Stochastic Optimization
Sample Average Approximation

Fabian Bastin
fabian.bastin@cirrelt.ca
Université de Montréal – CIRRELT

Framework

Consider the stochastic program of the form

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

where $z \in \mathbb{R}^m$ is a decision vector, S is a compact subset of $\mathbb{R}^m$ representing the feasible solutions of the previous problem, ξ is a random vector defined on $(\Xi, \mathcal{F}, P)$ and taking values in $(\mathbb{R}^k, \mathcal{B}^k)$ ($\mathcal{B}^k$ is the Borel measure). $G : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}$ is a real-valued function, and $E_P[\cdot]$ is the expectation with respect to P.

We assume that for each $z \in S$, the expected value function $g(z)$ is well defined, i.e. the function $G(z, \cdot)$ is $\mathcal{F}$-measurable and P-integrable.

For simplicity, we assume for now that S is deterministic.

Framework (cont'd)

If the cumulative distribution function of ξ is continuous or discrete with a large number of possible realizations, $g(z)$ is usually very difficult to evaluate.

Solving the previous problem is therefore very difficults and we have to use some approximations as the Monte Carlo methods: the original problem is replaced by successive approximations obtained by drawing $\xi_1, \ldots, \xi_N$. The approximation for a sample of size N is

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

We sometimes refer to the original program as the true program, or expected value program, and the approximate program as the sample average approximation (SAA).

Two-stage stochastic programming

This framework include the two-stage programs that we have analyzed, as

$$\min_x \; c^T x + E[\mathcal{Q}(x)]$$
$$\text{s.t. } Ax = b, \; x \geq 0$$

is equivalent to

$$\min_{x \in S} E[c^T x + \mathcal{Q}(x)],$$

where $S = \{x \,|\, Ax = b, x \geq 0\}$

First-order convergence

Investigate the convergence of solutions and optimal values of the sequence of SAA problems towards a solution and optimal value of the expected value program for $N \to \infty$.

Let z_N^* be a first-order critical point for the approximate problem, i.e. the KKT conditions are satisfied at z_N^*. In order to underline the dependency of z_N^* with respect to the successive draws $\xi_1, \ldots, \xi_N$, we will often use the notation $z_N^*(\xi_1, \ldots, \xi_N)$, or $z_N^*\left(\overline{\xi}\right)$, as $(\xi_1, \ldots, \xi_N)$ can be seen as the finite truncature of an infinite sequence $\overline{\xi} := \{\xi_k\}_{k=1}^{\infty}$.

As S is a compact set, the sequence of SAA solutions has a non-empty set of (finite) limit points.

Under which conditions such a limit point is a first-order critical point for the true problem?

Formalization

As the set of limit points depends on the realizations sequence $\overline{\xi}$, that is not known in advance, we have to introduce an adapted probability space over which we can define random variables whose realization 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 (i.i.d.).

Probability space of infinite dimension

From the i.i.d. property and the Kolmogorov consistency theorem (see for instance Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, 1967, Chapiter V, Theorem 5.1), we can build the probability space of infinite dimension

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

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 $B_i \in \mathcal{F}$, $i = 1, \ldots, j$.

Probability space of infinite dimension (cont'd)

In other terms, the marginal measures defined on $\prod_{i=1}^{j}(\Xi, \mathcal{F})$, with finite j ($j = 1, \ldots$), correspond to the product measures $\prod_{i=1}^{j} P$, as expected.

We therefore can see $\overline{\xi}$ as a random variable on $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, whose realizations are processes formed by the successive draws ξ_k, $k = 1, \ldots, \infty$, i.e. $\overline{\xi} = \{\xi_k\}_{k=1}^{\infty}$.

Notations:

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

Probability notions: types of convergence

Definition (Almost sure convergence (or with probability one))

A sequence of random variables $\{\xi_n\}_{n=1}^{+\infty}$ *converges almost surely to a random variable* ξ *if*

$$P_{\xi}[\{\omega | \, \xi_n(\omega) \to \xi(\omega) \text{ as } n \to \infty\}] = 1.$$

A weaker convergence mode is defined below.

Definition (Convergence in probability)

The sequence of random variables $\{\xi_n\}_{n=1}^{+\infty}$ *converges in probability to the random variable* ξ *if, for all* $\epsilon > 0$,

$$\lim_{n\to\infty} P_{\xi}[\{\omega \, | \, |\xi_n(\omega) - \xi(\omega)| \geq \epsilon\}] = 0.$$

Probability notions: types of convergence (cont'd)

Almost sure convergence implies converge in probability; the converse is not true.

