is continuous on $\mathcal{S}$,
(c) almost surely, $\emptyset \neq S_N^* \subseteq \mathcal{S}$ for sufficiently large N, and
(d) $\hat{\phi}_N(z)$ converges to $\phi(z)$ almost surely uniformly on $\mathcal{S}$ as $N \to \infty$.

Then $D(S_N^*, S^*) \to 0$ almost surely as $N \to \infty$.

3.1. Deterministic and convex constraints

When S is convex, we can rewrite the first-order criticality conditions for some point z^* as the requirement that $-\nabla_z g(z^*)$ belongs to the normal cone to S at z^*, denoted by $\mathcal{N}_S(z^*)$. If S is moreover deterministic, the feasible sets are the same for the true and SAA problems. Theorem 3.1 then allows an easy proof of almost sure first-order convergence. Consider the choice $\Gamma(\cdot) = \mathcal{N}_S(\cdot)$; $\phi(z^*)$ belongs to $\Gamma(z^*)$ if and only if

$$\langle \phi(z^*), u - z^*\rangle \leq 0, \quad \forall u \in S.$$

Following Shapiro [38], we refer to such variational inequalities as stochastic variational inequalities and note that the assumption (a) of Theorem 3.1 always holds in this case. Take $\Phi(z,\xi) = -\nabla_z G(z,\xi)$ and let S^* and S_N^* represent the set of first-order critical points of the true (3.2) and SAA (3.3) generalized equations, respectively. Then under **A.0**–**A.3**, we have that $\phi(z) = -\nabla_z g(z)$, and that $\phi(z)$ is a continuous random vector on S, yielding assumption (b). Assumption (d) results from the ULLN, while **A.1** and the compacity of S ensure assumption (c) by setting $\mathcal{S} = S$. Thus Theorem 3.1 guarantees first-order criticality in the limit as $N \to \infty$, almost surely.

3.2. Stochastic constraints

Under stronger assumptions, it is also possible to prove almost-sure first-order convergence when S is nonconvex or non-deterministic. We now suppose that the feasible set can be described by equality and inequality constraints. The original problem is then

$$\begin{aligned} \min_{z\in V}\ & g(z) = E_P[G(z,\boldsymbol{\xi})], \\ \text{subject to } & c_j(z) \geq 0, \quad j = 1,\ldots,k, \\ & c_j(z) = 0, \quad j = k+1,\ldots,M, \end{aligned} \tag{3.4}$$

where V is a compact subset of $\mathbb{R}^m$. The corresponding SAA problem is defined as

$$\begin{aligned} &\min_{z\in V} \ \hat{g}_N(z), \\ \text{subject to} \quad &\hat{c}_{jN}(z) \geq 0, \quad j = 1, \ldots, k, \\ &\hat{c}_{jN}(z) = 0, \quad j = k+1, \ldots, M. \end{aligned} \tag{3.5}$$

Here, for every $j = 1, \ldots, M$, $\{\hat{c}_{jN}(\cdot)\}$ is a sequence of real-valued (random) functions converging asymptotically to the corresponding function $c_j(\cdot)$ as $N \to \infty$. We assume that the functions $c_j(\cdot)$ can be represented in the form of expected values:

$$c_j(z) = E_P[H_j(z, \boldsymbol{\xi})], \quad j = 1, \ldots, M.$$

These functions can then be estimated by the corresponding sample mean functions

$$\hat{c}_{jN}(z) = \frac{1}{N} \sum_{i=1}^{N} H_j(z, \xi_i).$$

For simplicity, we will consider the more general parametric mathematical programming problem (that will also be of interest in the next section)

$$\begin{aligned} &\min_{z\in V} \ \hat{g}(z, \boldsymbol{\epsilon}), \\ \text{subject to} \quad &\hat{c}_j(z, \boldsymbol{\epsilon}) \geq 0, \quad j = 1, \ldots, k, \\ &\hat{c}_j(z, \boldsymbol{\epsilon}) = 0, \quad j = k+1, \ldots, M, \end{aligned} \tag{3.6}$$

where $\boldsymbol{\epsilon}$ is a random vector of parameters giving perturbations of the program (3.6), and $g(\cdot)$, $\hat{g}(\cdot, \boldsymbol{\epsilon})$, $c_j(\cdot)$, $\hat{c}_j(\cdot, \boldsymbol{\epsilon})$ are assumed to be twice continuously differentiable with respect to z. We will assume that the perturbation is of the form

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(z, \overline{\xi}) = \left(\boldsymbol{\epsilon}_g, \ \boldsymbol{\epsilon}_{c_1}, \ldots, \boldsymbol{\epsilon}_{c_M}, \ \boldsymbol{\epsilon}^T_{\nabla g}, \ \boldsymbol{\epsilon}^T_{\nabla c_1}, \ldots, \boldsymbol{\epsilon}^T_{\nabla c_M}\right)^T,$$

where each component is a function from $\mathbb{R}^m \times \prod_{i=1}^{\infty} \mathbb{R}^k$ to $\mathbb{R}$ or $\mathbb{R}^m$, and

$$\begin{aligned} &\hat{g}(z, \boldsymbol{\epsilon}) = g(z) + \boldsymbol{\epsilon}_g, \quad \nabla_z \hat{g}(z, \boldsymbol{\epsilon}) = \nabla_z g(z) + \boldsymbol{\epsilon}_{\nabla g}, \\ &\hat{c}_j(z, \boldsymbol{\epsilon}) = c_j(z) + \boldsymbol{\epsilon}_{c_j}, \ \nabla_z \hat{c}_j(z, \boldsymbol{\epsilon}) = \nabla_z c_j(z) + \boldsymbol{\epsilon}_{\nabla c_j}, \quad j = 1, \ldots, M. \end{aligned}$$

We also define $\epsilon_N(z, \overline{\xi})$ as

$$\epsilon_N(z, \overline{\xi}) = \begin{pmatrix} \hat{g}_N(z) - g(z) \\ \hat{c}_{jN}(z) - c_j(z), \quad j = 1, \ldots, M \\ \nabla_z \hat{g}_N(z) - \nabla_z g(z) \\ \nabla_z \hat{c}_{jN}(z) - \nabla_z c_j(z), \quad j = 1, \ldots, M \end{pmatrix},$$

and we will denote the corresponding random vector by $\boldsymbol{\epsilon}_N(z, \overline{\boldsymbol{\xi}})$. We will assume that $\boldsymbol{\epsilon}_N(z, \overline{\boldsymbol{\xi}})$ converges uniformly on V to the null function almost surely as N tends to infinity. In other terms, we assume that the ULLN holds for the objective and the constraints, as well as for the corresponding derivatives. We finally assume that the feasible sets for

the original and approximate problems (3.6) are nonempty. The Lagrangian functions associated with (3.4) and (3.6) are respectively

$$\mathcal{L}(z,\lambda) = g(z) - \sum_{j=1}^{M} [\lambda]_j c_j(z) \quad \text{and} \quad L(z,\lambda,\boldsymbol{\epsilon}) = \hat{g}(z,\boldsymbol{\epsilon}) - \sum_{j=1}^{M} [\lambda]_j \hat{c}_j(z,\boldsymbol{\epsilon}).$$

Let $z^*(\boldsymbol{\epsilon})$ denote a first-order critical point for program (3.6), and assume that $z^*(\boldsymbol{\epsilon})$ belongs to the interior of V, which we denote by $\overset{o}{V}$. Then there exist Lagrange multipliers $\lambda^*(\boldsymbol{\epsilon})$ such that $(z^*(\boldsymbol{\epsilon}), \lambda^*(\boldsymbol{\epsilon}))$ satisfy the KKT solutions; in other terms $(z^*(\boldsymbol{\epsilon}), \lambda^*(\boldsymbol{\epsilon}))$ is solution of the system

$$\begin{aligned}
&\nabla_z L(z,\lambda,\boldsymbol{\epsilon}) = 0, \\
&[\lambda]_j \, \hat{c}_j(z,\boldsymbol{\epsilon}) = 0, \quad j = 1,\ldots,M, \\
&\hat{c}_j(z,\boldsymbol{\epsilon}) = 0, \quad j = k+1,\ldots,M, \\
&\hat{c}_j(z,\boldsymbol{\epsilon}) \geq 0, \quad j = 1,\ldots,k, \\
&[\lambda]_j \, (\boldsymbol{\epsilon}) \geq 0, \quad j = 1,\ldots,k.
\end{aligned}$$

Consider now a particular sampling process $\overline{\xi}$. To clarify the dependency of the first-order critical points on the sampling process, we write $z_N^*(\overline{\xi})$ for $z^*(\epsilon_N)$ and $\lambda_N^*(\overline{\xi})$ for $\lambda^*(\epsilon_N)$. Let $\mathcal{Z}\left(\{z_N^*(\overline{\xi})\}\right)$ represent the set of accumulation points of the sequence $\{z_N^*(\overline{\xi})\}_{N=1}^{\infty}$. The compacity of V implies that for each $\overline{\xi}$, $\mathcal{Z}\left(\{z_N^*(\overline{\xi})\}\right)$ is not empty. We now prove almost-sure first-order convergence for the general case.

