CONVERGENCE THEORY
FOR NONCONVEX STOCHASTIC PROGRAMMING
WITH AN APPLICATION TO MIXED LOGIT

by F. Bastin[1], C. Cirillo[2] and Ph. L. Toint[2]

Report 03/19 14 October 2003

[1] Belgian National Fund for Scientific Research
email: fabian.bastin@fundp.ac.be

[2] Department of Mathematics,
University of Namur,
61, rue de Bruxelles, B-5000 Namur, Belgium,
Email: cinzia.cirillo@fundp.ac.be, philippe.toint@fundp.ac.be

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

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

October 14, 2003

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 minimizers of the sample average approximation (SAA) problems as well as of the true problem. In this paper we extend these results to the case where computed solutions are only local or even first-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 recognized as an important area of operations research (see the books by Birge and Louveaux [10] and by Kall and Wallace [26], 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 (see Shapiro [35] for an overview of this topic). 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, or in the closure of an open bounded subset containing the set of minimizers of the true problem (see for instance Gürkan, Özge and Robinson [19]). As a result, they have been applied mostly to linear or convex problems, where this assumption is not restrictive.

It is our purpose to extend the theoretical understanding of this class of methods to the case where this global minimization assumption does no longer hold, i.e. where we allow the minimization of the subproblem to converge to a local minimizer, irrespectively to the set of minimizers of the true problem. This investigation is worthwhile, in particular because it opens the possibility to consider stochastic problems with nonconvex objective functions. We first give consistency results when only first-order critical

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

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 since in nonlinear programming the number of local solutions can be high, and they often do not share the same constraint qualification properties. Our results avoid the need to consider the complete set of minimizers of the true problem by focusing only on the limit points of the computed approximating solutions.

Such 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 off their inherent complexity, they are becoming very popular among researchers and practitioners in economics and transportation (see, for instance, Montmarquette, Cannings and Mahseredjian [28], Bhat and Castelar [8], Brownstone, Bunch and Train [12], Cirillo and Axhausen [13], Hensher and Greene [21], Hensher and Sullivan [22], Hess and Polak [23]). 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 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 generated from a given density and, for each draw, observed parts of the alternatives utilities, conditionally to this realization, are calculated 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 [20] show that the computed estimators are, under reasonable assumptions, asymptotically consistent and efficient. But even in this form the 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 [37] advocate Halton sequences for mixed logit models and find that they perform much better than random draws in simulation estimation. However Bhat [7] have pointed out that Halton sequences rapidly deteriorate in the coverage of the integration domain for high integration dimensions and proposed 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 [24] 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 [25] have proposed the use of randomly shifted uniform vectors 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, practioners report that Monte-Carlo methods are again competitive compared to quasi Monte-Carlo approaches (Deak [15], Hess, Train and Polak [25]). 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 [38]), 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 the simulation variance.

The paper is organized as follows. We introduce the general stochastic problem and its application to mixed logit models in Section 2. Section 3 discusses our convergence theory for the general problem, while Section 5 applies those results 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 of a random variable (see Birge and Louveaux [10] or Kall and Wallace [26] for a more complete exposition):

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

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 $\boldsymbol{\omega}$ is a random vector defined on the probability space $(\Omega, \mathcal{F}, P)$ and where $G : \mathbb{R}^m \times \Omega \to \mathbb{R}$ is a real valued function. 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. A convenient class of functions for our analysis is the class of Carathéodory function. G is a Carathéodory function if $G(z, \boldsymbol{\omega})$ is continuous in z and measurable in $\boldsymbol{\omega}$. For simplicity, we restrict ourselves in a first step to the case where the set S is deterministic and convex.

If the distribution function of $\boldsymbol{\omega}$ is continuous or discrete with a large number of possible realizations, $g(z)$ is usually very difficult to evaluate. Solving the problem (2.1) directly thus becomes difficult and we have to turn to approximations as Monte-Carlo methods (see Shapiro [35] for a review). In these methods, the original problem (2.1) is replaced by successive approximations obtained by generating samples $\omega_1, \ldots, \omega_N$. The approximation for a sample of size N is

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

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 [27] and Allenby and Rossi [1]) or between transportation modes for travel (see Sheffi [36], Ortúzar and Willumsen [30]) 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 denote $\mathcal{A}$ 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 maximizes 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 althernative 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.

If we now assume that the residuals ϵ_{ij} are independently Gumbel distributed (with mean 0 and scale factor 1.0), the probability that the individual i chooses the alternative j can be expressed as

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

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. It expresses that, if some alternatives are removed from a choice set, the relative choice probabilities in the reduced choice set then 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, in particular, that its validity depends on the structure on the choice set, and also that it is 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 component structure models) (see Bhat and Kopelman [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 interalternative 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 each parameters 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 $\boldsymbol{\gamma}$ and a parameters vector θ, which we express as

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

γ typically specifies the random nature of the model while θ quantifies the population characteristics for the model. Usually, $\boldsymbol{\beta}$ follows itself a probability distribution, and θ specifies the parameters of this distribution. Therefore we have that $f(\boldsymbol{\beta}|\theta) = f(h(\boldsymbol{\gamma}, \theta))$, where f designs the underlying distribution function. We will nevertheless use the notation (2.3) in order to emphasize 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 independant normal variables whose k-th component is $N(\mu_k, \sigma_k)$. 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 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 standard logit formula

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

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

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

where P is the probability measure associated to γ and $f(\cdot)$ is its distribution function.

The unknown values of θ are estimated by maximizing 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

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

where R is the number of random draws γ_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

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

However, since I can be large (usually in the thousands), the evaluation of $SLL^R(\theta)$ may remain very expensive, even on modern computers, as pointed out by Hensher and Greene [21].

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

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

The associated sample average approximation problem is then written as

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

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 approaching infinity. Our framework extends the classical theory to the weaker algorithmic requirement that subproblems are only solved locally instead of globally, which makes it applicable to nonconvex problems.

