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

Bastin, Fabian; Cirillo, Cinzia; Toint, P.L.

Published in:
Mathematical Programming

DOI:
10.1007/s10107-006-0708-6
10.1007/s10107-006-0708-6

Publication date:
2006

Document Version
Peer reviewed version

Mathematical Programming manuscript No.
(will be inserted by the editor)

Fabian Bastin*[1] · Cinzia Cirillo[2] · Philippe L. Toint[2]

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

the date of receipt and acceptance should be inserted later

Abstract. Monte Carlo methods have been used extensively in the area of stochastic programming. As with other methods that involve a level of uncertainty, theoretical properties are required in order to give an indication of their performance. Traditional convergence properties of Monte Carlo methods in stochastic programming consider global minimisers or first-order critical points of the sample average approximation (SAA) problems as well as of the true problem. In this paper we review the first-order critical convergence and extend these results to the case where computed solutions are second-order critical, in turn allowing for problems whose objective function is nonconvex. As an application, we use the proposed framework in the estimation of mixed logit models for discrete choice, guaranteeing almost sure convergence of the solutions of the successive SAA problems. The result is observed to hold both for constrained and unconstrained problems. Finally, we produce estimates of the simulation bias and variance.

1. Introduction

Stochastic programming, that is mathematical programming where uncertainty is introduced in the problem by the use of random variables, is today recognised as an important area of operations research (see the books by Birge and Louveaux [10] and by Kall and Wallace [29], 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 realisations, or continuous. However, to our knowledge, the convergence theory for these methods has so far been limited to the case where the minimisation of the approximating subproblems is assumed to produce a global minimum in all the feasible set (see for instance Shapiro [39]), a first-order critical point (Gürkan, Özge and Robinson [21,22], Shapiro [40]) or a solution in a complete local minimising set with respect to some nonempty open bounded set (Robinson [36]). As a result, they have been applied mostly to linear or convex problems, where such assumptions are not restrictive.

It is our purpose to extend the theoretical understanding of this class of methods to the case where this global minimisation assumption no longer holds: the minimisation of the subproblem is allowed to converge to a local minimiser, irrespectively of the set of minimisers of the true problem. This investigation is worthwhile, in particular because it opens the possibility to consider stochastic problems with nonconvex objective

Department of Mathematics, University of Namur, Belgium

Transportation Research Group, Department of Mathematics, University of Namur, Belgium.

* Research Fellow of the Belgian National Fund for Scientific Research

Correspondence to: fbas@math.fundp.ac.be

functions. We introduce our approach by reviewing consistency results when only first-order critical points are considered. We next examine second-order criticality 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 practise, 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 [31], Bhat and Castelar [8], Brownstone, Bunch and Train [12], Cirillo and Axhausen [13], Hensher and Greene [24], Hensher and Sullivan [25], Hess and Polak [26]). Their advantages include the possibility to estimate taste variations, to take into account 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 implies 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 realisation and inserted in the logit formula. The integral giving the probability choice for this individual is then approximated by the mean of these results. Hajivassiliou and McFadden [23] show that the computed estimators are, under reasonable assumptions, asymptotically consistent and efficient. 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 [6] and Train [42] advocate Halton sequences for mixed logit models and find that they perform much better than random draws in simulation estimation. However Bhat [7] has pointed out that Halton sequences rapidly deteriorate in the coverage of the integration domain for high integration dimensions and has proposed using scrambled Halton sequences. He also randomised these sequences in order to allow the computation of the simulation variance of the model parameters. Hess, Polak and Daly [27] have shown that scrambled Halton methods can be very sensitive to the number of draws, and can behave poorly when this number increases. Recently Hess, Train and Polak [28] have proposed the use of modified latin hypercube sequences and have reported better results than with any of the Halton based approaches.

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 [15], Hess, Train and Polak [28]). 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 [43]), 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 of 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 organised 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 minimisation of the expectation of some function depending on a random variable (see Birge and Louveaux [10] or Kall and Wallace [29] 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, where S is a compact subset of $\mathbb{R}^m$ representing feasible solutions of the above problem, where $\boldsymbol{\xi}$ is a real random vector defined on the probability space $(\Xi, \mathcal{F}, P)$ and taking values in $\left(\mathbb{R}^k, \mathcal{B}^k\right)$, where $G : \mathbb{R}^m \times \Xi \to \mathbb{R}$ is a real valued function, and where $E_P[\cdot]$ 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 realisations, $g(z)$ 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 [39,40] 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

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

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 [2], McFadden and Train [30] and Allenby and Rossi [1]) or between transportation modes for travel (see Sheffi [41], Ortúzar and Willumsen [33]) 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 the relative attractiveness of the alternative.

Let us denote by $\mathcal{A}$ the set of alternatives and by I the population size. The set of alternatives available for individual i ($i = 1, \ldots, I$) is represented by $\mathcal{A}(i) \subset \mathcal{A}$. 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, 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. A popular and simple expression for V_{ij} ($j = 1, \ldots, |\mathcal{A}(i)|$) is the linear utility

$$V_{ij}(\beta_j, x_{ij}) = \beta_j^T x_{ij} = \sum_{k=1}^{K_j} \beta_j^k x_{ij}^k,$$

where K_j is the number of observed attributes for alternative j ($j = 1, \ldots, |\mathcal{A}(i)|$), but our analysis does not rely on this form. 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 maximises his/her utility. In other terms, he/she chooses A_j if and only if

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

Thus the probability of choosing alternative A_j is given by

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

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 to 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 parameters 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 given by

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

where η is a location parameter and $\mu > 0$ is a scale factor. Its mean is

$$\eta + \frac{\gamma_E}{\mu},$$

where $\gamma_E \approx 0.57721$ is the Euler constant. The popularity of the Gumbel 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

$$\frac{e^{V_{ij}}}{\sum_{m=1}^{|\mathcal{A}(i)|} e^{V_{im}}}.$$

This is the multinomial logit model. This model 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 more formal description of this property and associated difficulties can be found for instance in Ben-Akiva and Lerman [5], who show that its validity depends on the structure on the choice set, and also 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, including the mixed logit models (or error components models) (see Bhat and Koppelman [9] 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. More precisely they suppress 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 realisation of a random vector $\boldsymbol{\beta}$. Furthermore, $\boldsymbol{\beta}$ is itself assumed to be derived from a random vector $\boldsymbol{\gamma}$ and a parameters vector θ, which we express as

$$\boldsymbol{\beta} = h(\boldsymbol{\gamma}, \theta). \tag{2.3}$$

$\boldsymbol{\gamma}$ typically specifies the random nature of the model while θ quantifies the population characteristics for the model. Usually, $\boldsymbol{\beta}$ follows itself some probability distribution, and θ specifies the parameters of this distribution. We have therefore that $f(\boldsymbol{\beta}|\theta) = f(h(\boldsymbol{\gamma}, \theta))$, where f denotes the underlying distribution function. We will nevertheless use the notation (2.3) in order to emphasise that the random part can be expressed by a non-parametric vector, as in (2.1). For example, assume 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 $\boldsymbol{\beta}_k$, $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots, \mu_K, \sigma_K)$. Therefore, (2.3) can be written in this case as $\boldsymbol{\beta} = (\mu_1 + \sigma_1\boldsymbol{\gamma}_1, \mu_2 + \sigma_2\boldsymbol{\gamma}_2, \ldots, \mu_K + \sigma_K\boldsymbol{\gamma}_K)$.