Theorem 3.2. *Assume that for almost every $\overline{\zeta}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, $\epsilon_N(z,\overline{\zeta}) \to 0$ uniformly on the compact set V, as $N \to \infty$, and let $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$ satisfying this uniform convergence assumption. If $z^* \in \mathcal{Z}\left(\{z_N^*(\overline{\xi})\}\right)$ belongs to $\overset{o}{V}$, and one subsequence $\{z_\ell^*(\overline{\xi})\}_{\ell=1}^{\infty} \subseteq \{z_N^*(\overline{\xi})\}_{N=1}^{\infty}$ converging to z^* is associated with a sequence of Lagrange multipliers $\{\lambda_\ell^*(\overline{\xi})\}_{\ell=1}^{\infty}$ having at least one limit point, then z^* is a first-order critical point for* (3.4)*.*

Proof. From our assumptions, the sequence $\left\{(z_\ell^*(\overline{\xi}), \lambda_\ell^*(\overline{\xi})\right\}$, $\ell = 1,\ldots,\infty$, has some limit point (z^*, λ^*). The uniform convergence property implies that (z^*, λ^*) satisfies the KKT conditions for the true problem, so that, $z^*(\in \overset{o}{V})$ is a first-order critical point for the original problem. □

Note that Lagrange multipliers convergence assumption always holds if the multipliers remain bounded. This stronger assumption allows the use of Theorem 3.1 to prove first-order criticality of the limit points, as in Shapiro [38]. Let $\mu := (z,\lambda) \in \mathbb{R}^{m+M}$ and $\mathcal{K} := \mathbb{R}^m \times \mathbb{R}_+^k \times \mathbb{R}^{M-k} \subset \mathbb{R}^{m+M}$. Define

$$\phi(\mu) = \left(\nabla_z \mathcal{L}(z,\lambda), c_{k+1}(z), \ldots, c_M(z)\right),$$

and

$$\hat{\phi}_N(\mu) = \left(\nabla_z L(z,\lambda,\boldsymbol{\epsilon}_N), \hat{c}_{k+1}(z,\boldsymbol{\epsilon}_N), \ldots, \hat{c}_M(z,\boldsymbol{\epsilon}_N)\right).$$

The variational inequality $\phi(\mu) \in \mathcal{N}_{\mathcal{K}}(\mu)$ then represents the KKT optimality conditions for the true optimization problem, is $S^* \subseteq \overset{o}{V}$, and Theorem 3.1 then implies almost sure first-order criticality, with $\Gamma(\mu) := \mathcal{N}_{\mathcal{K}}(\mu)$. Assumptions (a) and (d) of Theorem 3.1 are satisfied since $\boldsymbol{\epsilon} \to 0$ almost surely, and assumption (c) holds because the feasible sets for the original and approximating problems are nonempty.

4. Second-order convergence

4.1. Deterministic constraints

Provided that we strengthen our assumptions, we can also show that, almost surely, there exists limit points in $\mathcal{Z}\left(\{z_N^*(\overline{\boldsymbol{\xi}})\}\right)$ which are local minimizers. We first consider the case where S is deterministic and assume that, for a particular sampling process $\overline{\xi}$, $z_N^*(\overline{\xi})$ is a local minimizer of $\hat{g}_N(z)$ over S. This is to say that

$$\exists\, \delta_N(\overline{\xi}) \text{ s.t. } \forall z \in B\left(z_N^*(\overline{\xi}), \delta_N(\overline{\xi})\right) \cap S, \quad \hat{g}_N\left(z_N^*(\overline{\xi})\right) \leq \hat{g}_N(z), \tag{4.1}$$

where $B(x, d)$ is the open ball centered at x and of radius d.

In order to show that $z^* \in \mathcal{Z}\left(\{z_N^*(\overline{\xi})\}\right)$ is a local minimizer of $g(\cdot)$ over S, we must therefore prove that for some subsequence $\{z_\ell^*(\overline{\xi})\}_{\ell=1}^{\infty} \subseteq \{z_N^*(\overline{\xi})\}_{N=1}^{\infty}$ converging to z^*, the neighbourhood in which $z_\ell^*(\overline{\xi})$ is a local minimizer does not shrink to a singleton as $\ell \to \infty$. We express this requirement by the following technical assumption.

A.4 There exists some subsequence $\left\{z_\ell^*(\overline{\xi})\right\}_{\ell=1}^{\infty}$ converging to some z^*, with $\left\{z_\ell^*(\overline{\xi})\right\} \subseteq \left\{z_N^*(\overline{\xi})\right\}$, and some constants $\delta_{z^*\overline{\xi}} > 0$ and $\ell_{z^*\overline{\xi}} > 0$ such that for all $\ell \geq \ell_{z^*\overline{\xi}}$,

$$\forall z \in B\left(z_\ell^*(\overline{\xi}), \delta_{z^*\overline{\xi}}\right) \cap S, \quad \hat{g}_\ell\left(z_\ell^*(\overline{\xi})\right) \leq \hat{g}_\ell(z).$$

This allows us to write a basic second-order convergence theorem.

Theorem 4.1. *Assume that **A.0–A.3** hold. Then for almost every sampling process $\overline{\xi}$, $\left\{z^* \in \mathcal{Z}\left(\{z_N^*(\overline{\xi})\}\right)$ satisfying **A.4**$\right\}$ is a set of local minima of $g(\cdot)$ over S.*

Proof. Consider some realization $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$ and $z^* \in \mathcal{Z}\left(\{z_N^*(\overline{\xi})\}\right)$ such that

$$\sup_{z \in S} \left|\hat{g}_N(z) - g(z)\right| \to 0, \tag{4.2}$$

$$\hat{g}_\ell(z_\ell^*) \to g(z^*), \tag{4.3}$$

for some subsequence $\{\hat{g}_\ell(z_\ell^*)\}_{\ell=1}^{\infty} \subseteq \{\hat{g}_N(z_N^*)\}_{N=1}^{\infty}$. From the ULLN property, almost every $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$ allows these requirements to be satisfied. Assume moreover that **A.4** is fulfilled at z^*. For simplicity of notation, we will write δ instead of $\delta_{z^*,\overline{\xi}}$. Let z' be a minimizer of g in $\mathcal{K} := B\left(z^*, \frac{\delta}{2}\right) \cap S$. We first show that, for ℓ sufficiently large,

$$z_\ell^* \in \mathcal{K} \subseteq B(z_\ell^*, \delta). \tag{4.4}$$

Since z^* is the limit point of $\{z_\ell^*\}_{\ell=0}^{\infty}$, the first inclusion of (4.4) must hold for ℓ large enough. Consider now $z \in \mathcal{K}$. We have that $\left|z - z_\ell^*\right| \leq |z - z^*| + \left|z^* - z_\ell^*\right|$, and thus that, for large l,

$$\left|z - z_\ell^*\right| < \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Therefore $z \in B(z_\ell^*, \delta)$, completing our proof that (4.4) holds for ℓ sufficiently large. We now verify that

$$\left|\hat{g}_\ell(z_\ell^*) - g(z')\right| \to 0. \tag{4.5}$$

Assume first that $\hat{g}_\ell(z_\ell^*) \leq g(z')$. Since z' minimizes $g(\cdot)$ in $\mathcal{K}$ and, from (4.4), $z_\ell^* \in \mathcal{K}$, we have that

$$0 \leq \left|\hat{g}_\ell(z_\ell^*) - g(z')\right| = g(z') - \hat{g}_\ell(z_\ell^*) \leq g(z_\ell^*) - \hat{g}_\ell(z_\ell^*) \leq \sup_{z \in S} \left|\hat{g}_\ell(z) - g(z)\right|.$$

The limit (4.2) then implies (4.5). Assume now that $\hat{g}_\ell(z_\ell^*) \geq g(z')$. Since $z' \in \mathcal{K}$, we deduce from the second part of (4.4) and **A.4** that $\hat{g}_\ell(z_\ell^*) \leq \hat{g}_\ell(z')$. Therefore

$$0 \leq \hat{g}_\ell(z_\ell^*) - g(z') \leq \hat{g}_\ell(z') - g(z') \leq \sup_{z \in S} \left|\hat{g}_\ell(z) - g(z)\right|,$$

and we again deduce (4.5). Taking (4.3) into account, we have that $g(z') = g(z^*)$, and $g(z^*) \leq g(z)$, for all $z \in \mathcal{K}$. In other terms, z^* is a local solution of problem (2.1). Since this reasoning is valid for almost every sampling process $\overline{\xi}$, our proof is complete. □