3.1. Deterministic and convex constraints

Let z_N^* be a first-order critical point of $\hat{g}_N(\cdot)$ as defined in (2.2) and consider the sequence

$$\{z_N^*\}_{N=1}^{\infty},$$

constructed by taking larger and larger samplings. Since S is a compact set, this sequence has some limit point z^*. Without loss of generality we identify the sequence $\{z_N^*\}_{N=1}^{\infty}$ with one of its subsequences converging to z^* from now on. Our first aim is to show that, under reasonable assumptions, z^* is a first-order critical point for the true problem (2.1).

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

- i.i.d. for independently and identically distributed;
- $a.e.$ for almost everywhere;
- $\xrightarrow{a.s.}$ for almost sure convergence;
- $\xrightarrow{p}$ for convergence in probability;
- $\Rightarrow$ for convergence in distribution;
- $N(\mu, \sigma)$ for normal distribution of mean μ and standard deviation σ.

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. 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 now state our assumptions.

A.0 The random draws are independently and identically distributed.
A.1 For P-almost every ω, the function $G(\cdot, \omega)$ is continuously differentiable on S.
A.2 The family $G(z, \omega)$, $z \in S$, is dominated by a P-integrable function $K(\omega)$, i.e. $E_P[K]$ is finite and $|G(z, \omega)| \leq K(\omega)$ for all $z \in S$ and P-almost every ω.
A.1 obvioulsy implies that $G(\cdot, \omega)$ is continuous almost surely. This and **A.2** are typical assumptions of stochastic programming theory (see Rubinstein and Shapiro [33], for instance). 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 [33] 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 allows us to develop the following technical lemma.

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

$$\hat{g}_N(z_N^*) \xrightarrow{a.s.} g(z^*). \tag{3.1}$$

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

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

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

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

then, for any real scalar a,

$$a\hat{h}_N\left(z_N^*\right) + \hat{g}_N\left(z_N^*\right) \xrightarrow{a.s.} ah\left(z^*\right) + g\left(z^*\right), \tag{3.3}$$

and

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

Proof. We first prove (3.1), or, more precisely, that

$$\forall \delta > 0, \exists N_1 \text{ s.t. } \forall N > N_1, \left|\hat{g}_N\left(z_N^*\right) - g\left(z^*\right)\right| < \delta \quad \text{a.e.} \tag{3.5}$$

The triangle inequality gives that

$$\left|\hat{g}_N\left(z_N^*\right) - g\left(z^*\right)\right| \leq \left|\hat{g}_N\left(z_N^*\right) - g\left(z_N^*\right)\right| + \left|g\left(z_N^*\right) - g\left(z^*\right)\right|. \tag{3.6}$$

¿From the ULLN property, we deduce that $\sup_{z\in S}|\hat{g}_N(z) - g(z)| \xrightarrow{a.s} 0$. Therefore, there exists some $N_1 > 0$ such that for all $N > N_1$,

$$\left|\hat{g}_N\left(z_N^*\right) - g\left(z_N^*\right)\right| \leq \sup_{z\in S}\left|\hat{g}_N\left(z\right) - g\left(z\right)\right| < \frac{\delta}{2} \quad \text{a.e.} \tag{3.7}$$

Moreover, the continuity of $g(z)$ on S and the fact that z^* is the limit point of $\{z_N^*\}_{N=1}^{\infty}$, together imply that for N large enough, say $N > N_2 > 0$,

$$\left|g\left(z_N^*\right) - g\left(z^*\right)\right| < \frac{\delta}{2}. \tag{3.8}$$

Combining (3.6), (3.7) and (3.8), we have $|\hat{g}_N(z_N^*) - g(z^*)| < \delta$ almost everywhere, for N larger than $\max(N_1, N_2)$, which proves (3.5).

We now prove (3.2). Let $\delta > 0$. Then, by continuity,

$$\exists\, \alpha_\delta \text{ s.t. } \forall x, y \in dom(f),\ |x - y| < \alpha_\delta \Rightarrow |f(x) - f(y)| < \delta.$$

But, since $\hat{g}_N(z_N^*)$ and $g(z^*)$ both belong to the domain of f, we have that

$$\exists\, N_\delta \text{ s.t. } \forall N > N_\epsilon,\ |\hat{g}_N(z_N^*) - g(z^*)| < \alpha_\delta \quad \text{a.e.}$$

Combining these two inequalities, we obtain that

$$\exists\, N_\delta \text{ s.t. } \forall N > N_\delta,\ |f(\hat{g}_N(z_N^*)) - f(g(z^*))| < \delta \quad \text{a.e,}$$

which is (3.2).

Finally, we prove (3.3) and (3.4). The result (3.3) is immediate. (3.4) follows easily from the following relations:

$$\begin{aligned}
&|h(z_N^*)g(z_N^*) - h(z^*)g(z^*)| \\
&\quad = |(h(z_N^*) - h(z^*))(g(z_N^*) - g(z^*)) + g(z^*)(h(z_N^*) - h(z^*)) \\
&\qquad + h(z^*)(g(z_N^*) - g(z^*))| \\
&\quad \leq |h(z_N^*) - h(z^*)|\,|g(z_N^*) - g(z^*)| + |g(z^*)|\,|h(z_N^*) - h(z^*)| \\
&\qquad + |h(z^*)|\,|g(z_N^*) - g(z^*)|.
\end{aligned}$$

□

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 $\frac{\partial}{\partial z_l} G(z, \omega)$, $l = 1, \ldots, m$, are dominated by a P-integrable function.

This new assumption allows us to apply the results of Rubinstein and Shapiro [33], page 71, and deduce that the expected value function $g(z)$ is continuous and differentiable at y almost surely, and that expectation can be interchanged in the expression of the gradient, that is

$$\nabla_z g(y) = E\left[\nabla_z G(y, \omega)\right].$$

We then have that

$$E\left[\nabla \hat{g}_N(z^*)\right] = \frac{1}{N}\sum_{i=1}^{N} E\left[\nabla_z G(z^*, \omega^i)\right] = \frac{1}{N}\sum_{i=1}^{N} \nabla_z E\left[G(z^*, \omega^i)\right] = \nabla g(z^*).$$

i.e. $\nabla \hat{g}_N(z^*)$ is an unbiaised estimator of $\nabla g(z^*)$. We may now verify that z^* is a first-order critical point.