Definition (Convergence in distribution)

Given a sequence of random variables $\{\xi_n\}_{n=1}^{\infty}$ *and a random variable* ξ, ξ_n *converges in distribution to* ξ *if* ξ_n *if*

$$\lim_{n\to\infty} F_n(x) \to F(x)$$

for each $x \in \mathbb{R}$ *where* F *is continuous.* F_n *and* F *are the distribution functions of* ξ_n *and* ξ, *respectively.*

We will also note $\xi_n \Rightarrow \xi$.

If both ξ and ζ follow the distribution specified by F, then $\xi_n \Rightarrow \xi$ and $\xi_n \Rightarrow \zeta$ are equivalent statements.

Notations

In what follows, and except explicitly stated, we will assume that the expressions *almost everywhere* and *almost surely* refer to the infinite dimension space $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, allowing to consider sets of realizations of the form $\{\xi_n\}_{n=1}^{\infty}$.

(In other words, the related results hold for almost every sampling process.)

A reference to another probability space will be denoted by prefixing the terms *almost surely* and *almost surely* by the measure defined on this probability space, and the expression *almost all sampling process* will be prefixed by the probability measure associated to the probability space of each process element.

Assumptions

A.0 The random draws $\{\xi_k\}_{k=1}^{\infty}$ are independent and identically distributed.

A.1 For P-almost all 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 each ξ.

A.1 obviously implies that $G(\cdot, \xi)$ is P-almost surely continuous. This last property and **A.2**, are typical assumptions in stochastic programming (see for instance Rubinstein and Shapiro, *Discrete Event Systems*, John Wiley & Sons, 1993). The stronger form **A.1** is justified by our interest in the first-order optimality conditions, which are expressed in term of the gradient of the objective.

Uniform law of large numbers

A.0–A.2 together imply that there exists an uniform law of large numbers (ULLN) on S, for the approximation $\hat{g}_n(z)$ of $g(z)$:

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

This also implies that $g(z)$ is continuous on S.

Reminder: the law of large numbers concerns the convergence of an empirical mean to the expectation of the random variable under consideration. The convergence can be in probability (weak law) or almost surely (strong law).

Our motivation to study the first-order conditions leads us to introduce an additional assumption on the gradient.

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

Consequences

This new assumption allows us to deduce that the expected value function $g(z)$ is continuously differentiable on S, and that the gradient and expectation operators can be interchanged, leading to

$$\nabla_z g(z) = E_P \left[\nabla_z G(z, \xi) \right].$$

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

The first-order convergence can be derived from the stochastic variational inequalities, as presented in A. 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.

Stochastic variational inequalities

Consider a mapping $\Phi : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^m$ and a multi-function $\Gamma : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$, and suppose that $\phi(z) := E_P[\Phi(z, \xi)]$ is well defined. We denote

$$\phi(z) \in \Gamma(z)$$

as the true generalized equation, or in expected value. $z^* \in \mathbb{R}^m$ is a solution of the generalized equation if $\phi(z^*) \in \Gamma(z^*)$.

If $\{\xi_1, \ldots, \xi_n\}$ is a random sample, we denote

$$\hat{\phi}_n(z) \in \Gamma(z)$$

as the generalized SAA equation, where $\hat{\phi}_n(z) = n^{-1} \sum_{i=1}^n \Phi(z, \xi_i)$. We denote by S^* and S_n^* the set of (all) solutions of the generalized equations in expected value and from the SAA, respectively.

First-order convergence

Denote by $d(x, A) := \inf_{x' \in A} \|x - x'\|$, the distance from $x \in \mathbb{R}^m$ to A, and $D(A, B) := \sup_{x \in A} D(x, B)$, the deviation of the set A from the set B. We have the following result (Shapiro, 2003).

Theorem (Generalized equations)

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

(a) *the multi-function $\Gamma(z)$ is closed, i.e. if $z_k \to z$, $y_k \in \Gamma(z_k)$ and $y_k \to y$, then $y \in \Gamma(z)$,*

(b) *the mapping $\phi(z)$ is continuous on $\mathcal{S}$,*

(c) *almost surely, $\emptyset \neq S_n^* \subseteq \mathcal{S}$ for n large enough, and*

(d) *$\hat{\phi}_n(z)$ converges to $\phi(z)$ almost surely uniformly of $\mathcal{S}$ as $n \to \infty$.*

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

First-order optimality conditions

We denote the feasible set of a mathematical program by $\mathcal{A}$. We will suppose here that $\mathcal{A}$ is closed.