Classical results, where global minimizers are considered, express that $d(z_N^*, S^*)$ converges almost surely to zero as $N \to \infty$, where S^* is the set of minimizers of the true problem (see for instance Theorem 3.1). Robinson [33] shows that, under mild regularity conditions, if the true problem has a complete local minimizing (CLM) set with respect to a nonempty open bounded set $\mathcal{G}$, then for large N, the approximating problem has almost surely a CLM set with respect to $\mathcal{G}$ such that the distance between the CLM set associated with the true problem and the one corresponding to the approximating problem tends to 0 as $N \to \infty$. Moreover, the approximating infimum over the closure of $\mathcal{G}$ almost surely converges to a finite minimum for the true problem over the closure of $\mathcal{G}$. While this proves the existence of solutions for the approximating problem, it does not imply that the distance from (local) minimizer of the approximating problem to the set of true local minimizers converges almost surely to zero. Consider for instance the problem

$$\min_{z \in [-1,1]} z^3 - \frac{z}{2} E_P[\boldsymbol{\xi}], \tag{4.6}$$

where $\Xi = \{-1, 1\}$ and $P[\xi = -1] = P[\xi = 1] = 0.5$, so $E_P[\boldsymbol{\xi}] = 0$, and (4.6) has only one local minimizer, which is also global, at $z^* = -1$. The SAA problem is then

$$\min_{z \in [-1,1]} z^3 - \frac{z}{2N} \sum_{i=1}^{N} \xi_i. \tag{4.7}$$

The problem (4.7) has two (isolated) local minimizers,

$$\left\{-1, \sqrt{\frac{\sum_{i=1}^{N} \xi_i}{6N}}\right\},$$

when $\sum_{i=1}^{N} \xi_i > 0$. We have that $P\left[\sum_{i=1}^{N} \xi_i > 0\right] \to 0.5$ when $N \to \infty$, but from the strong law of large numbers, $\frac{1}{6N} \sum_{i=1}^{N} \xi_i \to \frac{1}{6} E_P[\boldsymbol{\xi}] = 0$ almost surely as $N \to \infty$. However zero is a saddle point of the true problem (4.6), not a minimizer, even locally, and the distance to $S^* = \{-1\}$ is then equal to 1. Note that in this example the ULLN holds for the objective function as well as for all its derivatives.

It can be shown, under some mild regularity conditions, that the SAA has almost surely a solution in a neighbourhood of a local solution of the true solution, when N is sufficiently large (Shapiro [38]). However, the previous example illustrates that care must be exercised when solving the SAA problem for N fixed since we can find approximating local minimizers that are not close to true local minimizers.

4.2. Stochastic constraints

Assumption **A.4** is somewhat artificial and it is thus of interest to search for more elegant conditions. While our arguments will be similar to those presented in perturbation analysis, as for instance in Rubinstein and Shapiro [34], it is important to note that perturbation analysis assumes the existence of a solution for the true problem and then studies the existence and behaviour of solutions for the perturbed problem in a neighbourhood of this original solution. At variance with this approach, we focus here on conditions under which the limit point of a sequence of approximating solutions is itself a solution of the true problem. The difference will be more formally illustrated at the end of the section.

We consider the case where the feasible set is described by a set of equality and inequality constraints, as in (3.4). As before, we assume that $\boldsymbol{\epsilon}_N$ converges to zero uniformly on V almost surely. Consider a particular sampling process $\bar{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, and $z^* \in \mathcal{Z}\left(\{z_N^*(\bar{\xi})\}\right)$. Under some conditions, if the subsequence $\{z_\ell^*\}_{\ell=1}^{\infty}$ converges to z^*, the associated Lagrange vectors sequence $\{\lambda_\ell^*\}$ also converges to some Lagrange multiplier vector λ^* associated with z^* for the true problem, as expressed below.

Lemma 4.1. *Consider a particular sampling process $\bar{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$ such that $\epsilon_N\left(z, \bar{\xi}\right) \to 0$ uniformly on V as $N \to \infty$. If $z^* \in \mathcal{Z}\left(\{z_N^*(\bar{\xi})\}\right) \cap \overset{o}{V}$, $\{z_\ell^*\}_{\ell=1}^{\infty} \subseteq \{z_N^*\}_{N=1}^{\infty} \cap \overset{o}{V}$ converges to z^*, and there is a unique Lagrange multipliers vector λ^* associated with z^* satisfying the KKT conditions, then $\lambda_\ell^*(\bar{\xi})$ converges to λ^* as $\ell \to \infty$.*

Proof. From the uniqueness of λ^*, the Mangasarian-Fromowitz constraint qualification (MFCQ) holds at z^*, and therefore in a neighbourhood of z^* (while the converse is not necessarily true, as shown by Gugat [21]). Hence the Lagrange multipliers are uniformly bounded for ϵ close to zero. It is therefore sufficient to show that every limit point of the sequence $\{\lambda_\ell^*(\bar{\xi})\}$, $\ell = 1, \ldots, \infty$, is equal to λ^*. Let λ' be such a limit point. By continuity, (z^*, λ') satisfies the KKT conditions, so λ' is equal to λ^*. □

The uniqueness of λ^* can be ensured with a suitable constraint qualification, as the linear independence constraint qualification (LICQ). This constraint qualification will be particularly convenient for our discussion. First of all we recall the notion of active set. Consider the program (3.4). The active set $\mathcal{A}(z)$ at any feasible z is the union of set of indices of equality constraints with the indices of active inequality constraints:

$$\mathcal{A}(z) = \{i \in \{1, \ldots, k\} \mid c_i(z) = 0\} \cup \{k+1, \ldots, M\}.$$

Definition 4.1. *Given the point z^* and the active set $\mathcal{A}(z^*)$ we say that the linear independence constraint qualification (LICQ) holds at z^* if the set of active constraint gradients $\{\nabla c_j(z^*),\ j \in \mathcal{A}(z^*)\}$ is linearly independent.*

For a discussion of LICQ and other constraint qualifications, see for instance Nocedal and Wright [30]. Another useful concept for our purposes is the strict complementarity condition.

Definition 4.2. *Given z^* and a vector λ^* satisfying the KKT conditions, we say that the strict complementarity condition holds if exactly one of $[\lambda]_j^*$ and $c_j(z^*)$ is zero for each index $j = 1, \ldots, k$, i.e. we have that $[\lambda]_j^* > 0$ for each $j \in \{1, \ldots, k\} \cap \mathcal{A}(z^*)$.*

Consider again a particular sampling process $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$ such that $\epsilon_N(z, \overline{\xi}) \to 0$ as $N \to \infty$. If the assumptions of Lemma 4.1 hold for some subsequence $\{z_\ell^*\}_{\ell=1}^\infty \to z^*$, the Lagrangian gradient $\nabla L\left(z_\ell^*(\overline{\xi}), \lambda_\ell^*(\overline{\xi})\right)$ converges to $\nabla \mathcal{L}(z^*, \lambda^*)$, when ℓ tends to infinity, with $\lambda_\ell^*(\overline{\xi}) \to \lambda^*$. Assume that the strict complementarity condition holds at z^* for problem (3.6). We obtain that, for ℓ sufficiently large, $[\lambda_\ell^*]_j(\overline{\xi})$, $j \in \{1, \ldots, k\} \cap \mathcal{A}(z^*)$, are strictly positive and hence the corresponding constraints are active at $z_\ell^*(\overline{\xi})$. Moreover, since $\epsilon_N(z, \overline{\xi}) \to 0$, $\hat{c}_j\left(z_\ell^*(\overline{\xi}), \epsilon_\ell(z_\ell^*(\overline{\xi}), \overline{\xi})\right) \to c_j(z^*)$, and, for ℓ large enough, $\mathcal{A}\left(z_\ell^*(\overline{\xi})\right) = \mathcal{A}(z^*)$, so the strict complementarity condition holds at $\left(z_\ell^*(\overline{\xi}), \lambda_\ell^*(\overline{\xi})\right)$ for problem (3.5). This allows us to state the theorem below.

Theorem 4.2 (Second-order convergence). *Assume that, for almost every sampling process $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, there exists some $z^* \in \mathcal{Z}\left(\{z_N^*(\overline{\xi})\}\right) \cap \overset{o}{V}$ associated with a unique Lagrange multipliers vector λ^* and some subsequence $\{z_\ell^*\}_{\ell=1}^\infty \subseteq \{z_N^*\}_{N=1}^\infty$, such that $z_\ell^*(\overline{\xi}) \to z^*$, and*

(a) $\epsilon_N\left(z_N^*(\overline{\xi}), \overline{\xi}\right) \to 0$ *uniformly on V, as* $N \to \infty$,
(b) $z_N^*(\overline{\xi}) \in \overset{o}{V}$, $N = 1, \ldots,$
(c) $\nabla_{zz}^2 \hat{g}\left(z_\ell^*(\overline{\xi}), \epsilon_\ell(z_\ell^*(\overline{\xi}), \overline{\xi})\right) \to \nabla_{zz}^2 g(z^*)$ *as* $\ell \to \infty$,
(d) $\nabla_{zz}^2 \hat{c}_j\left(z_\ell^*(\overline{\xi}), \epsilon_\ell(z_\ell^*(\overline{\xi}), \overline{\xi})\right) \to \nabla_{zz}^2 c_j(z^*)$ $(j = 1, \ldots, M)$ *as* $\ell \to \infty$.