Theorem 3.1. *Assume that* ***A.0–A.3*** *hold. Then z^* is a first-order critical point of* (2.1) *almost surely.*

Proof. Since z_N^* is a first-order critical point of $\hat{g}_N(z)$, we have that (see e.g. Nocedal and Wright [29])

$$-\nabla_x \hat{g}_N(z_n^*) \in \mathcal{N}(z_N^*),$$

where $\mathcal{N}(z_N^*)$ is the normal cone to S at z_N^*. Since S is convex, this implies (see Rockafellar and Wets [32], Section 6.9) that each z_N^* satisfies the following first-order conditions for (2.2), i.e.

$$\langle \nabla_z \hat{g}_N(z_N^*), u - z_N^* \rangle \geq 0, \text{ for all } u \in S. \tag{3.9}$$

¿From Lemma 3.1, applied to the derivatives, we have $\nabla_z \hat{g}_N(z_N^*) \xrightarrow{a.s.} \nabla_z g(z^*)$. Therefore, taking the limit as N tends to infinity in (3.9), we obtain that

$$\langle \nabla_z g(z^*), u - z^* \rangle \, , \text{ for all } u \in S, \quad \text{a.e.}$$

In other terms, $-\nabla_z g(z^*) \in \mathcal{N}(z^*)$ almost everywhere. Thus z^* is a first-order critical point for (2.1) almost surely, as desired. □

3.2. Stochastic constraints

It is also possible to prove first-order convergence when S is nonconvex or non-deterministic under stronger assumptions. We suppose that the feasible set can be described by equality and inequality constraints. The original problem is then stated as follows

$$\begin{aligned} &\min_{z \in V} g(z) = E_P[G(z, \boldsymbol{\omega})], \\ &\text{subject to } c_j(z) \geq 0, \ j = 1, \ldots, k, \\ &\qquad c_j(z) = 0, \ j = k+1, \ldots, M, \end{aligned} \tag{3.10}$$

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

$$\begin{aligned} &\min_{z \in V} \hat{g}_N(z), \\ &\text{subject to } \hat{c}_{jN}(z) \geq 0, \ j = 1, \ldots, k, \\ &\qquad \hat{c}_{jN}(z) = 0, \ j = k+1, \ldots, M. \end{aligned} \tag{3.11}$$

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

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

The expected value functions $c_j(\cdot)$, $j = 1, \ldots, M$, can be estimated by the corresponding sample mean functions

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

For simplicity, we will consider the more general parametric mathematical programming problem

$$\begin{aligned} &\min_{z \in V} \hat{g}(z, \boldsymbol{\epsilon}), \\ &\text{subject to } \hat{c}_j(z, \boldsymbol{\epsilon}) \geq 0, \ j = 1, \ldots, k, \\ &\qquad \hat{c}_j(z, \boldsymbol{\epsilon}) = 0, \ j = k+1, \ldots, M, \end{aligned} \tag{3.12}$$

where $\boldsymbol{\epsilon}$ is a vector of parameters giving perturbations of the program (3.12); $g(\cdot)$, $\hat{g}(\cdot, \boldsymbol{\epsilon})$, $c_j(\cdot)$, $\hat{c}_j(\cdot, \boldsymbol{\epsilon})$ are assumed to be twice continuously differentiable.

We will assume that the perturbation is of the form

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

and

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

In the Monte-Carlo context, we can define $\boldsymbol{\epsilon}_N$ as

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

We will assume that $\boldsymbol{\epsilon}_N$ converges uniformly on V to 0 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.

The Lagrangian functions associated to (3.10) and (3.12) are respectively

$$\mathcal{L}(z, \lambda) = g(z) - \sum_{j=1}^{M} \lambda_j c_j(z)$$

and

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

Let $z^*(\boldsymbol{\epsilon})$ denotes a first-order critical point for program (3.12). Therefore there exists 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

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

As before, since V is compact, $z^*(\epsilon_N)$ has some limit point z^* as $N \to \infty$, and without loss of generality, we can assume that $z^*(\epsilon_N) \to z^*$ as $N \to \infty$. We can now prove first-order convergence for the general case.

Theorem 3.2. *Assume that*

(a) $\boldsymbol{\epsilon}_N \xrightarrow{a.s.} 0$ *uniformly on V, as* $N \to \infty$,
(b) $\lambda^*(\boldsymbol{\epsilon}_N)$ *has some limit point* λ^* *almost surely as* $N \to \infty$,

Then z^* is a first-order critical point for (3.10)*.*

Proof. Consider the sequence $\{(z^*(\epsilon_N), \lambda^*(\epsilon_N))\}$, $N = 1, \ldots, \infty$. From (b), there exists a subsequence converging almost surely to (z^*, λ^*). (a) and Lemma 3.1 implies that (z^*, λ^*) satisfies the KKT conditions for the true problem. $\square$

Note that assumption (b) always holds if the multipliers remain bounded.

4. Second-order convergence

4.1. Deterministic constraints

We may even verify that z^* is a local minimizer, if we further strengthen our assumptions. We first consider the case where S is deterministic and assume that z_N^* is a local minimum of $\hat{g}_N(x)$. This is to say that

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

where $B(x, d)$ is the open ball centered at x and of radius d. In order to show that z^* is a local minimum of $g(\cdot)$, we must therefore have that the neighbourhood in which z_N^* is a local minimum does not shrink to a singleton when $N \to \infty$. We express this requirement by the following technical assumption.