If we knew the realisation $\gamma(i)$, 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 standard logit formula

$$L_{ij}\left(\gamma,\theta\right)=\frac{e^{V_{ij}(\beta(i),x_{ij})}}{\sum_{m=1}^{|\mathcal{A}(i)|}e^{V_{im}(\beta(i),x_{im})}}. \tag{2.4}$$

However, since $\boldsymbol{\beta}$ is random, we need to calculate the associated unconditional probability, which is obtained by integrating (2.4) over $\boldsymbol{\gamma}$:

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

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

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

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

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. Therefore, we use a Monte Carlo estimate of $P_{ij_i}(\theta)$ obtained by sampling over γ, and given by

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

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 of the simulated log-likelihood problem

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

However, since I can be large (typically in the thousands), the evaluation of $SLL^{R}(\theta)$ may remain very expensive, even on modern computers, as pointed out by Hensher and Greene [24].

We finally notice that the mixed logit problem (2.5)–(2.6) can be viewed as a generalised stochastic programming problem (2.1). Indeed, we may write (2.5)–(2.6) as

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

The associated sample average approximation problem is then written as

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

where $N=R\,I$. We will denote by θ^* a solution of (2.7) and by θ_R^* a solution of (2.8). The generalisation is minor since it only consists in optimising 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 approaching infinity. We introduce the basic concepts used in this paper by reviewing first-order convergence.

Let z_N^* be a first-order critical point of $\hat{g}_N(\cdot)$, as defined in (2.2). 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 $\overline{\xi} \stackrel{def}{=} \{\xi_k\}_{k=1}^{\infty}$. Since S is a compact set, the sequence of SAA solutions has some limit point $z^*\left(\overline{\xi}\right)$. By identifying this sequence with one its subsequences if necessary, we can therefore assume that $z_N^*\left(\overline{\xi}\right) \to z^*\left(\overline{\xi}\right)$ as $N \to \infty$. Our first aim is to show that, under reasonable assumptions, $z^*\left(\overline{\xi}\right)$ is a first-order critical point for the true problem (2.1).

Since $z^*(\overline{\xi})$ depends from the sequence of realisations $\overline{\xi}$, which is not known a priori, we have to introduce a suitable probability space on which we can define some random variable, whose realisations 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 [34], Chapter V, Theorem 5.1), we can construct the infinite-dimensional probability space

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

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 products measures $\prod_{i=1}^{j} P$, as expected. An element of (3.1) is therefore a process

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

formed by the successive draws ξ_k, $k = 1, \ldots, \infty$.

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

- $a.e.$ for almost everywhere;
- $\stackrel{a.s.}{\longrightarrow}$ for almost sure convergence;
- $\stackrel{p}{\longrightarrow}$ for convergence in probability;
- $\Rightarrow$ for convergence in distribution.

We refer the reader to Davidson [14] 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 otherwise, we will assume that the terms *almost everywhere* et *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 everywhere* et *almost surely* by the measure defined on this probability space. As above, we continue to use bold symbols to denote random variables, while a realisation of such a variable is represented in standard font.

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 almost surely. This and **A.2** are typical assumptions of stochastic programming theory (see for instance Rubinstein and Shapiro [37]). The stronger form of **A.1** is justified by our interest in first-order optimality conditions, which are expressed in terms of the objective function's gradient.

It is important to note (see [37] 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 $g(z)$, that is

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

They also imply that $g(z)$ is then continuous on S.

The ULLN property corresponds to the stochastic version of the uniform convergence of a sequence of functions. Therefore, we recall the following results.

Lemma 3.1. *Assume that **A.0**–**A.2** hold. Then*

$$\hat{g}_N(z_N^*) \xrightarrow{a.s.} g(z^*). \tag{3.2}$$

Furthermore, if $f(\cdot)$ is a continuous function defined on some convex domain that includes $\hat{g}_N(z_N^*)$ ($N = 1, \ldots, \infty$) and $g(z^*)$, then

$$f(\hat{g}_N(z_N^*)) \xrightarrow{a.s.} f(g(z^*)). \tag{3.3}$$

If $\hat{h}_N(\cdot)$ ($N = 1, \ldots, \infty$), and $h(\cdot)$ are functions such that

$$\hat{h}_N(z_N^*) \xrightarrow{a.s.} h(z^*),$$

then, for any real scalar α,

$$\alpha \hat{h}_N(z_N^*) + \hat{g}_N(z_N^*) \xrightarrow{a.s.} \alpha h(z^*) + g(z^*), \tag{3.4}$$

and

$$\hat{h}_N(z_N^*)\, \hat{g}_N(z_N^*) \xrightarrow{a.s.} h(z^*)\, g(z^*). \tag{3.5}$$

As first-order conditions are generally expressed in terms of the objective function's gradient, we need a further assumption on this gradient.

A.3 The gradient components $\frac{\partial}{\partial z_l} G(z, \xi)$ $(l = 1, \ldots, m)$, $z \in S$, are dominated by a P-integrable function.

This new assumption allows us to apply the results of Rubinstein and Shapiro [37], 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

$$\nabla_z g(y) = E_P \left[\nabla_z G \left(y, \boldsymbol{\xi} \right) \right].$$

This also implies that $\nabla \hat{g}_N(z^*)$ is an unbiased estimator of $\nabla g\left(z^*\right)$.

First-order convergence can be derived from stochastic variational inequalities, as presented in Shapiro [40]. Consider a mapping $\Phi : \mathbb{R}^n \times \Xi_\Pi \to \mathbb{R}^n$ and a multifunction $\Gamma : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. Suppose that the expectation $\phi(z) := E_{P_\Pi}[\Phi(z, \boldsymbol{\xi})]$ is well defined. We refer now to

$$\phi(z) \in \Gamma(z) \tag{3.6}$$

as the true, or expected value, generalised equation and say that a point $z^* \in \mathbb{R}^n$ is a solution of (3.6) if $\phi(z^*) \in \Gamma(z^*)$. If $\{\xi_1, \ldots, \xi_N\}$ is a random sample, we refer to

$$\hat{\phi}_N(z) \in \Gamma(z) \tag{3.7}$$

as the SAA generalised 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 (3.6) and SAA (3.7) generalised equations, respectively.

Let denote by

$$d(x, A) \overset{def}{=} \inf_{x' \in A} \|x - x'\|,$$

the distance from $x \in \mathbb{R}^n$ to A, and

$$d(A, B) \overset{def}{=} \sup_{x \in A} d(x, B),$$

the deviation of the set A from the set B. We then have the following result (Shapiro [40]):

Theorem 3.1. *Let $\mathcal{S}$ be a compact subset of $\mathbb{R}^n$ such that $S^* \subset \mathcal{S}$. 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, $\emptyset \neq S_N^* \subset \mathcal{S}$ 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, the first-order critical condition for some point z^* is equivalent to require 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, irrespective of the number of draws. Theorem 3.1 then allows an easy proof of almost sure first-order convergence. Consider the choice $\Gamma(\cdot) = \mathcal{N}_S(\cdot)$. Then $\phi(z^*) \in \Gamma(z^*)$ if and only if

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

Following Shapiro [40], 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 (3.6) and SAA (3.7) generalised 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 S compact 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 stated as follows

$$\begin{aligned} \min_{z \in V} \ & g(z) = E_P[G(z, \boldsymbol{\xi})], \\ \text{subject to } & c_j(z) \geq 0, \ j = 1, \ldots, k, \\ & c_j(z) = 0, \ j = k+1, \ldots, M, \end{aligned} \tag{3.8}$$