Suppose also that the strict complementarity condition and the LICQ hold at (z^*, λ^*) for (3.4)*. Then, for almost every sampling process $\overline{\xi}$,*

(i) the LICQ holds at $z_\ell^*(\overline{\xi})$, for ℓ large enough,
(ii) (z^*, λ^*) satisfies the second-order necessary condition for (3.4)*:*

$$w^T \nabla_{zz}^2 \mathcal{L}(z^*, \lambda^*) w \geq 0, \text{ for all } w \in Null\left[\nabla_z c_j(z^*)^T\right]_{j \in \mathcal{A}(z^*)}. \tag{4.8}$$

If furthermore there exists some constant $\alpha_{\overline{\xi}} > 0$ *such that, for all* ℓ *large enough,*

$$w^T \nabla^2_{zz} L_\ell \left(z^*_\ell(\overline{\xi}), \lambda^*_\ell(\overline{\xi})\right) w > \alpha_{\overline{\xi}}, \text{ for all } w \in Null\left[\nabla_z \hat{c}_j(z^*_\ell\left(\overline{\xi}\right))^T\right]_{j\in\mathcal{A}(z^*_\ell(\overline{\xi}))}, \quad \|w\| = 1, \tag{4.9}$$

then (z^*, λ^*) *almost surely satisfies the second-order sufficient conditions for problem* (3.4)*, that is*

$$\text{(iii) } w^T \nabla^2_{zz} \mathcal{L}(z^*, \lambda^*) w > 0, \text{ for all } w \in Null\left[\nabla_z c_j(z^*)^T\right]_{j\in\mathcal{A}(z^*)}, \|w\| = 1. \tag{4.10}$$

In other terms, z^* *is an isolated local minimizer for the problem* (3.4)*.*

Proof. Consider a sampling process $\overline{\xi}$ and some limit point z^* such that the assumptions (a)–(d) are satisfied. For simplicity, we drop the dependence on $\overline{\xi}$ in our notation. In order to show (i), consider

$$\{\nabla_z c_j(z^*)\}_{j\in\mathcal{A}(z^*)}, \tag{4.11}$$

the matrix formed by the gradients of active constraints at z^* for (3.4). From the strict complementarity conditions and convergence of Lagrange multipliers, the active set of program (3.5) at z^*_ℓ is asymptotically the same as the active set of program (3.4) at z^*. Since $\epsilon_\ell \to 0$ uniformly on V, we have that the matrix formed by the active constraints of the perturbed problem,

$$\{\nabla_z \hat{c}_j(z^*_\ell, \epsilon_\ell)\}_{j\in\mathcal{A}(z^*_\ell)}, \tag{4.12}$$

converges to (4.11) as ℓ tends to infinity:

$$\{\nabla_z \hat{c}_j(z^*_\ell, \epsilon_\ell)\}_{j\in\mathcal{A}(z^*_\ell)} \to \{\nabla_z c_j(z^*)\}_{j\in\mathcal{A}(z^*)}. \tag{4.13}$$

The LICQ amounts to say that at least one square submatrix of (4.11) is nonsingular. From (4.13), the same is true for (4.12) for ℓ large enough, i.e. with have (i).

We now show (ii). From (4.13) we may associate a basis K_ℓ with the null space of (4.12) such that

$$K_\ell \to K, \tag{4.14}$$

where K is a basis of $\text{Null}[\nabla_z c_j(z)^T]_{j\in\mathcal{A}(z^*)}$ (see Gill et al. [17]). Using the strict complementarity condition and LICQ, the fact that $(z^*_\ell, \lambda^*_\ell)$ satisfies the second-order necessary conditions can now be expressed as

$$K_\ell^T \nabla^2_{zz} L(z^*_\ell, \lambda^*_\ell, \epsilon_\ell) K_\ell \text{ is positive semi-definite.}$$

From (4.14) and Assumptions (a)–(d), we have that

$$K_\ell^T \nabla^2_{zz} L(z^*_\ell, \lambda^*_\ell, \epsilon_\ell) K_\ell \to K^T \nabla^2_{zz} \mathcal{L}(z^*, \lambda^*) K,$$

as $\ell \to \infty$. We therefore have (4.8), and (ii) follows. The reasoning is identical for proving (iii), except that one now uses the lower bound $\alpha_{\overline{\xi}}$ on the eigenvalues of

$$K_\ell^T \nabla^2_{zz} L(z^*_\ell, \lambda^*_\ell, \epsilon_\ell) K_\ell$$

to obtain (4.10). □

Note that the LICQ and strict complementarity conditions imply that the minimizer is isolated while the second-order sufficient condition is usually used to characterize strict local minimizers. In other terms, there exists a neighbourhood $\mathcal{V}_S$ of z^* such that z^* is the only local minimizer in $\mathcal{V}_S$. Recall that isolated local minimizers are also strict local minimizers but that the inverse is not always true (Nocedal and Wright [30], page 14). If $z^* \in \mathcal{Z}\left(\{z^*_N(\overline{\xi})\}\right) \cap \overset{o}{V}$ is a strict but not isolated local minimizer, every neighbourhood of z^* contains other local minimizers that are candidates to be limit points of the sequences of solutions of the SAA problems (2.2) as N tends to infinity, and z^* can therefore be difficult to identify.

The non-degeneracy assumption (4.9) can also be replaced by requiring that the Jacobian of the equality equations involved in the KKT conditions associated with the program (3.4),

$$\begin{aligned} \nabla_z \mathcal{L}(z, \lambda) &= 0, \\ [\lambda]_j \, c_j(z) &= 0, \quad j = 1, \ldots, M, \\ c_j(z) &= 0, \quad j = k+1, \ldots, M \end{aligned} \tag{4.15}$$

is nonsingular at (z^*, λ^*), as shown in the corollary below.

Corollary 4.1. *Assume that, for almost every sampling process $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, there exists some $z^* \in \mathcal{Z}\left(\{z^*_N(\overline{\xi})\}\right) \cap \overset{o}{V}$ associated with a unique Lagrange multipliers vector λ^* and some subsequence $\{z^*_\ell\}_{\ell=1}^\infty \subseteq \{z^*_N\}_{N=1}^\infty$, such that $z^*_\ell(\overline{\xi}) \to z^*$, that assumptions (a)–(d) of Theorem 4.2 hold and that the strict complementarity condition holds at (z^*, λ^*) for the program* (3.4)*. Assume furthermore that the Jacobian of* (4.15) *is nonsingular at (z^*, λ^*). Then there exists almost surely some z^* in $\mathcal{Z}\left(\{z^*_N(\overline{\xi})\}\right) \cap \overset{o}{V}$, associated with some λ^*, that satisfies (4.10), the second-order sufficient conditions for the program* (3.4)*.*

Proof. Consider a sampling process $\overline{\xi}$ such that our assumptions are met. In order to prove second-order sufficiency, we rewrite the KKT conditions at z^* as

$$\begin{aligned} \nabla_z \mathcal{L}(z^*, \lambda^*) &= 0, \\ c_j(z^*) &= 0, \quad j \in \mathcal{A}(z^*), \\ [\lambda]^*_j &= 0, \quad j \notin \mathcal{A}(z^*), \end{aligned} \tag{4.16}$$

where we have used the strict complementarity condition when eliminating Lagrange multipliers in active inequality constraints. We renumber the active constraints such that

$\mathcal{A}(z^*) = \{1, \ldots, n_a\}$, while the inactive constraints are now numbered from $n_a + 1$ to M. The Jacobian of (4.17) is then

$$\begin{pmatrix} \nabla_{zz}\mathcal{L}(z^*) & -\nabla_z c_1(z^*) & \cdots & -\nabla_z c_{n_a}(z^*) & -\nabla_z c_{n_a+1}(z^*) & \cdots & -\nabla_z c_M(z^*) \\ \nabla_z^T c_1(s^*) & 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & 0 \\ \nabla_z^T c_{n_a}(z^*) & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 1 \end{pmatrix},$$

which is nonsingular if and only if

$$\begin{pmatrix} \nabla_{zz}\mathcal{L}(z^*) & \nabla_z c_1(z^*) & \cdots & \nabla_z c_{n_a}(z^*) \\ \nabla_z^T c_1(s^*) & 0 & \ldots & 0 \\ \vdots & & & \\ \nabla_z^T c_{n_a}(z^*) & 0 & \ldots & 0 \end{pmatrix}, \tag{4.17}$$

is itself nonsingular. From Sylvester's law of inertia, (4.17) is nonsingular if and only if

$$w^T \nabla^2_{zz}\mathcal{L}(z^*, \lambda^*) w \neq 0, \text{ for all } w \neq 0 \in \text{Null}[\nabla_z c_j(z^*)^T]_{j \in \mathcal{A}(z^*)}$$

(see Gould [18]). Note that (4.17) also implies that the LICQ holds at z^*. From Theorem 4.2, (z^*, λ^*) also satisfies the second-order necessary conditions for the program (3.4), so the second-order sufficient conditions (4.10) are verified. □

The converse of Theorem 4.2 can be obtained from classical results of perturbation analysis (Fiacco [14], Theorem 3.2.2), which we restate for completeness. More developments in the context of stochastic programming can be found in Rubinstein and Shapiro [34] and Shapiro [36].