Necessary condition: if x^* is a local minimizer,

$$-\nabla_x f(x^*) \in \mathcal{N}_{\mathcal{A}}(x^*),$$

where $\mathcal{N}_{\mathcal{A}}(x^*)$ is the normal cone to $\mathcal{A}$ at x^*.

Cone

A subset C of a vectorial space V is a (linear) cone iff αx belongs to C for any $x \in C$ and any strictly positive scalar α. It is pointed if α can be equal to zero.

A convex cone is a cone closed under convex combinations, i.e. iff $\alpha x + \beta y \in C$, $\forall \alpha,\ \beta > 0$, with $\alpha + \beta = 1$.

Tangent cone, normal cone: general framework

We first say that a vector $w \in \mathbb{R}^n$ is tangent to $\mathcal{A}$ at $x \in \mathcal{A}$ if for all sequences of vectors $\{x_i\}$ with $x_i \to x$, and $x_i \in \mathcal{A}$, and all sequences of positive scalars $t_i \downarrow 0$, there exists a sequence $w_i \to w$ such that $x_i + t_i w_i \in \mathcal{A}$ for all i.

The tangent cone $T_{\mathcal{A}}(x)$ is the collection of vectors tangent to $\mathcal{A}$ at x.
The normal cone $N_{\mathcal{A}}(x)$ is the orthogonal complement, i.e.

$$N_{\mathcal{A}}(x) = \{v \mid v^T w \leq 0, \ \forall w \in T_{\mathcal{A}}(x)\}.$$

OK, but in practice???

Tangent cone, normal cone (cont'd)

If $\mathcal{A}$ is convex, $\mathcal{N}_{\mathcal{A}}(x) = \left\{ v \mid v^T(x - x_0) \geq 0,\ \forall x_0 \in \mathcal{A} \right\}$.

The condition becomes $\nabla_x f(\hat{x})(x - \hat{x}) \geq 0, \forall x \in \mathcal{A}$.

If f is convex (i.e. we are in the convex programming framework), the condition is also sufficient.

Deterministic and convex constraints

Therefore, when S is convex, we can rewrite the first-order criticality conditions at some point z^* as the requirement that $-\nabla_z g(z^*)$ belongs to the normal cone to S at z^*, denoted by $\mathcal{N}_S(z^*)$.

If moreover S is deterministic, the feasible sets are the same for the true and SAA problems.

The previous theorem allows us to easily establish the first-order convergence. Consider the choice $\Gamma(\cdot) = \mathcal{N}_S(\cdot)$; $\phi(z^*)$ belongs to $\Gamma(z^*)$ iff

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

We design such variational inequalities as stochastic variational inequalities.

Convergence

The assumption (a) of the theorem always holds in this case. Take $\Phi(z, \xi) = -\nabla_z G(z, \xi)$ and represent by S^* and S_n^* the set of first-order critical points of the true and SAA generalized equations, respectively.

- The assumption (a) holds by definition of the normal cone.
- 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, giving assumption (b).
- Assumption (d) follows from the ULLN, while **A.1** and the compacity of S ensure assumption (c) by setting $\mathcal{S} = S$.

The previous theorem ensures therefore the first-order criticality at the limit as $n \to \infty$, almost surely.

Remark

The theorem establishes that $D(S_n^*, S^*) \to 0$ when $n \to \infty$, almost surely. However, we cannot write $D(S^*, S_n^*) \to 0$ when $n \to \infty$, almost surely.

Consider for instance the optimization problem

$$\min_{x \in S} |E[\xi]| x^2,$$

where $S = [-a, a]$, $a > 0$. If $\xi \sim N(0, \sigma^2)$, $S^* = S$. The SAA approximation is

$$\min_{x \in S} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right| x^2,$$

and almost surely, $S_n^* = \{0\}$. Therefore, a.s. $\lim_{n \to \infty} S_n^* = \{0\} \neq S^*$, and $D(S^*, \{0\}) = a$.

Stochastic constraints

Under stronger assumptions, it is also possible to prove the first-order convergence almost surely when S is non-convex or non-deterministic.

We now assume that the feasible set can be described by equality and inequality constraints. The original problem now becomes

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

where V is a compact subset of $\mathbb{R}^m$.