A.4 There exists $a > 0$ and $N_a > 0$ such that, for all $N \geq N_a$,

$$\exists \delta_N \geq a \text{ s.t. } \forall x \in B(z_N^*, \delta_N) \cap S, \hat{g}_N(z_N^*) \leq \hat{g}_n(z).$$

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

Theorem 4.1. *Assume that **A.0–A.4** hold. Then z^* is a local minimum of $g(\cdot)$ almost surely.*

Proof. Let z' be a minimizer of g in $K = \overline{B}\left(z^*, \frac{a}{2}\right) \cap S$, where a is defined as in **A.4**. Then for N sufficiently large, we have that,

$$\overline{B}\left(z^*, \frac{a}{2}\right) \subset B\left(z_N^*, \epsilon_N\right) \text{ and } z_N^* \in \overline{B}\left(z^*, \frac{a}{2}\right). \tag{4.1}$$

Since z^* is the limit point of $\{z_N^*\}_{N=0}^{\infty}$, the second part of (4.1) is satisfied for N sufficiently large. Take now $z \in \overline{B}\left(z^*, \frac{a}{2}\right)$. We have that, $|z - z_N^*| \leq |z - z^*| + |z^* - z_N^*|$. Thus

$$|z - z_N^*| \leq \frac{a}{2} + \frac{a}{2} = a \leq \delta_N.$$

Therefore $z \in \overline{B}\left(z_N^*, \epsilon_N\right)$, so the first part of (4.1) is also satisfied when N is sufficiently high.

We now verify that

$$\left|\hat{g}_N\left(z_N^*\right) - g\left(z'\right)\right| \xrightarrow{a.s.} 0. \tag{4.2}$$

Assume indeed that $\hat{g}_N\left(z_N^*\right) \leq g(z')$, so

$$|\hat{g}_N(z_N^*) - g(z')| = g(z') - \hat{g}_N(z_N^*)$$

Since z' minimizes $g(\cdot)$ in K and, from (4.1), $z_N^* \in K$, we have that

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

But the ULLN property implies that

$$\sup_{z \in S} |\hat{g}_N(z) - g(z)| \xrightarrow{a.s.} 0,$$

and therefore

$$|g(z') - \hat{g}_N(z_N^*)| \leq 0 \quad \text{a.e.} \tag{4.3}$$

Assume now that $\hat{g}_N(z_N^*) \geq g(z')$. Since $z' \in K$, from the first part of (4.1), $\hat{g}_N(z_N^*) \leq \hat{g}_N(z')$. Therefore we can write

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

and we deduce that

$$|g(z') - \hat{g}_N(z_N^*)| \geq 0 \quad \text{a.e.} \tag{4.4}$$

Combining (4.3) and (4.4), we obtain (4.2). ¿From Lemma 3.1, $\hat{g}_N(z_N^*) \xrightarrow{a.s} g(z^*)$, so we have that $g(z') = g(z^*)$. In other terms, z^* is a local solution of (2.1) almost surely.

Classical results, where global minimizers are considered, express that dist(z_N^*, S^*) converges almost surely to 0 as N tends to infinity, where S^* is the set of minimizers of the true problem (Rubinstein and Shapiro [33], Section 6.1, Gürkan, Özge and Robinson [19]). However, Assumption **A.4** reflects that this property is not necessarily true if z_N^* is a local minimizer of the approximating problem. Consider for instance the problem

$$\min_{z \in [-1,1]} z^3 - \frac{z}{2} E[\boldsymbol{\omega}], \tag{4.5}$$

where $\Omega = \{-1, 1\}$ and $P[\omega = -1] = P[\omega = 1] = 0.5$, so $E[\boldsymbol{\omega}] = 0$. Therefore (4.5) has only one local minimizer, which is also global, at $z^* = -1$. The SAA problem is then

$$\min_{z \in [-1,1]} z^3 - \frac{z}{2N} \sum_{i=1}^{N} \omega_i. \tag{4.6}$$

(4.6) has two (isolated) local minimizers,

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

when $\sum_{i=1}^N \omega_i > 0$. We have that $P\left[\sum_{i=1}^N \omega_i > 0\right] \to 0.5$ when $N \to \infty$, but

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

since, from the strong law of large numbers, $\frac{1}{N}\sum_{i=1}^{N}\omega_i \to E[\boldsymbol{\omega}] = 0$ almost surely as $N \to \infty$. But 0 is a saddle point of the true problem (4.5), 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 objectives as well as for all derivatives.

Assumption **A.4** is somewhat artificial so it is valuable 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 [33], it is important to note that sensitivity 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 neighborhood of this original solution. On the other hand, we will focus here on conditions under which the limit point of a sequence of approximating solutions is a solution of the true problem irrespective of the existence of a set of minimizers for the original problem. The difference be will more formally illustrated at the end of the section where we compare our developments to some sentivity analysis results.

We consider the case where the feasible set is described by a set a equality and inequality constraints, as in (3.10). Without loss of generality, we can assume that $z^*(\boldsymbol{\epsilon}_N) \to z^*$ almost surely, as N tends to infinity. If there is a unique Lagrange multipliers vector $\lambda^* \equiv \lambda^*(0)$ associated to z^* that satisfies the KKT conditions, then under some conditions $\lambda^*(\boldsymbol{\epsilon}_N) \to \lambda^*$ almost surely as $N \to \infty$.

Lemma 4.1. *Assume that the Jacobian of*

$$\begin{aligned} \nabla_z \mathcal{L}(z,\lambda) &= 0, \\ \lambda_j c_j(z) &= 0, \quad j = 1,\ldots,M, \\ c_j(z) &= 0, \quad j = k+1,\ldots,M, \end{aligned} \tag{4.7}$$

is well defined at (z^*,λ^*) and nonsingular. Then $\lambda^*(\boldsymbol{\epsilon}_N)$ converges almost surely to λ^* as N tends to infinity.

Proof. Consider the linear system

$$\begin{aligned} \nabla_z L(z,\lambda,\boldsymbol{\epsilon}) &= 0, \\ \lambda_j \hat{c}_j(z,\boldsymbol{\epsilon}) &= 0, \quad j = 1,\ldots,M, \\ c_j(z,\boldsymbol{\epsilon}) &= 0, \quad j = k+1,\ldots,M. \end{aligned} \tag{4.8}$$

¿From our initial assumptions, this system is continuously differentiable in all the arguments. Moreover, the Jacobian of the left-hand side with respect to (z,λ) is invertible at (z^*,λ^*).