Theorem 4.3. *Suppose that the following assumptions hold:*

(a) the functions defining (3.6) *are twice continuously differentiable in z and their gradients with respect to z and the constraints are once continuously differentiable in* ϵ *in a neighbourhood of* $(z^*, 0)$ $(z^* \in \overset{o}{V})$,

(b) the second-order sufficient conditions for a local minimum of (3.6) *hold at* z^*, *with associated Lagrange multipliers* λ^*,

(c) the LICQ holds at z^*,

(d) the strict complementarity condition holds at $(z^*, 0)$,

then

(i) z^* *is a local isolated minimum of* (3.6) *with* $\epsilon = 0$ *and the associated Lagrange multipliers* λ^* *are unique,*

(ii) for ϵ *in a neighbourhood of* 0, *there exists a unique, once continuously differentiable vector function* $\gamma(\epsilon) = (z(\epsilon), \lambda(\epsilon))^T$ *satisfying the second-order sufficient conditions for a local minimum of problem* (3.6) *such that* $\gamma(0) = (z^*, \lambda^*)^T$, *and hence* $z(\epsilon)$ *is a local isolated minimizer of problem* (3.6) *with associated unique Lagrange multipliers* $\lambda(\epsilon)$, *and*

(iii) for ϵ near 0, the set of active constraints is unchanged, strict complementarity conditions hold, and the LICQ holds at $z^*(\epsilon)$.

This theorem must of course be applied for a fixed sampling process $\overline{\xi}$, and the results of interest are only true almost surely. Note that the second-order sufficiency property is now taken as an assumption, so that z^* is assumed to be a local solution. More general results of perturbation analysis can also be obtained by using epi-continuity arguments and the concept of complete local minimizing set (Robinson [32, 33]).

5. Application to mixed logit problems

5.1. Convergence

We now apply the above results to the framework of mixed logit models. This is possible because we have already seen in Section 2.2 that the mixed-logit problem is a generalization of the stochastic program (2.1).

In this context, **A.0** should now be understood as the requirement that the different samples used to compute the choice probabilities are identically distributed and independent both for each individual and across them. We next note that, at variance with the stochastic programming case where the compacity assumption for the set S ensures that the solutions of problem (2.2) remain in a bounded domain of $\mathbb{R}^m$, our formulation of the mixed logit problem does not include any such safeguard. We therefore complete our assumptions with the following additional requirement.

A.5 For almost every sampling process $\overline{\gamma} = \{\gamma_{i,r}\}_{i=1,\ r=1}^{I,\ \infty}$, the solution $\theta_R^*(\overline{\gamma})$ of the simulated mixed-logit problem (2.8) remains in some convex compact set S (independent of $\overline{\gamma}$) for all R sufficiently large.

The set S can explicitly be expressed using convex constraints (bound constraints are typical) on the problem or be implicit for an unconstrained problem. In the latter case, **A.5** indicates that the solutions are almost surely uniformly bounded for sufficiently large sampling sizes, the bound being independent of the sampling process. Such an assumption is reasonable to avoid pathological cases where some components of θ_R^* converge towards infinity. As in the stochastic programming case, this assumption implies that, for almost every sampling process $\overline{\gamma}$, the corresponding sequence $\{\theta_R^*(\overline{\gamma})\}$ has limit points, whose set is represented by $\mathcal{Z}\left(\{\theta_R^*(\overline{\gamma})\}\right)$.

In order to obtain the first-order convergence, we also need to translate Assumptions **A.1**–**A.3**. We first ensure **A.1** and **A.2** by imposing suitable conditions on the I problem's components $E_P[L_{ij_i}(\boldsymbol{\gamma}, \theta)]$.

A.1ml The utilities $V_{ij}(\gamma, \cdot, x_{ij})$ $(i = 1, \ldots, I,\ j = 1, \ldots, J)$ are continuously differentiable for P-almost every γ.

That **A.1ml** implies **A.1** immediately results from the property of the logit formula, which ensures that

$$\frac{\partial}{\partial [\theta]_t} L_{ij_i}(\gamma, \theta) = L_{ij_i}(\gamma, \theta) \sum_{s \neq j_i} L_{is}(\gamma, \theta) \times \frac{\partial}{\partial [\theta]_t} \left(V_{ij_i}\left(\gamma, \theta, x_{ij_i}\right) - V_{is}(\gamma, \theta, x_{is}) \right). \tag{5.1}$$

The assumption **A.2** is automatically satisfied since $|L_{ij_i}(\gamma, \theta)| \leq 1$ for all θ and 1 is obviously P-integrable with unit expectation. We obtain from **A.5** and **A.1ml** that, for almost every $\overline{\gamma}$, if $\{\theta_\ell^*\} \subseteq \{\theta_R^*\}$ converges to some θ^*, then for all individuals i $(i = 1, \ldots, I)$,

$$SP_{ij_i}^{\ell}(\theta_\ell^*(\overline{\gamma})) \to P_{ij_i}(\theta^*) \quad \text{and } SLL^{\ell}(\theta_\ell^*(\overline{\gamma})) \to LL(\theta^*).$$

We now turn to Assumption **A.3** by examining the derivatives of the true and SAA problems. For $t = 1, \ldots, m$, we have

$$\frac{\partial}{\partial[\theta]_t} LL(\theta) = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{E_P\left[L_{ij_i}(\boldsymbol{\gamma}, \theta)\right]} \frac{\partial}{\partial[\theta]_t} E_P\left[L_{ij_i}(\boldsymbol{\gamma}, \theta)\right],$$

and

$$\frac{\partial}{\partial[\theta]_t} SLL(\theta) = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{SP_{ij_i}^{R}(\theta)} \frac{1}{R} \sum_{r=1}^{R} \frac{\partial}{\partial[\theta]_t} L_{ij_i}(\gamma_{i,r}, \theta).$$

Assumption **A.3** now becomes

A.3ml For $t = 1, \ldots, m$, $\frac{\partial}{\partial[\theta]_t} L_{ij_i}(\gamma, \theta)$ $(i = 1, \ldots, I)$ is dominated by a P-integrable function.

From (5.1) we see that this property holds in particular if

A.3ml' For $t = 1, \ldots, m$, $\frac{\partial}{\partial[\theta]_t} V_{ij}\left(\gamma, \theta, x_{ij}\right)$ $(i = 1, \ldots, I,\ j = 1, \ldots, J)$ is dominated by a P-integrable function.

If the utilities are linear in θ, as is often the case in applications, the derivatives are independent of θ. All what we then have to assume is that the expectation of the absolute partial derivatives is finite, which is usually not restrictive. If the utilities are nonlinear, we observe that **A.3ml'** is satisfied if, for $t = 1, \ldots, m$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $E_P[K(\boldsymbol{\gamma})]$ is finite, where $K(\gamma) = \sup_\theta \left| \frac{\partial}{\partial[\theta]_t} V_{ij}\left(\gamma, \theta, x_{ij}\right) \right|$. Under **A.1ml**, and the assumption that $\theta \in S$, where S is compact, $K(\gamma)$ is finite for P-almost every γ, and its expectation is usually finite. We deduce that for almost every $\overline{\gamma}$, if $\{\theta_\ell^*\} \subseteq \{\theta_R^*\}$ converges to some θ^*.

$$\nabla_\theta SLL^{\ell}(\theta_\ell^*(\overline{\gamma})) \to \nabla_\theta LL(\theta^*),$$

as $\ell \to \infty$. We can again apply Theorem 3.1 in order to deduce the following result.

Theorem 5.1 (First-order convergence for mixed logit). *Assume that **A.0**, **A.5**, **A.1ml** and **A.3ml** hold. Then, for almost every sampling process $\overline{\gamma} = \{\gamma_{i,r}\}$, if $\theta^* \in \mathcal{Z}\left(\{\theta_R^*(\overline{\gamma})\}\right)$, θ^* is a first-order critical point of problem* (2.7).