Stochastic constraints (cont'd)

The corresponding SAA problem is defined as

$$\begin{aligned} &\min_{z \in V} \hat{g}_n(z), \\ \text{such that } &\hat{c}_{jn}(z) \geq 0, \ j = 1, \ldots, k, \\ &\hat{c}_{jn}(z) = 0, \ j = k+1, \ldots, M. \end{aligned}$$

Here, for each $j = 1, \ldots, M$, $\{\hat{c}_{jn}(\cdot)\}$ is a sequence of (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 form of expectations:

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

Stochastic constraints (cont'd)

These functions can then be approximated by corresponding sample average approximations:

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

For simplicity, we consider the more general parametric mathematical problem

$$\begin{aligned} &\min_{z \in V} \hat{g}(z, \epsilon), \\ \text{such that } &\hat{c}_j(z, \epsilon) \geq 0, \ j = 1, \ldots, k, \\ &\hat{c}_j(z, \epsilon) = 0, \ j = k+1, \ldots, M, \end{aligned}$$

where ϵ is a parametric random vector giving the perturbations of the approximated program, and $g(\cdot)$, $\hat{g}(\cdot, \epsilon)$, $c_j(\cdot)$, $\hat{c}_j(\cdot, \epsilon)$ are assumed to be of class C^2 with respect to z.

Random perturbations

We assume that the perturbation has the form

$$\epsilon = \epsilon(z, \bar{\xi}) = \left(\epsilon_g,\ \epsilon_{c_1},\ \ldots,\ \epsilon_{c_M},\ \epsilon_{\nabla g}^T,\ \epsilon_{\nabla c_1}^T,\ \ldots,\ \epsilon_{\nabla c_M}^T\right)^T,$$

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

$$\hat{g}(z, \epsilon) = g(z) + \epsilon_g,\ \nabla_z \hat{g}(z, \epsilon) = \nabla_z g(z) + \epsilon_{\nabla g},$$
$$\hat{c}_j(z, \epsilon) = c_j(z) + \epsilon_{c_j},\ \nabla_z \hat{c}_j(z, \epsilon) = \nabla_z c_j(z) + \epsilon_{\nabla c_j},\ j = 1, \ldots, M.$$

We also define $\epsilon_n(z, \bar{\xi})$ as

$$\epsilon_n(z, \bar{\xi}) = \begin{pmatrix} \hat{g}_n(z) - g(z) \\ \hat{c}_{jn}(z) - c_j(z),\ j = 1, \ldots, M \\ \nabla_z \hat{g}_n(z) - \nabla_z g(z) \\ \nabla_z \hat{c}_{jn}(z) - \nabla_z c_j(z),\ j = 1, \ldots, M \end{pmatrix},$$

Random perturbations (cont'd)

We denote the corresponding random vector by $\epsilon_n(z, \overline{\xi})$, and we will assume the ULLN holds for the objective and the constraints, as well as for their corresponding derivatives.

We finally assume that the feasible set for the original and approximated problems are non-empty. The Lagrangians of the true and approximated problems are respectively

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

and

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

KKT conditions

Let $z^*(\epsilon)$ be a first-order critical point of the approximated program, and assume that $z^*(\epsilon)$ belongs to the interior of V, that we denote by $\overset{o}{V}$.

There exist Lagrange multipliers $\lambda^*(\epsilon)$ such that $(z^*(\epsilon), \lambda^*(\epsilon))$ satisfies the KKT conditions; in other terms $(z^*(\epsilon), \lambda^*(\epsilon))$ is solution to the system

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

Solution characteristics

Consider now the specific sampling process $\overline{\xi}$. In order to emphasize the dependency of the first-order critical points to the samplibg process, we denote $z_n^*(\overline{\xi})$ for $z^*(\epsilon_n)$ and $\lambda_n^*(\overline{\xi})$ for $\lambda^*(\epsilon_n)$.

Let $\mathcal{Z}\left(\{z_n^*(\overline{\xi})\}\right)$ the set of accumulation points of the sequence $\{z_n^*(\overline{\xi})\}_{n=1}^{\infty}$. The compacity of V implies that for each $\overline{\xi}$, $\mathcal{Z}\left(\{z_n^*(\overline{\xi})\}\right)$ is non-empty.

We can now prove the first-order convergence for the general case.

First-order convergence

Theorem

Assume that for almost every $\overline{\zeta}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, $\epsilon_n(z, \overline{\zeta}) \to 0$ uniformly on the compact set V, as $n \to \infty$, and let $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$ satisfying the assumption of uniform convergence.

If $z^* \in \mathcal{Z}\left(\{z_n^*(\overline{\xi})\}\right)$ belongs to $\overset{o}{V}$, and a sub-sequence $\{z_\ell^*(\overline{\xi})\}_{\ell=1}^{\infty} \subseteq \{z_n^*(\overline{\xi})\}_{n=1}^{\infty}$ converging to z^* is associated to a sequence of Lagrange multipliers $\{\lambda_\ell^*(\overline{\xi})\}_{\ell=1}^{\infty}$ with a least one limit point, then z^* is a first-order critical point of the true problem.