¿From the implicit function theorem (see for instance Fiacco [16], page 36), in a neighborhood $\mathcal{B}$ of $(z^*,\lambda^*,0)$, there exists a unique continuously differentiable function $\gamma(\boldsymbol{\epsilon}) = (z^*(\boldsymbol{\epsilon}),\lambda^*(\boldsymbol{\epsilon}))^T$ satisfying (4.8) with $(z^*(0),\lambda^*(0)) = (z^*,\lambda^*)$. Since $z^*(\boldsymbol{\epsilon}_N) \to z^*$ almost surely, $z^*(\boldsymbol{\epsilon}_N)$ belongs almost surely to $\mathcal{B}$ for N large enough. But $z^*(\boldsymbol{\epsilon}_N), \lambda^*(\boldsymbol{\epsilon}_N)$ satisfies (4.8), so from the unicity and continuity of $\gamma(\boldsymbol{\epsilon})$, $\lambda^*(\boldsymbol{\epsilon}_N)$ tends to λ^* almost surely as N tends to infinity.

Therefore, $\nabla L(z^*(\boldsymbol{\epsilon}_N),\lambda^*(\boldsymbol{\epsilon}_N))$ converges almost surely to $\nabla\mathcal{L}(z^*,\lambda^*)$, when N tends to infinity. The unicity of λ^* can be ensured with a suitable constraint qualification, as the linear independence constrained 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 (3.10). 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:

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

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 [29]. 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 $j \in \{1, \ldots, k\} \cap \mathcal{A}(z^*)$.*

Assume that the strict complementarity condition and the LICQ hold at (z^*, λ^*) for program (3.12). If $\lambda^*(\boldsymbol{\epsilon}_N) \to \lambda^*$ almost surely we obtain then that almost surely $\lambda_N(\boldsymbol{\epsilon}_N)$, $j \in J(z^*)$, are strictly positive and hence the corresponding constraints are active at $z^*(\boldsymbol{\epsilon}_N)$ for all N large enough. Moreover, since $\hat{c}_j(z^*(\boldsymbol{\epsilon}_N), \boldsymbol{\epsilon}_N) \to c(z^*)$ almost surely, we have that for all N large enough, almost surely, $\mathcal{A}(z^*(\boldsymbol{\epsilon}_N)) = \mathcal{A}(z^*)$ and strict complementarity condition holds at $z^*(\boldsymbol{\epsilon}_N)$, associated to the Lagrangian multipliers $\lambda^*(\boldsymbol{\epsilon}_N)$, for program (3.11).

Theorem 4.2 (Second-order convergence). *Assume that λ^* is the unique vector of Lagrangian multipliers associated to program* (3.10) *at z^*, and that*

(a) $\boldsymbol{\epsilon}_N \xrightarrow{a.s.} 0$ uniformly on V, as $N \to \infty$,
(b) $z^*(\boldsymbol{\epsilon}_N) \xrightarrow{a.s.} z^*$, $\lambda^*(\boldsymbol{\epsilon}_N) \xrightarrow{a.s.} \lambda^*$ as $N \to \infty$,
(c) $\nabla^2_{zz}\hat{g}(z^*(\boldsymbol{\epsilon}_N), \boldsymbol{\epsilon}_N) \xrightarrow{a.s.} \nabla^2_{zz} g(z^*)$ as $N \to \infty$,
(d) $\nabla^2_{zz}\hat{c}_j(z^*(\boldsymbol{\epsilon}_N), \boldsymbol{\epsilon}_N) \xrightarrow{a.s.} \nabla^2_{zz} c_j(z^*)$, $j = 1, \ldots, M$, as $N \to \infty$.

Suppose also that the strict complementarity condition and the LICQ holds at (z^*, λ^*) for (3.10)*. Then, almost surely,*

(i) the LICQ holds at $(z^*(\boldsymbol{\epsilon}_N), \lambda^*(\boldsymbol{\epsilon}_N))$,
(ii) (z^*, λ^*) satisfy the second-order necessary condition for (3.10)*:*

$$w^T \nabla^2_{zz}\mathcal{L}(z^*, \lambda^*) w \geq 0, \text{ for all } w \in Null[\nabla_z \hat{c}_j(z^*)^T]_{j \in \mathcal{A}(z^*)}. \tag{4.9}$$

If furthermore there exists some $\alpha > 0$ such that for all N large enough

$$w^T \nabla^2_{zz} L_N(z_N^*, \lambda_N^*) w > \alpha, \text{ for all } w \in Null[\nabla_z \hat{c}_j(z^*(\boldsymbol{\epsilon}_N))^T]_{j \in \mathcal{A}(z^*(\boldsymbol{\epsilon}_N))},$$

then z^*, associated to λ^*, satisfies the second-order sufficient conditions for program (3.10)*:*

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

In other terms v^* is an isolated local minimizer of (3.10)*.*

Proof. In order to show (i), consider the matrix formed by the gradients of active constraints at z^* for (3.10):

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

From the strict complementarity conditions and convergence of Lagrange multipliers, the active set of program (3.11) at z_N^* is asymptotically the same as the active set of program (3.10) at z^*, almost surely. Since $\boldsymbol{\epsilon}_N \to 0$ uniformly on V, almost surely, we have that the matrix formed by the active constraints of the perturbed problem,

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

converges to (4.11) almost surely as N tends to infinity:

$$\{\nabla_z \hat{c}_j(z^*(\boldsymbol{\epsilon}_N),\boldsymbol{\epsilon}_N)\}_{j\in\mathcal{A}(z^*(\boldsymbol{\epsilon}_N))} \xrightarrow{a.s.} \{\nabla_z c_j(z^*)\}_{j\in\mathcal{A}(z^*)}. \tag{4.13}$$

The LICQ amounts to say that any square submatrice of the Jacobian of (4.11) is nonsingular. From (4.13), the same is true almost surely for (4.12) for N large enough. Thus the LICQ holds almost surely for the approximating problems when N is sufficiently large.