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

$$\begin{aligned} \min_{z \in V} \ & \hat{g}_N(z), \\ \text{subject to } & \hat{c}_{jN}(z) \geq 0, \ j = 1, \ldots, k, \\ & \hat{c}_{jN}(z) = 0, \ j = k+1, \ldots, M. \end{aligned} \tag{3.9}$$

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:

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

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

$$\begin{aligned} &\min_{z\in V} \hat{g}(z,\boldsymbol{\epsilon}), \\ &\text{subject to } \hat{c}_j(z,\boldsymbol{\epsilon}) \geq 0,\ j=1,\ldots,k, \\ &\qquad\qquad \hat{c}_j(z,\boldsymbol{\epsilon}) = 0,\ j=k+1,\ldots,M, \end{aligned} \tag{3.10}$$

where $\boldsymbol{\epsilon}$ is a random vector of parameters giving perturbations of the program (3.10); $g(\cdot)$, $\hat{g}(\cdot,\boldsymbol{\epsilon})$, $c_j(\cdot)$, $\hat{c}_j(\cdot,\boldsymbol{\epsilon})$ are assumed to be twice continuously differentiable with respect to z. We will assume that the perturbation is of the form

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

where each component is a function from $\mathbb{R}^m \times \Xi_{\Pi}$ to $\mathbb{R}$, and

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

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

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

and we will denote the corresponding random vector by $\boldsymbol{\epsilon}_N\left(z,\overline{\boldsymbol{\xi}}\right)$. We will assume that $\boldsymbol{\epsilon}_N\left(z,\overline{\boldsymbol{\xi}}\right)$ converges uniformly on V to $0 = 0(z)$ 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 approximating problems are nonempty. The Lagrangian functions associated to (3.8) and (3.10) are respectively

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

Let $z^*(\boldsymbol{\epsilon})$ denote a first-order critical point for program (3.10): there therefore exist Lagrange multipliers $\lambda^*(\boldsymbol{\epsilon})$ such that $(z^*(\boldsymbol{\epsilon}),\lambda^*(\boldsymbol{\epsilon}))$ satisfy the Karush-Kuhn-Tucker (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 $\lambda^*(\epsilon_N)$. As before, since V is compact, $z_N^*(\overline{\xi})$ has some limit point $z^*\left(\overline{\xi}\right)$ as $N \to \infty$. Without loss of generality, we can assume that $z_N^*(\overline{\xi})$ converges to $z^*\left(\overline{\xi}\right)$ as $N \to \infty$, by considering a subsequence if necessary. We can now prove almost-sure first-order convergence for the general case.

Theorem 3.2. *Assume that*

(a) $\boldsymbol{\epsilon}_N \xrightarrow{a.s.} \mathbf{0}$ *uniformly on V, as* $N \to \infty$,
(b) for almost every $\overline{\xi}$ *in* $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, $\lambda_N^*(\overline{\xi})$ *has some limit point* $\lambda^*\left(\overline{\xi}\right)$ *as* $N \to \infty$,

Then, for almost every $\overline{\xi}$, $z^*\left(\overline{\xi}\right)$ *is a first-order critical point for* (3.8)*.*

Proof. From (b), for almost every $\overline{\xi}$, the sequence $\{(z_N^*(\overline{\xi}), \lambda_N^*(\overline{\xi})\}$, $N = 1, \ldots, \infty$, has some limit point $\left(z^*\left(\overline{\xi}\right), \lambda^*\left(\overline{\xi}\right)\right)$. (a) and Lemma 3.1 imply that $\left(z^*\left(\overline{\xi}\right), \lambda^*\left(\overline{\xi}\right)\right)$ satisfies the KKT conditions for the true problem. □

Note that assumption (b) always holds if the multipliers remain bounded. If this stronger assumption is made, it is also possible to use Theorem 3.1 to show that z^* is first-order critical, as in Shapiro [40]. 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 optimisation problem, and Theorem 3.1 then implies almost sure first-order criticality, with $\Gamma(\mu) := \mathcal{N}_\mathcal{K}\mu$. Assumptions (a) and (c) are satisfied since $\boldsymbol{\epsilon} \to 0$ almost surely, and. (c) is ensured by the assumption that the feasible sets for the original and approximating problems are nonempty.

4. Second-order convergence

4.1. Deterministic constraints

If we are ready to further strengthen our assumptions, we now show that, almost surely, there exists a limit point $z^*\left(\overline{\boldsymbol{\xi}}\right)$ which is a local minimizer. 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 minimiser of $\hat{g}_N(z)$. This is to say that

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

where $B(x, d)$ is the open ball centred at x and of radius d. As before, we also assume that $z_N^*(\overline{\xi})$ converges to some $z^*\left(\overline{\xi}\right)$ as N tends to infinity. In order to show that $z_N^*\left(\overline{\xi}\right)$ is a local minimiser of $g(\cdot)$, we must therefore have that the neighbourhood in which

$z_N^*(\overline{\xi})$ is a local minimiser does not shrink to a singleton when $N \to \infty$. We express this requirement by the following technical assumption.

A.4 For almost every sampling process $\overline{\xi}$, there exists a $\delta(\overline{\xi}) > 0$ and an $N(\overline{\xi}) > 0$ such that, for all $N \geq N(\overline{\xi})$,

$$\forall z \in B(z_N^*(\overline{\xi}), \delta(\overline{\xi})) \cap S, \quad \hat{g}_N(z_N^*(\overline{\xi})) \leq \hat{g}_N(z). \tag{4.2}$$

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

Theorem 4.1. *Assume that **A.0–A.4** hold. Then for almost every sampling process $\overline{\xi}$, $z^*\left(\overline{\xi}\right)$ is a local minimum of $g(\cdot)$.*

Proof. Consider a particular realisation $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$ such that

$$\sup_{z \in S} |\hat{g}_N(z) - g(z)| \to 0, \tag{4.3}$$

$$\hat{g}_N(z_N^*) \longrightarrow g(z^*) \tag{4.4}$$

and such that the $\delta(\overline{\xi})$ given by **A.4** exists. From the ULLN property, Lemma 3.1 and **A.4**, almost every $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$ satisfies these requirements. For simplicity of notation, we will write z^* instead of $z^*\left(\overline{\xi}\right)$, and δ instead of $\delta(\overline{\xi})$. Let z' be a minimiser of g in $\mathcal{K} \stackrel{def}{=} B\left(z^*, \frac{\delta}{2}\right) \cap S$. Then we first show that, for N sufficiently large,

$$z_N^* \in \mathcal{K} \subseteq B(z_N^*, \delta). \tag{4.5}$$

Since z^* is the limit point of $\{z_N^*\}_{N=0}^{\infty}$, the first inclusion of (4.5) must hold for N sufficiently large. Consider now $z \in \mathcal{K}$ We have that $|z - z_N^*| \leq |z - z^*| + |z^* - z_N^*|$, and thus that

$$|z - z_N^*| < \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Therefore $z \in B(z_N^*, \delta)$, completing our proof of (4.5) holds for N is sufficiently large. We now verify that

$$|\hat{g}_N(z_N^*) - g(z')| \longrightarrow 0. \tag{4.6}$$

Assume first that $\hat{g}_N(z_N^*) \leq g(z')$. Since z' minimises $g(\cdot)$ in $\mathcal{K}$ and, from (4.5), $z_N^* \in \mathcal{K}$, we have that

$$0 \leq |\hat{g}_N(z_N^*) - g(z')| = g(z') - \hat{g}_N(z_N^*) \leq g(z_N^*) - \hat{g}_N(z_N^*) \leq \sup_{z \in S} |\hat{g}_N(z) - g(z)|.$$