We have therefore proved that any limit point of a sequence of first-order critical simulated estimators is almost surely a first-order critical solution for the true maximum likelihood problem, allowing the inclusion of convex constraints on θ. Classical results (see Chapter 10 of Train [41]) show convergence in distribution and in probability

asymptotically when the population size increases. The asymptotic behaviour is briefly discussed in Section 5.3.

The extension of Theorem 4.2 establishing second-order convergence to the mixed-logit problem is immediate, as well as Theorem 4.1, as long as the corresponding assumptions are made. In particular, assuming that the utilities are twice continuously differentiable P-almost surely, we have that

$$\frac{\partial}{\partial[\theta]_u \partial[\theta]_t} LL(\theta) = \frac{1}{I} \sum_{i=1}^{I} \frac{\frac{\partial}{\partial[\theta]_u} \frac{\partial}{\partial[\theta]_t} E_P \left[L_{ij_i}(\boldsymbol{\gamma}, \theta) \right]}{P_{ij_i}(\theta)} - \frac{1}{I} \sum_{i=1}^{I} \frac{\frac{\partial}{\partial[\theta]_u} E_P \left[L_{ij_i}(\boldsymbol{\gamma}, \theta) \right] \frac{\partial}{\partial[\theta]_t} E_P \left[L_{ij_i}(\boldsymbol{\gamma}, \theta) \right]}{\left(P_{ij_i}(\theta) \right)^2},$$

and

$$\begin{aligned} &\frac{\partial}{\partial[\theta]_u \partial[\theta]_t} SLL(\theta) \\ &= \frac{1}{I} \sum_{i=1}^{I} \frac{\frac{1}{R} \sum_{r=1}^{R} \frac{\partial}{\partial[\theta]_u} \frac{\partial}{\partial[\theta]_t} L_{ij_i}(\gamma_{i,r}, \theta)}{SP_{ij_i}^R(\theta)} \\ &\quad - \frac{1}{I} \sum_{i=1}^{I} \frac{\left(\frac{1}{R} \sum_{r=1}^{R} \frac{\partial}{\partial[\theta]_u} L_{ij_i}(\gamma_{i,r}, \theta) \right) \left(\frac{1}{R} \sum_{r=1}^{R} \frac{\partial}{\partial[\theta]_t} L_{ij_i}(\gamma_{i,r}, \theta) \right)}{SP_{ij_i}^R(\theta)^2}. \end{aligned}$$

We therefore have to require that for almost every $\overline{\gamma}$, if $\{\theta_\ell^*\} \subseteq \{\theta_R^*\} \to \theta^*$,

$$\frac{1}{\ell} \sum_{r=1}^{\ell} \frac{\partial}{\partial[\theta]_u} \frac{\partial}{\partial[\theta]_t} L_{ij_i}(\gamma_{i,r}, \theta_\ell^*(\overline{\gamma})) \longrightarrow \frac{\partial}{\partial[\theta]_u} \frac{\partial}{\partial[\theta]_t} E_P \left[L_{ij_i}(\boldsymbol{\gamma}, \theta^*) \right],$$

for $i = 1, \ldots, I, t, u = 1, \ldots, m$, which is usually the case, for instance if the second-order derivatives are dominated by integrable functions.

5.2. Estimation of the simulation's variance and bias

We now further investigate the question of estimating the error made by using the SAA problem (2.8) instead of the true problem (2.7) as a function of the sampling size R, for a fixed population size. Due to the stochastic nature of the approximation, the size of the error can only be assessed by providing a (hopefully high) probability that it is within some confidence interval asymptotically centered at zero and of radius Δ. In practice, we first fix some probability level $\alpha > 0$ and determine the value of Δ such that, for given θ (and dropping the dependence on the sampling process $\overline{\gamma}$), $P\left[\left|LL(\theta) - SLL^R(\theta)\right| \leq \Delta\right] \geq \alpha$. Developing this expression we have that $\left|LL(\theta) - SLL^R(\theta)\right|$ is smaller than Δ if and only if

$$\left| \frac{1}{I} \sum_{i=1}^{I} \ln P_{ij_i}(\theta) - \frac{1}{I} \sum_{i=1}^{I} \ln SP_{ij_i}^R(\theta) \right| \leq \Delta.$$

Consider now individual i. We are interested in the asymptotic behaviour of

$$\ln P_{ij_i}(\theta) - \ln SP^R_{ij_i}(\theta),$$

for a given θ (such as the solution of the SAA problem). Since the logarithm is continuously differentiable on $\mathbb{R}_0^+$ and $E_P\left[L_{ij_i}(\boldsymbol{\gamma},\theta)^2\right]$ is finite, we can use the Delta method (see for instance Borovkov [10], page 44, for the one-dimensional case or Rubinstein and Shapiro [34] Section 6.3, for the multi-dimensional case) to conclude that

$$\sqrt{R}\left(\ln P_{ij_i}(\theta) - \ln SP^R_{ij_i}(\theta)\right) \Rightarrow \frac{d}{dP_{ij_i}} \ln P_{ij_i}(\theta) N\left(0, \sigma^2_{ij_i}(\theta)\right),$$

where $\sigma^2_{ij_i}(\theta)$ is the variance of $L_{ij_i}(\gamma,\theta)$. As the samples are independent between individuals, so are the normal distributions in this last limit, and we thus have that

$$\sqrt{R}\left(LL(\theta) - SLL^R(\theta)\right) \Rightarrow N\left(0, \frac{1}{I^2}\sum_{i=1}^{I} \frac{\sigma^2_{ij_i}(\theta)}{\left(P_{ij_i}(\theta)\right)^2}\right). \tag{5.2}$$

Let α_δ be the quantile of a $N(0,1)$ associated with some level of significance δ, i.e. $P[-\alpha_\delta \le X \le \alpha_\delta] = \delta$, where $X \sim N(0,1)$. The associated asymptotic value of the confidence interval radius Δ is then given by

$$\Delta^R_\delta(\theta) = \alpha_\delta \frac{1}{I}\sqrt{\sum_{i=1}^{I} \frac{\sigma^2_{ij_i}(\theta)}{R\left(P_{ij_i}(\theta)\right)^2}}. \tag{5.3}$$

Typically, one chooses $\alpha_{0.9} \approx 1.64$ of $\alpha_{0.95} \approx 1.96$. In practice we evaluate this accuracy $\Delta^R_\delta(\theta)$ by taking the SAA estimators $\sigma^R_{ij_i}(\theta)$ and $SP^R_{ij_i}(\theta)$, where $\sigma^R_{ij_i}(\theta)$ is the sample standard deviation of $L_{ij_i}(\gamma_{i,r},\theta)$, $r = 1,\ldots,R$.

Equation (5.3) gives us important information on the quality of the approximation. The accuracy can be improved if we take a larger sampling size R, but, as in other basic Monte Carlo methods, the convergence is only in $O\left(\sqrt{R}\right)$ (Fishman [15], page 8). However, the population size also has an influence on the quality of the approximation. First of all, we note that

$$0 \le \Delta^R_\delta(\theta) \le \alpha_\delta \frac{1}{I\sqrt{R}} \sum_{i=1}^{I} \frac{\sigma_{ij_i}(\theta)}{P_{ij_i}(\theta)}.$$

If the total population, denoted by $\mathcal{I}$, is assumed to be infinite, then we may consider a population of size I as an independent and identically distributed sample within it, drawn following some probability distribution $P_{\mathcal{I}}$. From now on, we also assume that $\frac{\sigma_{ij_i}(\theta)}{P_{ij_i}(\theta)}$ has finite mean and variance for all θ in S. We also note that for every individual i, $\sigma_{ij_i}(\theta)/P_{ij_i}(\theta)$ is continuous on S, since $L_{ij_i}(\boldsymbol{\gamma},\theta)$ has continuous first and second-order moments (and $L_{ij_i}(\boldsymbol{\gamma},\theta)$ is strictly positive on S, so is $P_{ij_i}(\theta)$). $L_{ij_i}(\gamma,\theta)$ is indeed P-dominated by the constant function $f(\gamma) = 1$, and is continuous for every indivual i

from **A.1ml**. We obtain from the strong law of large numbers that, for $P_{\mathcal{I}}$-almost every sampling process, for fixed R and I tending to ∞,

$$0 \leq \Delta_{\delta}^{R}(\theta) \leq \frac{\alpha_{\delta}}{\sqrt{R}} E_{P_{\mathcal{I}}}\left[\frac{\sigma_{ij_i}(\theta)}{P_{ij_i}(\theta)}\right] \leq \alpha_{\delta}\frac{\kappa}{\sqrt{R}}, \tag{5.4}$$

where κ is defined as

$$\kappa = \sup_{\theta \in S} E_{P_{\mathcal{I}}}\left[\frac{\sigma_{ij_i}(\theta)}{P_{ij_i}(\theta)}\right].$$

We assume that κ is finite, which is satisfied if $\sigma_{ij_i}(\theta)/P_{ij_i}(\theta)$ is dominated by a $P_{\mathcal{I}}$-integrable function (for instance a constant function) on S, since the expectation of $\sigma_{ij_i}(\theta)/P_{ij_i}(\theta)$ is then continuous on the compact set S.