We show now (ii). From (4.13) we can associate a basis K_N to the nullspace of (4.12) such that

$$K_N \xrightarrow{a.s.} K, \tag{4.14}$$

where K is a basis of Null$[\nabla_z c_j(z)^T]_{j\in\mathcal{A}(z^*)}$ (see Gill and al [18]).

$(z^*(\boldsymbol{\epsilon}_N),\lambda^*(\boldsymbol{\epsilon}_N))$ satisfy the second-order necessary conditions, that can be expressed as, from the strict complementarity conditions and LICQ,

$$K_N^T \nabla^2_{zz} L(z^*(\boldsymbol{\epsilon}_N),\lambda^*(\boldsymbol{\epsilon}_N),\boldsymbol{\epsilon}_N) K_N \text{ is semi-positive definite.}$$

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

$$K_N^T \nabla^2_{zz} L(z^*(\boldsymbol{\epsilon}_N),\lambda^*(\boldsymbol{\epsilon}_N),\boldsymbol{\epsilon}_N) K_N \xrightarrow{a.s.} K^T \nabla^2_{zz}\mathcal{L}(z^*,\lambda^*) K.$$

Therefore we have (4.9) and (4.10).

Note that 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 neighborhood $\mathcal{N}$ of z^* such that z^* is the only local minimizer in $\mathcal{N}$. Recall that isolated local minimizers are also strict local minimizers but that the inverse is not always true (Nocedal and Wright [29], page 14). If z^* is a strict but not isolated local minimizer every neighborhood of z^* contains other local minimizers than z^*, 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^* can be difficult to identify.

Recall that we have proved the convergence of the Lagrange multipliers by assuming that the Jacobian of (4.7) is nonsingular. The same assumption can be used to replace the non-degeneracy hypothesis, as shown in the corollary below.

Corollary 4.1. *Assume that λ^* is the unique vector of Lagrangian multipliers associated to program* (3.10) *at z^*, and that*

(a) $\epsilon_N \xrightarrow{a.s.} 0$ *as* $N \to \infty$,
(b) $z^*(\epsilon_N) \xrightarrow{a.s.} z^*$, *as* $N \to \infty$,
(c) the Jacobian of (4.7) *is nonsingular,*
(d) $\nabla^2_{zz}\hat{g}(z^*(\epsilon_N), \epsilon_N) \xrightarrow{a.s.} \nabla^2_{zz}g(z^*)$ *as* $N \to \infty$,
(e) $\nabla^2_{zz}\hat{c}_j(z^*(\epsilon_N), \epsilon_N) \xrightarrow{a.s.} \nabla^2_{zz}c_j(z^*)$, $j = 1, \ldots, M$, *as* $N \to \infty$.

Suppose also that the strict complementarity condition and the LICQ holds at (z^*, λ^*) *for* (3.10). *Then* z^*, *associated to* λ^* *satisfies the second-order sufficient conditions for program* (3.10)*:*

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

Proof. ¿From Theorem 4.2, z^* associated to λ^* satisfies the second-order necessary conditions for program (3.10). But (c) implies that

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

Therefore (z^*, λ^*) satisfies the second-order sufficient property for program (3.10).

The converse of Theorem 4.2 can be obtained from classical results of sensitivity analysis (Fiacco [16], Theorem 3.2.2), that we restate for completeness. More developments in the context of stochastic programming can be found in Rubinstein and Shapiro [33] and Shapiro [34].

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

(a) the functions defining (3.12) *are twice continuously differentiable in* z *and their gradients with respect to* z *and the constraints are once continuously differentiable in* ϵ *in a neighborhood of* $(z^*, 0)$,
(b) the second-order sufficient conditions for a local minimum of (3.12) *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* (3.12) *with* $\epsilon = 0$ *and the associated Lagrange multipliers* λ^* *are unique,*
(ii) for ϵ *in a neighborhood of 0, there exists a unique, once continuously differentiable vector function* $\gamma(\epsilon) = (x(\epsilon), \lambda(\epsilon))^T$ *satisfying the second-order sufficient conditions for a local minimum of problem* (3.12) *such that* $\gamma(0) = (z^*, \lambda^*)^T$, *and hence* $z(\epsilon)$ *is a local isolated mimizer of problem* (3.12) *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* $z^*(\epsilon)$.

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 sensitivity analysis can be obtained by using epi-continuity arguments (Robinson [31]).

4.2. Asymptotic analysis of the optimal value

In this section we briefly discuss the asymptotic behaviour of the optimal values $\hat{g}(z_N^*)$ of the sample average approximations. Our differentiability assumptions allow us to extend some results given in Rubinstein and Shapiro [33], at the price of stronger assumptions.

We consider here the case where S is deterministic. In order to be able to derive an elegant asymptotic distribution of the optimal value, we require that z^* is the only limit point in a neighborhood of z^*:

A.5 z^* is an isolated local minimizer of $g(z)$ as defined in (2.1).

Consider the sample average approximation (2.2) of the true problem (2.1), and as before, without loss of generality, assume that z_N^* converges to z^* almost surely as N tends to infinity. We have the following theorem.