The limit (4.3) then implies (4.6). Assume now that $\hat{g}_N(z_N^*) \geq g(z')$. Since $z' \in \mathcal{K}$, we then deduce, from the second part of (4.5) and Assumption **A.4**, that $\hat{g}_N(z_N^*) \leq \hat{g}_N(z')$. Therefore

$$0 \leq \hat{g}_N(z_N^*) - g(z') \leq \hat{g}_N(z') - g(z') \leq \sup_{z \in S} |\hat{g}_N(z) - g(z)|,$$

and we again deduce (4.6) in this case. Taking now (4.4) into account, we deduce 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 reasonning is valid for almost every sampling process $\overline{\xi}$, our proof is complete. □

Classical results, where global minimisers are considered, express that $d(z_N^*, S^*)$ converges almost surely to zero as N tends to infinity, where S^* is the set of minimisers of the true problem (see for instance Theorem 3.1). Robinson [36] shows that, under mild regularity conditions, if the true problem has a complete local minimising (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 to the true problem and the one corresponding to the approximating problem tends to 0 as N tends to infinity. Moreover, the approximating infimum over the closure of $\mathcal{G}$ 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) minimiser of the approximating problem to the set of true local minimisers converges almost surely to zero. To see this, consider the problem

$$\min_{z\in[-1,1]} z^3 - \frac{z}{2}E_P[\boldsymbol{\xi}], \tag{4.7}$$

where $\Xi = \{-1, 1\}$ and $P[\xi = -1] = P[\xi = 1] = 0.5$, so $E_P[\boldsymbol{\xi}] = 0$. Therefore (4.7) has only one local minimiser, which is also global, at $z^* = -1$. The SAA problem is then

$$\min_{z\in[-1,1]} z^3 - \frac{z}{2N}\sum_{i=1}^{N}\xi_i. \tag{4.8}$$

(4.8) has two (isolated) local minimisers,

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

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

$$\sqrt{\frac{\sum_{i=1}^N \xi_i}{6N}} \xrightarrow{a.s.} 0$$

since, from the strong law of large numbers, $\frac{1}{N}\sum_{i=1}^N \xi_i \to E_P[\boldsymbol{\xi}] = 0$ almost surely as $N \to \infty$. But zero is a saddle point of the true problem (4.7), not a minimiser, 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 that in a neighbourhood of a local solution of the true solution, under some mild regularity the SAA has almost surely a solution when N is sufficiently large (Shapiro [40]). However, the previous example illustrates that care must be exercised when solving the SAA problem for N fixed since we can find approximating local minimisers that are not close to true local minimisers.

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 [37], it is important to note that perturbation analysis assumes the existence of a solution of the true problem and then studies the existence and behaviour of solutions of 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 a solution of the true problem. The difference will be more formally illustrated at the end of the section where we compare our developments to some sensitivity analysis results.

We consider the case where the feasible set is described by a set of equality and inequality constraints, as in (3.8). As before, we assume that $\boldsymbol{\epsilon}_N$ converges uniformly on V to zero almost surely. Consider a particular sampling process $\overline{\xi}$ in $(\Xi_{\Pi}, \mathcal{F}_{\Pi}, P_{\Pi})$; without loss of generality, we can assume that $z_N^*(\overline{\xi})$ converges to some $z^*\left(\overline{\xi}\right)$ as $N \to \infty$. Under some conditions, $\lambda_N^*(\overline{\xi})$ also converges to some Lagrange multipliers vector $\lambda^*\left(\overline{\xi}\right)$ associated to $z^*\left(\overline{\xi}\right)$ for the true problem, as expressed in the lemma below.

Lemma 4.1. *Consider a particular sampling process $\overline{\xi}$ in $(\Xi_{\Pi}, \mathcal{F}_{\Pi}, P_{\Pi})$ such that $\epsilon_N\left(z, \overline{\xi}\right) \to 0$ uniformly on V as $N \to \infty$ and that $z_N^*(\overline{\xi})$ converges to some $z^*\left(\overline{\xi}\right)$. Assume moreover that there is an unique Lagrange multipliers vector $\lambda^*\left(\overline{\xi}\right)$ associated to $z^*\left(\overline{\xi}\right)$ that satisfies the KKT conditions. Then $\lambda_N^*(\overline{\xi})$ converges to $\lambda^*\left(\overline{\xi}\right)$ as N tends to infinity.*

Proof. From the uniqueness of $\lambda^*\left(\overline{\xi}\right)$, the Mangasarian-Fromowitz constraint qualification (MFCQ) holds at $z^*(\overline{\xi})$, and therefore in a neighbourhood of $z^*\left(\overline{\xi}\right)$ (while the converse is not necessarily true, as shown by Gugat [20]). 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 $\{\lambda_N^*\}$, $N = 1, \ldots, \infty$, is equal to $\lambda^*\left(\overline{\xi}\right)$. Let λ' be such a limit point. By continuity, $(z^*\left(\overline{\xi}\right), \lambda')$ satisfies the KKT conditions, so λ' is equal to $\lambda^*\left(\overline{\xi}\right)$. □

The uniqueness of $\lambda^*\left(\overline{\xi}\right)$ 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 (3.8). 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 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 [32]. 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 λ_j^* and $c_j(z^*)$ is zero for each index $j = 1, \ldots, k$. In other words, we have that $\lambda_j^* > 0$ for each $j \in \{1, \ldots, k\} \cap \mathcal{A}(z^*)$.*

If the assumptions of Lemma 4.1 hold for almost every sampling process $\overline{\xi}$ in $(\Xi_{\Pi}, \mathcal{F}_{\Pi}, P_{\Pi})$, the gradient $\nabla L\left(z_N^*\left(\overline{\boldsymbol{\xi}}\right), \lambda_N^*\left(\overline{\boldsymbol{\xi}}\right)\right)$ converges almost surely to some $\nabla \mathcal{L}\left(z^*\left(\overline{\boldsymbol{\xi}}\right), \lambda^*\left(\overline{\boldsymbol{\xi}}\right)\right)$, when N tends to infinity. Consider again a particular sampling process $\overline{\xi}$ in $(\Xi_{\Pi}, \mathcal{F}_{\Pi}, P_{\Pi})$ such that $\epsilon_N\left(z, \overline{\xi}\right) \to 0$ as $N \to \infty$. Assume that the strict complementarity condition and the LICQ hold at $\left(z^*\left(\overline{\xi}\right), \lambda^*\left(\overline{\xi}\right)\right)$ for problem (3.10) at $\overline{\xi}$. If $\lambda_N^*(\overline{\xi}) \to \lambda^*\left(\overline{\xi}\right)$ we obtain then that $\lambda_N^*(\overline{\xi})$, $j \in \{1, \ldots, k\} \cap \mathcal{A}\left(z^*\left(\overline{\xi}\right)\right)$, are strictly positive and hence the corresponding constraints are active at $z_N^*(\overline{\xi})$ for N sufficiently large. Moreover, assuming $\hat{c}_j(z_N^*(\overline{\xi}), \epsilon_N(z_N^*(\overline{\xi}), \overline{\xi})) \to c(z^*(\overline{\xi}))$ (which is true for almost every $\overline{\xi}$), we have that for N large enough, $\mathcal{A}(z_N^*(\overline{\xi})) = \mathcal{A}(z^*(\overline{\xi}))$ and the strict complementarity condition holds at $z_N^*(\overline{\xi})$, with respect to the Lagrangian multipliers $\lambda_N^*(\overline{\xi})$, for problem (3.9). This allows us to state the theorem below.