Inequalities (5.4) suggest that the error decreases as the population size increases. We however have to remember that $E_P\left[SLL^R(\theta)\right] \neq LL(\theta)$, because of the logarithmic operator, and our confidence interval is thus centered at zero only asymptotically with R, leading to additional consistency issues when the population size I increases; we will briefly address them in the next section.

Since (5.2) implies that $LL(\theta) - SLL^R(\theta) \xrightarrow{p} 0$, when R tends to ∞ for a fixed population size I, the SAA the estimator is consistent, despite the bias introduced by the logarithm operator. The asymptotic bias can be shown to be of order $1/R$ for any unbiased estimator $\tilde{P}$ of P_{ij_i}, $i = 1, \ldots, I$ (Gouriéroux and Monfort [19]). In the Monte-Carlo situation, it is furthermore easy to numerically estimate the bias for a given finite R. We first compute the Taylor development of $\ln SP_{ij_i}^R$ around the true value P_{ij_i}, for some individual i:

$$\ln SP_{ij_i}^R(\theta) = \ln P_{ij_i}(\theta) + \frac{1}{P_{ij_i}(\theta)} h_{ij_i} - \frac{1}{2\left(P_{ij_i}(\theta)\right)^2} h_{ij_i}^2 + O\left(h_{ij_i}^3\right),$$

where $h_{ij_i} = SP_{ij_i}^R(\theta) - P_{ij_i}(\theta)$. Therefore, since $E_P\left[h_{ij_i}\right] = 0$,

$$E_P\left[\ln SP_{ij_i}^R(\theta)\right] - \ln P_{ij_i}(\theta) = -\frac{1}{2\left(P_{ij_i}(\theta)\right)^2} E_P\left[h_{ij_i}^2\right] + E_P\left[O\left(h_{ij_i}^3\right)\right].$$

From **A.0**, we then obtain that

$$E_P\left[h_{ij_i}^2\right] = \frac{1}{R}\sigma_{ij_i}^2(\theta).$$

Averaging now over the individuals, and neglecting the terms of order three and above, we obtain that the simulation bias B can be approximated by

$$B^R(\theta) := E_P[SLL^R(\theta)] - LL(\theta) = -\frac{1}{2IR}\sum_{i=1}^{I} \frac{\sigma_{ij_i}^2(\theta)}{\left(P_{ij_i}(\theta)\right)^2} \leq 0, \tag{5.5}$$

which can be easily computed from the estimated error as

$$B^R(\theta) = -\frac{I}{2\alpha_{\delta}^2}\left(\Delta_{\delta}^R(\theta)\right)^2. \tag{5.6}$$

Thus, (5.5) implies that, up to second order,

$$\max_{\theta} E_P[SLL^R(\theta)] \leq \max_{\theta} LL(\theta).$$

It is finally interesting to note from (5.3) that the confidence interval radius $\Delta_\delta^R(\theta)$ is small whenever the standard deviations are themselves small compared to the probability choices. Moreover (5.6) shows that the simulation bias decreases faster than the error. This suggests that the desired number of random draws should depend on the model nature: as expected, more variation of model parameters between the individuals imposes larger samples. The choice of a uniformly satisfying sample size across different models thus appears doubtful. This observation seems to support, for the case of the objective function value, the practical conclusions of Section 4.3 of Hensher and Greene [25].

Moreover, if we now make the additional assumption that the SAA problems are solved globally instead of locally, we obtain that, almost surely,

$$\max_{\theta \in S} E_P\left[SLL^R(\theta)\right] \leq E_P\left[\max_{\theta \in S} SLL^R(\theta)\right].$$

The maximization procedure itself can therefore produce another bias opposed to the simulation bias, a well-known phenomenon in stochastic programming. As a consequence, the optimal values of successive SAA problems do not necessarily increase monotonically with R, which makes bias tests based on this increase questionable.

5.3. Asymptotic behaviour for increasing population sizes

We finally devote a last paragraph to extending the results obtained by Gouriéroux and Monfort [19, 20] on the consistency and efficiency of the SAA problem when the population size becomes infinite. In particular, our results apply to the constrained case and the convergence results hold almost surely, instead of in distribution.

Due to the bias induced by the logarithm operator, the SAA estimator is inconsistent if R is fixed and I tends to infinity, while it is consistent when both R and I increase to infinity (Gouriéroux and Monfort [20], page 43). We however have to strengthen the assumptions in order to ensure that the SAA problem is asymptotically equivalent to the true problem. The following result is proved in Gouriéroux and Monfort [19] (in a more general setting); it can also be found in Hajivassiliou and McFadden [24], and is discussed in Train [41], page 288.

Proposition 5.1. *If I, $R \to \infty$ and $\sqrt{I}/R \to 0$, the SAA estimator is asymptotically equivalent to the true estimator.*

Note that this result is obtained using convergence in distribution of the solutions of the SAA problems. We provide, in the next theorem, a result of the same type, but now expressed almost surely, at the expense of not being directly computable.

Proposition 5.2. *Assume that a ULLN holds for the approximation $LL(\theta)$ of*

$$ELL(\theta) := E_{P_{\mathcal{I}}}\left[\ln P_{ij_i}(\theta)\right],$$

and another ULLN holds for the approximation $SLL^R(\theta)$ of $LL(\theta)$. Suppose furthermore that $SLL^R(\cdot, \gamma)$ is continuous on S for almost every sampling process γ, that $LL(\theta)$ is continuous on S for $P_{\mathcal{I}}$-almost every sampling process, and that $ELL(\theta)$ is continuous on S. Then

$$\sup_{\theta \in S} \left| SLL^R(\theta) - ELL(\theta) \right| \to 0$$

for $(P_{\mathcal{I}} \times P)$-almost every sampling process, if I tends to infinity and R tends to infinity sufficiently fast compared to I.

Proof. Let $\delta > 0$ be a small constant. From the ULLN assumption for $LL(\theta)$, we have that, for I sufficiently large,

$$\sup_{\theta} \left| E_{P_{\mathcal{I}}}[\ln P_{ij_i}(\theta)] - \frac{1}{I} \sum_{i=1}^{I} \ln P_{ij_i}(\theta) \right| < \frac{\delta}{2}$$

for $P_{\mathcal{I}}$-almost every sampling process. For such an I, we have, from the ULLN assumption for $SLL^R(\theta)$, that for R sufficiently high,

$$\sup_{\theta} \left| \frac{1}{I} \sum_{i=1}^{I} \ln P_{ij_i}(\theta) - \sum_{i=1}^{I} \ln \frac{1}{R} \sum_{r=1}^{R} P^R_{ij_i}(\theta) \right| < \frac{\delta}{2}$$

almost surely. Combining these two inequalities with the triangular inequality

$$\sup_{\theta} |SLL^R(\theta) - ELL(\theta)| \leq \sup_{\theta} |SLL^R(\theta) - LL(\theta)| + \sup_{\theta} |LL(\theta) - ELL(\theta)|,$$

we obtain that

$$\exists\, I_\delta \text{ s.t. } \forall I \geq I_\delta \;\exists\, R_I \text{ s.t. } \forall R \geq R_I, \sup_{\theta} |SLL^R(\theta) - ELL(\theta)| < \delta$$

for $(P_{\mathcal{I}} \times P)$-almost every sampling process. Now define some sequence $\{\delta_n\}_{n=1}^{\infty}$ converging to zero, and let $\{I_{\delta_n}\}$ be the corresponding population sizes as given by this last bound. If the population size I grows faster than I_{δ_n} and R faster than R_I, we see that

$$\sup_{\theta} \left| SLL^R(\theta) - LL(\theta) \right| \to 0 \tag{5.7}$$

for $(P_{\mathcal{I}} \times P)$-almost every sampling process, which implies the desired result in this case. If, on the other hand, I grows slower than I_{δ_n}, we identify an increasing subsequence of population sizes $\{I_n\} \subseteq \{I\}$ that grows faster than I_{δ_n}. For population sizes I' between I_n and I_{n+1}, (5.7) holds if we require $R_{I'}$ to be equal or larger than $R_{I_{n+1}}$. As a consequence, we obtain that (5.7) holds irrespective of the speed of growth of $\{I\}$ provided R grows sufficiently fast. □

Dropping again the explicit dependence on the sampling process, for $(P_{\mathcal{I}} \times P)$-almost every sampling process, if θ^* is a limit point of the sequence of SAA solutions $\{\theta^*_{I,R}\}$, for I tending to infinity, and R tending to infinity sufficiently fast compared to I, and if the subsequence $\{\theta^*_{I,\ell}\}$ of $\{\theta^*_{I,R}\}$ converges to θ^*, we have under the assumptions of the previous proposition that

$$SLL^{\ell}\left(\theta^*_{I,\ell}\right) \to ELL(\theta^*),$$

as $\ell \to \infty$. We may finally re-apply our convergence analysis to this framework, and obtain, under assumptions similar to those used above (we now need domination by functions that are $(P_{\mathcal{I}} \times P)$-integrable), that, $(P_{\mathcal{I}} \times P)$-almost surely, a sequence $\{\theta^*_{I,R}\}^{\infty,\ \infty}_{I=1,\ R=1}$ has a limit point θ^* that is first (second)-order critical if the $\theta^*_{I,R}$ are first (second)-order critical points.