Theorem 4.4. *Assume that **A.0** and **A.5** hold, and that $z_N^* \to z^*$ almost surely. Let $\mathcal{N}$ be a neighborhood of z^* where z^* is an isolated minimizer of $g(\cdot)$. Under the following assumptions:*

(a) for every $z \in S$, the function $G(z, \cdot)$ is $\mathcal{F}$-measurable;
(b) for some $z \in S \cap \mathcal{N}$, the expectation $E_P[G(z, \boldsymbol{\omega})^2]$ is finite;
(c) there exists an $\mathcal{F}$-measurable function $K : \Omega \to \mathbb{R}$ such that $E_P[K(\boldsymbol{\omega})^2]$ is finite and

$$|g(z_1, \omega) - g(z_2, \omega)| \leq K(\omega)\|z_1 - z_2\|$$

for all $z_1, z_2 \in S \cap \mathcal{N}$ and almost every $\omega \in \Omega$;
(d) for all N sufficiently large, $\hat{g}_N(z_N^*) \leq \hat{g}_N(z)$ for all $z \in \mathcal{N}$ if $z_N^* \in \mathcal{N}$,

we have that

$$\sqrt{N}(\hat{g}_N(z_N^*) - g(z^*)) \Rightarrow N(0, \sigma^2)$$

almost surely, where

$$\sigma^2 = E_P[G(z^*, \boldsymbol{\omega})^2] - g(z^*)^2.$$

Proof. Since $z_N^* \to z^*$ almost surely as $N \to \infty$, there exists some N_0 such that for all $N \geq N_0$, z_N^* belongs to $\mathcal{N}$ almost surely. Define $S' = S \cap \mathcal{N}$. z^* can be seen as the unique global minimizer of $g(z)$ over S' and almost surely z_N^* is also a global minimizer of $\hat{g}_N(z)$ over S'. The result then follows directly from Theorem 6.4.3 of Rubinstein and Shapiro [33].

Theorem 4.3 is stronger than the classical central limit theorem in the sense that the approximating and the true objectives are not evaluated at the same point. When global solutions are found more detailed results can be derived (Rubistein and Shapiro [33], section 6.5). They are still valid if for N large enough Assumption (d) of Theorem 4.3 is satisfied. Further research would include reexamination of this assumption and investigation of the complete local minimizers set concept, introduced by Robinson [31], to develop a better framework for local optimization.

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).

We first specify that **A.0** is now to be understood as the requirement that different samples used to compute the choice probabilities are identically distributed and independent both for each individual and accross them.

Next note that the assumption, in the stochastic programming case, that the set S is compact ensures that the solutions of problem (2.2) remain in a bounded domain of $\mathbb{R}^m$, but that 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.6 The solution θ_R^* of the simulated mixed-logit problem (2.8) remain in some convex compact set S 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.6** 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 plus or minus infinity. As for the stochastic programming case, this assumption implies that the sequence $\{\theta_R^*\}$ has limit points, and, again as above, we identify it, without loss of generality, to one of its convergent subsequences and assume that $\theta_R^* \to \theta^*$ as $R \to \infty$.

For obtaining convergence to first-order critical points, we also need to ensure **A.1** and **A.2** by imposing suitable conditions of the problem's components.

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

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

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

A.2 is automatically satisfied since $|SP_{ij_i}(\gamma, \theta)| \leq 1$ for all θ and 1 is obviously P-integrable with unit expectation. We obtain from **A.6**, **A.1ml** and Lemma 3.1 that, for all individuals i, $i = 1, \dots, I$,

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

A.3 now becomes

A.3ml For $t = 1, \ldots, m$, $\frac{\partial}{\partial \theta_t} V_{ij}\left(\gamma_r, \theta, x_{ij_i}\right)$, $j = 1, \ldots, N$, is dominated by a P-integrable function.

If the utilities are linear in θ, as is nearly always the case in applications, the derivatives are independent of θ. All we have to assume is then that the expectation of the absolute partial derivatives is finite, which is usually not restrictive.

We may now apply Lemma 3.1 and deduce that

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

as $R \to \infty$. Using the same arguments as in Theorem 3.1, we then obtain the following result.

Theorem 5.1 (First-order convergence for mixed logit). *Assume that* ***A.0, A.6, A.1ml*** *and* ***A.3ml*** *hold. Then θ^* is a first-order critical point of problem* (2.7) *almost surely.*

Therefore we have proved convergence of the simulated estimators to the true maximum likelihood estimators almost surely, and allows the inclusion of convex constraints on θ. Classical results (see Chapter 10 of Train [38]) shows convergence in distribution and in probability asymptotically when the population size increases. The asymptotic behaviour is briefly discussed in section 5.2.1.

The extension of Theorem 4.2 establishing second-order convergence to the mixed-logit problem is immediate, as well as Theorem 4.1, as soon as the corresponding assumptions are fulfilled.

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 been assessed by providing a (hopefully high) probability that it is within some confidence interval asymptotically centered at zero and of radius ϵ. In practice, we first fix some probability level $\alpha > 0$ and determine the value of ϵ such that, for given θ,

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

Developing this expression we have that $\left|LL(\theta) - SLL^R(\theta)\right|$ is smaller than ϵ if and only if

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

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

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

for a given θ (such as the solution of the SAA problem). Since the logarithm is continuously differentiable on $\mathbb{R}_0^+$ and since $E_P\left[L_{ij_i}\left(\gamma, \theta\right)^2\right]$ is finite, we can use the

Delta method (see for instance Borovkov [11], page 44, for the one-dimensional case or Rubinstein and Shapiro [33] for the multi-dimensional case, section 6.3) to conclude that

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

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

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

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

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

Let α_δ be the quantile of a $N(0,1)$ associated to some level of signification 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 ϵ is then given by

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

Typically, one chooses $\alpha_{0.9} \approx 1.64$ of $\alpha_{0.95} \approx 1.96$. In practice we evaluate this accuracy ϵ_δ by taking the SAA estimators $\sigma^R_{ij_i}(\theta)$ and $P^R_{ij_i}(\theta)$.