Theorem 4.2 (Second-order convergence). *Assume that, for almost every sampling process $\overline{\xi}$ in $(\Xi_{\Pi}, \mathcal{F}_{\Pi}, P_{\Pi})$, $\lambda^*\left(\overline{\xi}\right)$ is the unique vector of Lagrangian multipliers associated to program* (3.8) *at $z^*\left(\overline{\xi}\right)$, and that*

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

Suppose also that, almost surely, the strict complementarity condition and the LICQ hold at $\left(z^*\left(\overline{\xi}\right), \lambda^*\left(\overline{\xi}\right)\right)$ for (3.8)*. Then, for almost every sampling process $\overline{\xi}$,*

(i) the LICQ holds at $\left(z_N^*\left(\overline{\xi}\right), \lambda_N^*\left(\overline{\xi}\right)\right)$,
(ii) $\left(z^*\left(\overline{\xi}\right), \lambda^*\left(\overline{\xi}\right)\right)$ satisfies the second-order necessary condition for (3.8)*:*

$$w^T \nabla_{zz}^2 \mathcal{L}\left(z^*\left(\overline{\xi}\right), \lambda^*\left(\overline{\xi}\right)\right) w \geq 0, \text{ for all } w \in \text{Null}[\nabla_z c_j\left(z^*\left(\overline{\xi}\right)\right)^T]_{j \in \mathcal{A}\left(z^*\left(\overline{\xi}\right)\right)}. \tag{4.9}$$

If furthermore there exists almost surely some $\alpha(\overline{\xi}) > 0$ such that, for all N large enough,

$$w^T \nabla_{zz}^2 L_N(z_N^*(\overline{\xi}), \lambda_N^*(\overline{\xi})) w > \alpha(\overline{\xi}), \text{ for all } w \in \text{Null}[\nabla_z \hat{c}_j(z_N^*(\overline{\xi}))^T]_{j \in \mathcal{A}(z_N^*(\overline{\xi}))}, \ \|w\| = 1, \tag{4.10}$$

then $(z^*\left(\overline{\xi}\right), \lambda^*\left(\overline{\xi}\right))$ almost surely satisfies the second-order sufficient conditions for problem (3.8)*, that is*

$$\text{(iii) } w^T \nabla_{zz}^2 \mathcal{L}\left(z^*\left(\overline{\xi}\right), \lambda^*\left(\overline{\xi}\right)\right) w > 0, \text{ for all } w \in \text{Null}[\nabla_z c_j\left(z^*\left(\overline{\xi}\right)\right)^T]_{j \in \mathcal{A}\left(z^*\left(\overline{\xi}\right)\right)}, \ \|w\| = 1. \tag{4.11}$$

In other terms, $z^*\left(\overline{\xi}\right)$ is an isolated local minimiser of (3.8)*, for almost every sampling process $\overline{\xi}$.*

Proof. Consider a sampling process $\overline{\xi}$ such that the assumptions of the theorem are satisfied. For simplicity, we drop the dependence on $\overline{\xi}$ in our notation. In order to show (i), consider

$$\{\nabla_z c_j(z^*)\}_{j\in\mathcal{A}(z^*)}, \tag{4.12}$$

the matrix formed by the gradients of active constraints at z^* for (3.8). From the strict complementarity conditions and convergence of Lagrange multipliers, the active set of program (3.9) at z_N^* is asymptotically the same as the active set of program (3.8) at z^*. Since $\epsilon_N \to 0$ uniformly on V, we have that the matrix formed by the active constraints of the perturbed problem,

$$\{\nabla_z \hat{c}_j(z_N^*, \epsilon_N)\}_{j\in\mathcal{A}(z_N^*)} \tag{4.13}$$

converges to (4.12) as N tends to infinity:

$$\{\nabla_z \hat{c}_j(z_N^*, \epsilon_N)\}_{j\in\mathcal{A}(z_N^*)} \longrightarrow \{\nabla_z c_j(z^*)\}_{j\in\mathcal{A}(z^*)}. \tag{4.14}$$

The LICQ amounts to say that at least one square submatrix of (4.12) is nonsingular. From (4.14), the same is true for (4.13) for N large enough. Thus the LICQ holds for the approximating problems when N is sufficiently large. Conclusion (i) then follows from the fact that our assumptions on the sampling process $\overline{\xi}$ hold almost surely.

We show now (ii) and again consider a particular sampling process $\overline{\xi}$ satisfying our assumptions. From (4.14) we may associate a basis K_N with the null space of (4.13) such that

$$K_N \longrightarrow K, \tag{4.15}$$

where K is a basis of Null$[\nabla_z c_j(z)^T]_{j\in\mathcal{A}(z^*)}$ (see Gill et al. [18]). Using the strict complementarity condition and LICQ, the fact that (z_N^*, λ_N^*) satisfies the second-order necessary conditions can now be expressed as

$$K_N^T \nabla_{zz}^2 L(z_N^*, \lambda_N^*, \epsilon_N) K_N \text{ is positive semi-definite.}$$

From (4.15) and Assumptions (a)–(d), we have that

$$K_N^T \nabla_{zz}^2 L(z_N^*, \lambda_N^*, \epsilon_N) K_N \longrightarrow K^T \nabla_{zz}^2 \mathcal{L}(z^*, \lambda^*) K.$$

Therefore we have (4.9) and (ii) follows from the fact that our assumptions on the sampling process hold almost surely. The reasonning is identical for proving (iii), except that one now uses the lower bound α on the eigenvalues of $K_N^T \nabla_{zz}^2 L(z_N^*, \lambda_N^*, \epsilon_N) K_N$ to obtain (4.11). □

Theorem 4.2 expresses that, under some smoothness conditions, a limit point of a sequence of SAA second-order critical solutions is almost surely a solution of the true problem if some qualification constraints hold at this point.

Note that the LICQ and strict complementarity conditions imply that the minimiser is isolated while the second-order sufficient condition is usually used to characterise strict local minimisers. In other terms, there exists a neighbourhood $\mathcal{V}_S$ of $z^*\left(\overline{\xi}\right)$ such that $z^*\left(\overline{\xi}\right)$ is the only local minimiser in $\mathcal{V}_S$. Recall that isolated local minimisers are also strict local minimisers but that the inverse is not always true (Nocedal and Wright [32], page 14). If $z^*\left(\overline{\xi}\right)$ is a strict but not isolated local minimiser every neighbourhood of $z^*\left(\overline{\xi}\right)$ contains other local minimisers than $z^*\left(\overline{\xi}\right)$, that are candidates to

be limit points of the sequences of solutions of the SAA problems (2.2), as N tends to infinity, so $z^*\left(\overline{\xi}\right)$ can be difficult to identify.