6. Conclusion

We have first extended convergence properties known in stochastic programming in the case where minimizers of the approximating problems are global to the case where they are only local. This in turn allows for problems whose objective function is nonconvex.

In a second part, we have shown that the problem of estimating parameters in mixed logit models for discrete choices can be cast into this general stochastic programming framework. We then applied the new convergence properties to that case and strengthened existing results by proving almost sure convergence instead of convergence in distribution, both for constrained and unconstrained problems. The new theory also allows for general nonlinear utility functions. We finally derived computable estimates of the simulation bias and variance. These estimates provide information on the quality of the successive average approximation which can be used to improve efficiency of numerical estimation procedures, as done in AMLET (Bastin, Cirillo and Toint [3]).

Further research would be useful to alleviate assumptions needed for our consistency results, in particular when the feasible set S is nonconvex and/or stochastic, and to develop a more complete statistical inference theory for local minimization. Another point of interest is a better understanding of the bias and variance of the solutions of the successive average approximation solutions themselves (as opposed to values of the log-likelihood functions). A next step would also be to determine accurate bounds derived from quasi-Monte Carlo techniques instead of Monte Carlo samplings.

Acknowledgements. The authors would like to express their gratitude to Rüdiger Schultz who provided useful references and discussions for starting the work on Monte Carlo methods in stochastic programming as well as comments on the preprint version of this paper, and to Alexander Shapiro for suggesting use of stochastic variational inequalities and other important comments. Our thanks go also to Marcel Rémon and Thomas Bruss for their helpful comments on statistical theory and Stéphane Hess for his remarks on mixed logit theory, as well as to Serge Gratton and two anonymous referees for their relevant suggestions. We are also grateful to the Belgian National Fund for Scientific Research for the grant that made this research possible for the first author and for its support of the third author during his sabbatical mission.

References

1. Anderson, S.P., De Palma, A., Thisse, J.-F.: Discrete Choice Theory of Product Differentiation. MIT Press, Cambridge, Massachusetts, USA, 1992

2. Bastin, F., Cirillo, C., Hess, S.: Evaluation of optimisation methods for estimating mixed logit models. Transportation Research Record, vol. 1921, 35–43 (2005)
3. Bastin, F., Cirillo, C., Toint, P.L.: An adaptive monte carlo algorithm for computing mixed logit estimators. Computational Management Science, **3** (1), 55–79 (2006)
4. Ben-Akiva, M., Lerman, S.R.: Discrete Choice Analysis: Theory and Application to Travel Demand. The MIT Press, 1985
5. Bhat, C.R.: Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B, **35** (7), 677–693 (2001)
6. Bhat, C.R.: Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transportation Research Part B, **37** (3), 837–855 (2003)
7. Bhat, C.R., Castelar, S.: A unified mixed logit framework for modelling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco bay area. Transportation Research Part B, **36** (3), 593–616 (2002)
8. Bhat, C.R., Koppelman, F.S.: Activity-based modeling of travel demand. In: R.W. Hall, (ed.) Handbook of Transportation Science, Norwell, USA. Kluwer Academic Publisher, pp. 35–61, 1999
9. Birge, J.R., Louveaux, F.: Introduction to Stochastic Programming. Springer-Verlag, 1997
10. Borovkov, A.: Statistique mathmatique. Mir, 1987
11. Cirillo, C., Axhausen, K.W.: Mode choice of complex tour. In: Proceedings of the European Transport Conference (CD-ROM), Cambridge, UK, 2002
12. Davidson, J.: Stochastic Limit Theory. Oxford University Press, Oxford, England, 1994
13. Deák, I.: Multidimensional integration and stochastic programming. In: Y. Ermoliev, R.J.-B. Wets, (eds.), Numerical Techniques for Stochastic Optimization, Springer Verlag, pp. 187–200, 1988
14. Fiacco. A.V.: Introduction to Sensitivity and Stability Analysis in Nonlinear Programming. Academic, New York, USA, 1983
15. Fishman, G.S.: Monte Carlo: Concepts, Algorithms and Applications. Springer Verlag, New York, USA, 1996
16. Garrido. R.A.: Estimation performance of low discrepancy sequences in stated preferences. Paper presented at the 10th International Conference on Travel Behaviour Research, 2003
17. Gill, P.E., Murray, W., Saunders, M., Stewart G.W., Wright, M.H.: Properties of a representation of a basis for the null space. Mathematical Programming, **33** (2), 172–186 (1985)
18. Gould, N.I.M.: On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programming problem. Mathematical Programming, **32** (1), 90–99 (1985)
19. Gouriéroux, C., Monfort, A.: Simulation based econometrics in models with heterogeneity. Annales d'Economie et de Statistiques, **20** (1), 69–107 (1991)
20. Gouriéroux, C., Monfort, A.: Simulation-based Econometric Methods. Oxford University Press, Oxford, United Kingdom, 1996
21. Gugat, M.: A parametric view on the Mangasarian-Fromovitz constraint qualification. Mathematical Programming, **85** (3), 643–653, 1999
22. Gürkan, G., Özge, A.Y., Robinson, S.M.: Sample-path solution of stochastic variational inequalities. Mathematical Programming, **84** (2), 313–333 (1999)
23. Gürkan, G., Özge, A.Y., Robinson, S.M.: Solving stochastic optimization problems with stochastic constraints: an application in network design. In: D.T. Sturrock, P.A. Farrington, H.B. Nembhard, G.W. Evans, (eds.), Proceedings of the 1999 Winter Simulation Conference, USA, pp. 471–478, 1999
24. Hajivassiliou, V.A., McFadden. D.L.: The method of simulated scores for the estimation of LDV models. Econometrica, **66** (4), 863–896 (1998)
25. Hensher, D.A., Greene, W.H.: The mixed logit model: The state of practice. Transportation, **30** (2), 133–176 (2003)
26. Hess, S., Polak, J., Daly, A.: On the performance of shuffled Halton sequences in the estimation of discrete choice models. In: Proceedings of European Transport Conference (CD-ROM), Strasbourg, France, 2003. PTRC
27. Hess, S., Train, K., Polak, J.: On the use of a modified latin hypercube sampling (mlhs) approach in the estimation of a mixed logit model for vehicle choice. Transportation Research Part B, **40** (2), 147–163 (2006)
28. Kall, P., Wallace. S.W.: Stochastic Programming. John Wiley & Sons, 1994
29. Montmarquette, C., Cannings, K., Mahseredjian, S.: How do young people choose college majors? Economics of Education Review, **21** (6), 543–556 (2002)
30. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York, USA, 1999
31. Parthasarathy, K.R.: Probability Measures on Metric Spaces. Academic Press, 1967
32. Robinson, S.M.: Local epi-continuity and local optimization. Mathematical Programming, **37** (2), 208–222 (1987)
33. Robinson, S.M.: Analysis of sample-path optimization. Mathematics of Operations Research, **21** (3), 513–528 (1996)

34. Rubinstein, R.Y., Shapiro, A.: Discrete Event Systems. John Wiley & Sons, Chichester, England, 1993
35. Sándor, Z., Train, K.: Quasi-random simulation of discrete choice models. Transportation Research Part B, **38** (4), 313–327 (2004)
36. Shapiro, A.: Probabilistic constrained optimization: Methodology and applications. In: S. Uryasev, (ed.), Statistical inference of stochastic optimization problems. Kluwer Academic Publishers, pp. 282–304, 2000
37. Shapiro, A.: Stochastic programming by Monte Carlo simulation methods. SPEPS, 2000
38. Shapiro, A.: Monte Carlo sampling methods. In: A. Shapiro, A. Ruszczyński, (eds.), Stochastic Programming, Vol. **10** of Handbooks in Operations Research and Management Science. Elsevier, pp. 353–425, 2003
39. Sivakumar, A., Bhat, C.R., Ökten, G.: Simulation estimation of mixed discrete choice models using randomized quasi-monte carlo sequences: A comparison of alternative sequences, scrambling method, and uniform-to-normal variate transformation techniques. Transportation Research Record, vol. 1921, 112–122 (2005)
40. Train, K.: Halton sequences for mixed logit. Working paper No. E00-278, Department of Economics, University of California, Berkeley, 1999
41. Train, K.: Discrete Choice Methods with Simulation. Cambridge University Press, New York, USA, 2003