Equation (5.3) gives us important information on the quality of the approximation. The accurary 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

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

If the total population is assumed to be infinite, then we may consider a population of size I as a independent and identically distributed sample within it. We then obtain from the strong law of large numbers that, almost surely,

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

In other terms, for a fixed sampling size R and a fixed θ, if the expectation of individual errors is finite, ϵ_δ converges to some real value which is less than this expectation. Assume indeed that there exists some κ such that for all θ in S, and for each individual i, $i = 1, \ldots, I$,

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

Then, from (5.3),

$$\epsilon_\delta \leq \alpha_\delta \frac{\kappa}{\sqrt{IR}}.$$

This suggests that the error decreases as the population size increases. However, we must remember that $E\left[SLL^R(\theta)\right] \neq LL(\theta)$, because of the logarithm operator, and our confidence interval is thus centered 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 $\ln P^R_{ij_i}$ around the true value P_{ij_i}, for some individual i:

$$\ln P^R_{ij_i}(\theta) = \ln P_{ij_i}(\theta) + \frac{1}{P_{ij_i}(\theta)} h_{ij_i} - \frac{1}{2\left(P_{ij_i}(\theta)\right)^2} h^2_{ij_i} + O\left(h^3_{ij_i}\right),$$

where $h_{ij_i} = P^R_{ij_i}(\theta) - P_{ij_i}(\theta)$. Therefore, since $E\left[h_{ij_i}\right] = 0$,

$$E\left[\ln P^R_{ij_i}(\theta)\right] - \ln P_{ij_i}(\theta) = -\frac{1}{2\left(P_{ij_i}(\theta)\right)^2} E\left[h^2_{ij_i}\right] + E\left[O\left(h^3_{ij_i}\right)\right].$$

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

$$E\left[h^2_{ij_i}\right] = \frac{1}{R}\sigma^2_{ij_i}(\theta).$$

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

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

which can be easily computed from the estimated error as

$$B = -\frac{I}{2\alpha^2_\delta}\epsilon^2_\delta. \tag{5.5}$$

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

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

It is interesting to note from (5.3) that the confidence interval radius ϵ 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 [21].

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

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

Therefore the maximization 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.

Of course, our analysis is restricted to the bias and error on the log-likelihood values, but give so far no indication on the bias and error on the estimated parameters θ. All hope of obtaining such estimates is not lost however, since we anticipate the log-likelihood function to be locally concave near its maximizer, which would provide some metric relating variations of the function to variations in the parameters.

5.2.1. Asymptotic behaviour We finally devote a last paragraph to extending the results obtained by Hajivassiliou and McFadden [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 everywhere, instead of in distribution.

¿From the strong law of large numbers,

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

Therefore the problem is consistent if and only if R tends to infinity when N tends to infinity, as reported by Hajivassiliou and McFadden [20] and Train [38], page 257. Taking the Taylor expansion around the true maximum likelihood estimators of $L(\theta) = E_I\left[\ln P_{ij_i}(\theta)\right]$, 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 $\sqrt{N}$, the SAA problem is consistent but not asymptotically normal;
- if R rises faster than $\sqrt{N}$, 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 being directly computable.

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

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

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

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

we obtain that

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

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 the sequence of SAA solutions for I tending to infinity, and R tending to infinity sufficiently fast compared to I. Let θ^* be a limit point of this sequence and assume (wlog) that $\{\theta^*_{I,R}\}$ converges to θ^*. Then, under the assumptions of the previous proposition, we obtain from Lemma 3.1 that

$$SLL^R\left(\theta^*_{I,R}\right) \xrightarrow{a.s.} L(\theta^*).$$

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, θ^* is a first (second)-order critical point 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 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 provides 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 minimization. Another point of interest is a better understanding of 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 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. Our thanks go also to Marcel Remon for his helpful comments on statistical theory and Stéphane Hess for his remarks on mixed logit theory. 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, 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. Technical Report 03/10, Department of Mathematics, University of Namur, Namur, Belgium, 2003.
4. Fabian Bastin, Cinzia Cirillo, and Philippe L. Toint. A trust-region Monte-Carlo algorithm for mixed logit estimation. Technical report, Department of Mathematics, University of Namur, Namur, Belgium, Forthcoming.
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. Kopelman. 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*, Cambridge, UK, 2002.
14. James Davidson. *Stochastic Limit Theory*. Oxford University Press, Oxford, England, 1994, 1994.
15. István Deák. Multidimensional integration and stochastic programming. In Yuri Ermoliev and Roger 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. Gül Gürkan, A. Yonca Özge, and Stephen M. Robinson. Sample-path optimization in simulation. In S. Manivannan J.D. Tew, D.A. Sadowski, and A.F. Seila, editors, *Proceedings of the 1994 Winter Simulation Conference*, pages 247–254, Piscataway, New Jersey, USA, 1994. IEEE.
20. Vassilis A. Hajivassiliou and Daniel L. McFadden. The method of simulated scores for the estimation of LDV models. *Econometrica*, 66(4):863–896, 1998.
21. David A. Hensher and William H. Greene. The mixed logit model: The state of practice. *Transportation*, 30(2):133–176, 2003.
22. 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.
23. Stéphane Hess and John Polak. Mixed logit estimation of parking type choice. *Submitted to Transportation Research Board Annual Meeting*, 2004.
24. Stéphane Hess, John Polak, and Andrew Daly. On the performance of shuffled Halton sequences in the estimation of discrete choice models. *Submitted to the Proceeding of the European Transport Conference*, 2003.
25. Stéphane Hess, Kenneth Train, and John Polak. On the use of randomly shifted uniform vectors in the estimation of a mixed logit model for vehicle choice. *Transportation Research B*, Submitted.
26. Peter Kall and Stein W. Wallace. *Stochastic Programming*. John Wiley & Sons, 1994.
27. 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.
28. Claude Montmarquette, Kathy Cannings, and Sophie Mahseredjian. How do young people choose college majors? *Economics of Education Review*, 21(6):543–556, 2002.
29. Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, USA, 1999.
30. Juan de Dios Ortúzar and Luis G. Willumsen. *Modelling Transport*. John Wiley & Sons, 3rd edition, 2001.
31. Stephen M. Robinson. Local epi-continuity and local optimization. *Mathematical Programming*, 37(2):208–222, 1987.
32. R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.
33. Reuven Y. Rubinstein and Alexander Shapiro. *Discrete Event Systems*. John Wiley & Sons, Chichester, England, 1993.
34. 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.
35. Alexander Shapiro. Stochastic programming by Monte Carlo simulation methods. *SPEPS*, 2000.
36. Yosef Sheffi. *Urban Transportation Networks*. Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1985.
37. Kenneth Train. Halton sequences for mixed logit. Working paper No. E00-278, Department of Economics, University of California, Berkeley, 1999.
38. Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, New York, USA, 2003.