The non-degeneracy assumption (4.10) can also be replaced by requiring that the Jacobian of the equality equations involved in the KKT conditions associated to the program (3.8),

$$\begin{aligned} \nabla_z \mathcal{L}(z,\lambda) &= 0, \\ \lambda_j c_j(z) &= 0, j = 1,\ldots,M, \\ c_j(z) &= 0, j = k+1,\ldots,M \end{aligned} \tag{4.16}$$

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

Corollary 4.1. *Assume that, for almost every $\overline{\xi}$ in $(\Xi_{\Pi},\mathcal{F}_{\Pi},P_{\Pi})$, $\lambda^*\left(\overline{\xi}\right)$ is the unique vector of Lagrangian multipliers associated to program* (3.8) *at $z^*\left(\overline{\xi}\right)$, that assumptions (a)-(d) of Theorem 4.2 hold and that the strict complementarity condition holds at $\left(z^*\left(\overline{\xi}\right),\lambda^*\left(\overline{\xi}\right)\right)$ for* (3.8)*. Assume furthermore that the Jacobian of* (4.16) *is almost surely nonsingular at $\left(z^*\left(\overline{\boldsymbol{\xi}}\right),\lambda^*\left(\overline{\boldsymbol{\xi}}\right)\right)$. Then $(z^*(\overline{\boldsymbol{\xi}}),\lambda^*(\overline{\boldsymbol{\xi}}))$, almost surely satisfies (4.11), the second-order sufficient conditions for program* (3.8)*.*

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

$$\begin{aligned} \nabla_z \mathcal{L}(z^*) &= 0, \\ c_j(z^*) &= 0, \quad j \in \mathcal{A}(z^*), \\ \lambda_j^* &= 0, \quad j \notin \mathcal{A}(z^*), \end{aligned} \tag{4.17}$$

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

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

who is nonsingular if and only if

$$\begin{pmatrix} \nabla_{zz}\mathcal{L}(z^*) & \nabla_z c_1(z^*) & \cdots & \nabla_z c_{n_a}(z^*) \\ \nabla_z^T c_1(s^*) & 0 & \ldots & 0 \\ \vdots & & & \\ \nabla_z^T c_{n_a}(z^*) & 0 & \ldots & 0 \end{pmatrix}, \tag{4.18}$$

is itself nonsingular. From the Sylvester's law of inertia, (4.18) 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 [19]). Note that (4.18) also implies that the LICQ holds at (z^*, λ^*). From Theorem 4.2, (z^*, λ^*) also satisfies the second-order necessary conditions for program (3.8), that is (4.11). □

The converse of Theorem 4.2 can be obtained from classical results of perturbation analysis (Fiacco [16], Theorem 3.2.2), which we restate for completeness. More developments in the context of stochastic programming can be found in Rubinstein and Shapiro [37] and Shapiro [38].

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

(a) the functions defining (3.10) *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 $(z^*, 0)$,*
(b) the second-order sufficient conditions for a local minimum of (3.10) *hold at z^*, with associated Lagrange multipliers λ^*,*
(c) the LICQ holds at $(z^*, 0)$,
(d) the strict complementarity condition holds at $(z^*, 0)$,

then

(i) z^* is a local isolated minimum of (3.10) *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 (3.10) *such that $\gamma(0) = (z^*, \lambda^*)^T$, and hence $z(\epsilon)$ is a local isolated minimiser of problem* (3.10) *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 $z^*(\epsilon)$.

Of course, this theorem must be applied for a fixed sampling process $\overline{\xi}$, and the results of interest here are only true almost surely. Note that the second-order sufficiency property is now taken as an assumption, so that z^* is in fact 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 minimising set (Robinson [35, 36]).

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 generalisation 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 we assume that the set S is compact 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 on the latter by introducing it.

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 be explicitly expressed as convex constraints (bounds are typical) on the problem or be implicit for an unconstrained problem. In the latter case, **A.5** indicates that the solutions are uniformly bounded for sufficiently large sampling sizes. Such an assumption is reasonable to avoid pathological cases where some components of θ_R^* converge towards infinity. As for 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, and, again as above, we identify it, without loss of generality, to one of its convergent subsequences and assume that $\theta_R^*(\overline{\gamma}) \to \theta^*(\overline{\gamma})$ as $R \to \infty$.

In order to obtain convergence to first-order critical points, we also need to translate Assumptions **A.1**–**A.3**. We first ensure **A.1** and **A.2** by imposing suitable conditions on the 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

$$\frac{\partial}{\partial \theta_t} L_{ij_i}(\gamma, \theta) = L_{ij_i}(\gamma, \theta) \sum_{s \neq j_i} L_{is}(\gamma, \theta) \frac{\partial}{\partial \theta_t} \left(V_{ij_i}(\gamma, \theta, x_{ij_i}) - V_{is}(\gamma, \theta, x_{is}) \right). \quad (5.1)$$

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**, **A.1ml** and Lemma 3.1 that, for all individuals i $(i = 1, \ldots, I)$,

$$SP_{ij_i}^R(\theta_R^*(\overline{\boldsymbol{\gamma}})) \xrightarrow{a.s.} P_{ij_i}(\theta^*(\overline{\boldsymbol{\gamma}})) \text{ and } SLL^R(\theta_R^*(\overline{\boldsymbol{\gamma}})) \xrightarrow{a.s.} LL(\theta^*(\overline{\boldsymbol{\gamma}})).$$

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

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

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 $t=1,\ldots,m$, $\frac{\partial}{\partial\theta_t}V_{ij}\,(\gamma,\theta,x_{ij})$ $(i=1,\ldots,I,\,j=1,\ldots,J)$ 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. Then all we 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 $K(\gamma)=\max_\theta\left|\frac{\partial}{\partial\theta_t}V_{ij}\,(\gamma,\theta,x_{ij})\right|$. Under **A.1ml**, and the assumption that $\theta\in S$, where S is compact, $K(\gamma)$ is finite for almost every γ, and its expectation is usually finite.

We may now apply Lemma 3.1 and deduce that

$$\nabla_\theta SLL^R(\theta_R^*(\overline{\boldsymbol{\gamma}}))\xrightarrow{a.s.}\nabla_\theta LL(\theta^*(\overline{\boldsymbol{\gamma}})),$$

as $R\to\infty$. 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 $\overline{\gamma}=\{\gamma_{i,r}\}$, $\theta^*(\overline{\gamma})$ 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 [43]) 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

$$\begin{aligned}\frac{\partial}{\partial\theta_u\partial\theta_t}LL(\theta)=&\frac{1}{I}\sum_{i=1}^{I}\frac{P_{ij_i}(\theta)\frac{\partial}{\partial\theta_u}\frac{\partial}{\partial\theta_t}E_P\left[L_{ij_i}(\boldsymbol{\gamma},\theta)\right]}{\left(P_{ij_i}(\theta)\right)^2}\\&-\frac{1}{I}\sum_{i=1}^{I}\frac{\frac{\partial}{\partial\theta_u}E_P\left[L_{ij_i}(\boldsymbol{\gamma},\theta)\right]\frac{\partial}{\partial\theta_u}E_P\left[L_{ij_i}(\boldsymbol{\gamma},\theta)\right]}{\left(P_{ij_i}(\theta)\right)^2},\end{aligned}$$

and

$$\frac{\partial}{\partial\theta_u\partial\theta_t}SLL(\theta) = \frac{1}{I}\sum_{i=1}^{I}\frac{SP_{ij_i}^R(\theta)\frac{1}{R}\sum_{r=1}^{R}\frac{\partial}{\partial\theta_u}\frac{\partial}{\partial\theta_t}L_{ij_i}(\gamma_{i,r},\theta)}{SP_{ij_i}^R(\theta)^2} - \frac{1}{I}\sum_{i=1}^{I}\frac{\left(\frac{1}{R}\sum_{r=1}^{R}\frac{\partial}{\partial\theta_u}L_{ij_i}(\gamma_{i,r},\theta)\right)\left(\frac{1}{R}\sum_{r=1}^{R}\frac{\partial}{\partial\theta_t}L_{ij_i}(\gamma_{i,r},\theta)\right)}{SP_{ij_i}^R(\theta)^2}.$$

We therefore have to request that

$$\frac{1}{R}\sum_{r=1}^{R}\frac{\partial}{\partial\theta_u}\frac{\partial}{\partial\theta_t}L_{ij_i}(\gamma_{i,r},\theta_R^*(\overline{\boldsymbol{\gamma}})) \xrightarrow{a.s.} \frac{\partial}{\partial\theta_u}\frac{\partial}{\partial\theta_t}E_P\left[L_{ij_i}(\boldsymbol{\gamma},\theta^*(\overline{\boldsymbol{\gamma}}))\right],$$

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. 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 centred at zero and of radius Δ. In practise, 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 $\overline{\boldsymbol{\gamma}}$),

$$P\left[\left|LL(\theta) - SLL^R(\theta)\right| \leq \Delta\right] \geq \alpha.$$

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

$$\ln P_{ij_i}(\theta) - \ln SP_{ij_i}^R(\theta)$$

for a given θ (such as the solution of the SAA problem). Since the logarithm is continuously differentiable on $\mathbb{R}_0^+$ and since $E_P\left[L_{ij_i}(\boldsymbol{\gamma},\theta)^2\right]$ is finite, we can use the Delta method (see for instance Borovkov [11], page 44, for the one-dimensional case or Rubinstein and Shapiro [37] section 6.3, for the multi-dimensional case) to conclude that

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

where $\sigma^2_{ij_i}(\theta)$ is the variance of $L_{ij_i}(\gamma, \theta)$. In other terms, we have that

$$\ln P_{ij_i}(\theta) - \ln SP^R_{ij_i}(\theta) \Rightarrow \frac{1}{P_{ij_i}(\theta)\sqrt{R}} N\left(0, \sigma^2_{ij_i}(\theta)\right).$$

As samples are independent between individuals, so are the normal distributions in this last limit, and we thus have that

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

Let α_δ be the quantile of a $N(0,1)$ associated to some level of significance δ, i.e. $P[-\alpha_\delta \leq X \leq \alpha_\delta] = \delta$, 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 practise 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 bigger sampling size R, but, as in other basic Monte Carlo methods, the convergence is only in $O\left(\sqrt{R}\right)$ (Fishman [17], page 8). However the population size also has an influence on the quality of the approximation. First of all, we note that

$$0 \leq \Delta^R_\delta(\theta) \leq \alpha_\delta \frac{1}{I}\sum_{i=1}^{I} \sqrt{\frac{\sigma^2_{ij_i}(\theta)}{R\left(P_{ij_i}(\theta)\right)^2}}.$$

If the total population is assumed to be infinite, then we may consider a population of size I as an independent and identically distributed sample within it. From now on, we also assume that $\frac{\sigma_{ij_i}(\theta)}{P_{ij_i}(\theta)}$ has finite mean and variance. We obtain from the strong law of large numbers that, almost surely,

$$0 \leq \Delta^R_\delta(\theta) \leq \frac{\alpha_\delta}{\sqrt{R}} E_I\left[\frac{\sigma_{ij_i}(\theta)}{P_{ij_i}(\theta)}\right].$$

In other terms, for a fixed sampling size R and a fixed θ, $\Delta^R_\delta(\theta)$ converges almost surely to some real value which is less than the expectation of individual errors, defined by

$$\frac{\alpha_\delta}{\sqrt{R}} E_I\left[\frac{\sigma_{ij_i}(\theta)}{P_{ij_i}(\theta)}\right].$$

Assume now that these quantities are almost surely finite, i.e. that there exists some κ such that for all θ in S, and for almost every individual i in the (infinite) population,

$$\frac{\sigma_{ij_i}(\theta)}{P_{ij_i}(\theta)} \leq \kappa$$

Then, from (5.3),

$$\Delta_\delta^R(\theta) \leq \alpha_\delta \frac{\kappa}{\sqrt{IR}}$$

almost surely. This suggests that the error decreases as the population size increases. However, we must remember that $E_P\left[SLL^R(\theta)\right] \neq LL(\theta)$, because of the logarithmic operator, and our confidence interval is thus centred at zero only asymptotically. However, since (5.2) implies that $LL(\theta) - SLL^R(\theta) \xrightarrow{p} 0$, when R tends to ∞ for a fixed population size I, we deduce that the estimator is consistent. To 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 obtain then 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

$$B^R(\theta) := E_P[SLL^R(\theta)] - LL(\theta) = -\frac{1}{2IR}\sum_{i=1}^{I} \frac{\sigma_{ij_i}^2(\theta)}{\left(P_{ij_i}(\theta)\right)^2} \leq 0, \tag{5.4}$$

which can be easily computed from the estimated error as

$$B^R(\theta) = -\frac{I}{2\alpha_\delta^2}\left(\Delta_\delta^R(\theta)\right)^2. \tag{5.5}$$

Thus, (5.4) implies that, up to second order,

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

It is 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, (5.5) shows that the simulation bias decreases faster than the error. This suggests that the number of random draws is heavily related to the nature of the model: 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 [24].

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

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

Therefore the maximisation procedure itself can produce another bias opposed to the bias of simulation. As a consequence, the solutions of successive SAA problems do not necessarily increase monotonically when R grows, 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 Hajivassiliou and McFadden [23] 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 everywhere, instead of in distribution.

From the strong law of large numbers,

$$B^R(\theta) \xrightarrow{a.s.} -\frac{1}{2R} E_I \left[\frac{\sigma^2_{ij_i}(\theta)}{(P_{ij_i})^2} \right],$$

where we have again assumed that $\frac{\sigma_{ij_i}(\theta)}{P_{ij_i}(\theta)}$ has finite mean and variance. Therefore the problem is consistent if and only if R tends to infinity when I tends to infinity, as reported by Hajivassiliou and McFadden [23] and Train [43], page 288. Taking the Taylor expansion around the true parameters, that solve $ELL(\theta) := E_I \left[\ln P_{ij_i}(\theta) \right]$, the expectation of the logarithm of the probability choice for all individuals i, these authors conclude that

- if R is fixed, the SAA problem is inconsistent;
- if R rises slower than $\sqrt{I}$, the SAA problem is consistent but not asymptotically normal;
- if R rises faster than $\sqrt{I}$, the SAA problem is consistent, asymptotically normal and efficient, equivalent to the true problem.

Note that these results are obtained using convergence in distribution of the solutions of the SAA problems. We provide, in the next theorem, results of the same type. They are now expressed almost surely, at the expense of not being directly computable.

Proposition 5.1. *Assume that a ULLN holds for the approximation $LL(\theta)$ of $ELL(\theta)$ 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 almost every i, and that $ELL(\theta)$ is continuous on S. Then*

$$\sup_{\theta \in S} \left| SLL^R(\theta) - ELL(\theta) \right| \xrightarrow{a.s.} 0$$

as 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,

$$\sup_{\theta} \left| E_I \left[\ln P_{ij_i}(\theta) \right] - \frac{1}{I} \sum_{i=1}^{I} \ln P_{ij_i}(\theta) \right| < \frac{\delta}{2} \quad \text{a.e.}$$

For such an I, we have, from the ULLN assumption for $SLL^R(\theta)$, that for R sufficiently high,

$$\sup_{\theta} \left| \frac{1}{I} \sum_{i=1}^{I} \ln P_{ij_i}(\theta) - \sum_{i=1}^{I} \ln \frac{1}{R} \sum_{r=1}^{R} P_{ij_i}^{R}(\theta) \right| < \frac{\delta}{2} \quad \text{a.e.}$$

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

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

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

$$\sup_{\theta} \left| SLL^R(\theta) - LL(\theta) \right| \to 0, \quad \text{a.e.,} \tag{5.6}$$

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}, (5.6) holds if we require $R_{I'}$ to be equal or larger than $R_{I_{n+1}}$. As a consequence, we obtain that (5.6) holds irrespective of the speed of growth of $\{I\}$ provided R grows sufficiently fast. □

Let $\{\theta^*_{I,R}\}$ be a sequence of SAA solutions for I tending to infinity, and R tending to infinity sufficiently fast compared to I. Dropping again the explicit dependence on the sampling process, let θ^* be a limit point of this sequence and assume (without loss of generality) that $\{\theta^*_{I,R}\}$ converges to θ^*. Then, under the assumptions of the previous proposition, we obtain from Lemma 3.1 that

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

for almost every sequence $\{\theta^*_{I,R}\}$. 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 $(I \times P)$-integrable), that, almost surely, a sequence $\{\theta^*_{I,R}\}_{I=1,\,R=1}^{\infty,\,\infty}$ 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 minimisers of the approximating problems are global to the case where they are only local or even first-order critical. 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]), whose description and assessment are available in a companion paper [4].

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 minimisation. 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 for his helpful comments on statistical theory and Stéphane Hess for his remarks on mixed logit theory, as well as to 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. Greg M. Allenby and Peter E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89:57–78, 1999.
2. Simon P. Anderson, Andre De Palma, and Jacques-Francois Thisse. *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, Massachusetts, USA, 1992.
3. Fabian Bastin, Cinzia Cirillo, and Philippe L. Toint. Numerical experiments with AMLET, a new Monte-Carlo algorithm for estimating mixed logit models. Electronic Proceeding (CD-ROM) of the 10th International Conference on Travel Behaviour Research, 2003.
4. Fabian Bastin, Cinzia Cirillo, and Philippe L. Toint. An adaptive monte carlo algorithm for computing mixed logit estimators. *Computational Management Science*, Submitted.
5. Moshe Ben-Akiva and Steven R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, 1985.
6. Chandra R. Bhat. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research*, 35B(7):677–693, August 2001.
7. Chandra R. Bhat. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research B*, 37(3):837–855, 2003.

8. Chandra R. Bhat and Saul Castelar. 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 B*, 36(3):593–616, 2002.
9. Chandra R. Bhat and Frank S. Koppelman. Activity-based modeling of travel demand. In Randolph W. Hall, editor, *Handbook of Transportation Science*, pages 35–61, Norwell, USA, 1999. Kluwer Academic Publisher.
10. John R. Birge and Franois Louveaux. *Introduction to Stochastic Programming*. Springer-Verlag, 1997.
11. Alexander Borovkov. *Statistique mathmatique*. Mir, 1987.
12. David Brownstone, David S. Bunch, and Kenneth Train. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research B*, 34(5):315–338, 2000.
13. Cinzia Cirillo and Kay W. Axhausen. Mode choice of complex tour. In *Proceedings of the European Transport Conference (CD-ROM)*, Cambridge, UK, 2002.
14. James Davidson. *Stochastic Limit Theory*. Oxford University Press, Oxford, England, 1994.
15. István Deák. Multidimensional integration and stochastic programming. In Y. Ermoliev and R. J.-B. Wets, editors, *Numerical Techniques for Stochastic Optimization*, pages 187–200. Springer Verlag, 1988.
16. Anthony V. Fiacco. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. Academic, New York, USA, 1983.
17. George S. Fishman. *Monte Carlo: Concepts, Algorithms and Applications*. Springer Verlag, New York, USA, 1996.
18. Philip E. Gill, Walter Murray, Michael Saunders, G. W. Stewart, and Margaret H. Wright. Properties of a representation of a basis for the null space. *Mathematical Programming*, 33(2):172–186, 1985.
19. Nicholas I. M. Gould. On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programming problem. *Mathematical Programming*, pages 90–95, 1985.
20. Martin Gugat. A parametric view on the mangasarian-fromovitz constraint qualification. *Mathematic Programming*, 85(1999):643–653, 1999.
21. Gül Gürkan, A. Yonca Özge, and Stephen M. Robinson. Sample-path solution of stochastic variational inequalities. *Mathematical Programming*, 84(2):313–333, 1999.
22. Gül Gürkan, A. Yonca Özge, and Stephen M. Robinson. Solving stochastic optimization problems with stochastic constraints: an application in network design. In D.T. Sturrock P. A. Farrington, H. B. Nembhard and G. W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 471–478, USA, 1999.
23. Vassilis A. Hajivassiliou and Daniel L. McFadden. The method of simulated scores for the estimation of LDV models. *Econometrica*, 66(4):863–896, 1998.
24. David A. Hensher and William H. Greene. The mixed logit model: The state of practice. *Transportation*, 30(2):133–176, 2003.
25. David A. Hensher and Charles Sullivan. Willingness to pay for road curviness and road type for long-distance travel in New Zealand. *Transportation Research D*, 8(2):139–155, 2003.
26. Stéphane Hess and John Polak. Mixed logit estimation of parking type choice. *Presented at the 83rd Transportation Research Board Annual Meeting*, 2004.
27. Stéphane Hess, John Polak, and Andrew Daly. 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.
28. Stéphane Hess, Kenneth Train, and John Polak. On the use of a modified latin hypercube sampling (mlhs) approach in the estimation of a mixed logit model for vehicle choice. *Transportation Research B*, Submitted.
29. Peter Kall and Stein W. Wallace. *Stochastic Programming*. John Wiley & Sons, 1994.
30. Daniel L. McFadden and Kenneth Train. Consumers' evaluation of new products: learning from self and others. *Journal of Political Economy*, 104(4):683–703, 1996.
31. Claude Montmarquette, Kathy Cannings, and Sophie Mahseredjian. How do young people choose college majors? *Economics of Education Review*, 21(6):543–556, 2002.
32. Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, USA, 1999.
33. Juan de Dios Ortúzar and Luis G. Willumsen. *Modelling Transport*. John Wiley & Sons, 3rd edition, 2001.
34. K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 1967.
35. Stephen M. Robinson. Local epi-continuity and local optimization. *Mathematical Programming*, 37(2):208–222, 1987.
36. Stephen M. Robinson. Analysis of sample-path optimization. *Mathematics of Operations Research*, 21(3):513–528, 1996.
37. Reuven Y. Rubinstein and Alexander Shapiro. *Discrete Event Systems*. John Wiley & Sons, Chichester, England, 1993.

38. Alexander Shapiro. Probabilistic constrained optimization: Methodology and applications. In S. Uryasev, editor, *Statistical inference of stochastic optimization problems*, pages 282–304. Kluwer Academic Publishers, 2000.
39. Alexander Shapiro. Stochastic programming by Monte Carlo simulation methods. *SPEPS*, 2000.
40. Alexander Shapiro. Monte Carlo sampling methods. In A. Shapiro and A. Ruszczyński, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 353–425. Elsevier, 2003.
41. Yosef Sheffi. *Urban Transportation Networks*. Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1985.
42. Kenneth Train. Halton sequences for mixed logit. Working paper No. E00-278, Department of Economics, University of California, Berkeley, 1999.
43. Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, New York, USA, 2003.