Proof.

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

On the Lagrange multipliers

The assumption of Lagrange multipliers convergence always holds if the Lagrange multipliers are bounded. This stronger assumption allows us to use the convergence results on the variational inequalities in order to prove the first-order criticality of limit points (Shapiro, 2003).

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

$$\phi(\mu) = (\nabla_z \mathcal{L}(z, \lambda), c_{k+1}(z), \ldots, c_M(z)),$$
$$\hat{\phi}_n(\mu) = (\nabla_z L(z, \lambda, \epsilon_n), \hat{c}_{k+1}(z, \epsilon_n), \ldots, \hat{c}_M(z, \epsilon_n)).$$

The generalized equation $\phi(\mu) \in \mathcal{N}_{\mathcal{K}}(\mu)$ then represents the KKT optimality conditions for the original optimization problem, $S^* \subseteq \overset{o}{V}$, and the theorem on generalized equations then implies first-order criticality almost surely, with $\Gamma(\mu) := \mathcal{N}_{\mathcal{K}}(\mu)$.

On the Lagrange multipliers (cont'd)

Assumptions (a) and (d) of the theorem on generalized equations are sufficient as $\epsilon \to 0$ almost surely, and assumption (c) holds as feasible set for original and approximate problems are non-empty.

Trickier: second-order.

Second-order convergence

If we strengthen our assumptions, we can also show that, almost surely, there exists limit points in $\mathcal{Z}\left(\{z_n^*(\overline{\boldsymbol{\xi}})\}\right)$ that are local minimizers.

We first consider the case where S is deterministic and assume that, for a given sampling process $\overline{\xi}$, $z_n^*(\overline{\xi})$ is a local minimizer of $\hat{g}_n(z)$ over S. In other terms,

$$\exists\, \delta_n(\overline{\xi}) \text{ t.q. } \forall z \in B\left(z_n^*(\overline{\xi}), \delta_n(\overline{\xi})\right) \cap S,\ \hat{g}_n\left(z_n^*(\overline{\xi})\right) \leq \hat{g}_n(z),$$

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

A counter-example

Consider the problem

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

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

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

Problem (2) has two (isolated) local minimizers

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

when $\sum_{i=1}^n \xi_i > 0$. As $n \to \infty$, we have $P\left[\sum_{i=1}^n \xi_i > 0\right] \to 0.5$, but from the ULLN, $\frac{1}{6n}\sum_{i=1}^n \xi_i \to \frac{1}{6}E_P[\xi] = 0$ a.s.

Rigidity assumption

(Bastin, Cirillo, and Toint, 2006, Pasupathy, 2010)
In order to show that $z^* \in \mathcal{Z}\left(\{z_n^*(\overline{\xi})\}\right)$ is a local minimizer of $g(\cdot)$ over S, we have to show that for some subsequence $\{z_\ell^*(\overline{\xi})\}_{\ell=1}^{\infty} \subseteq \{z_n^*(\overline{\xi})\}_{n=1}^{\infty}$ converging to z^*, the neighborhood where $z_\ell^*(\overline{\xi})$ is a local minimizer do not reduce to a singleton as $\ell \to \infty$.

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

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

Basic theorem

Theorem

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

Proof.

Voir Bastin, Cirillo, and Toint, *Convergence theory for nonconvex stochastic programming with an application to mixed logit*, Mathematical Programming, 108(2–3):207–234 (2006). □

Stochastic constraints

Assumption **A.4** is somewhat artificial and it is therefore more interesting to look for more elegant conditions.

At the difference of classical sensitivity analysis, we focus here on conditions under which a limit point of a sequence of approximate solutions is a solution of the true problem.

As previously, we consider the case where the feasible set is described by a set of equality and inequality constraints, and we assume that ϵ_n converges to 0 almost surely, uniformly on V.

Lagrange multipliers convergence

Consider a particular sampling process $\overline{\xi}$ in $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, and $z^* \in \mathcal{Z}\left(\{z_n^*(\overline{\xi})\}\right)$. Under some conditions, if the subsequence $\{z_\ell^*\}_{\ell=1}^\infty$ converges to z^*, the sequence of associated Lagrange multipliers vectors $\{\lambda_\ell^*\}$ also converges to some Lagrange multipliers vector λ^* associated to z^* for the true problem.

Lemma

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 $n \to \infty$. If $z^* \in \mathcal{Z}\left(\{z_n^*(\overline{\xi})\}\right) \cap \overset{o}{V}$, $\{z_\ell^*\}_{\ell=1}^\infty \subseteq \{z_n^*\}_{n=1}^\infty \cap \overset{o}{V}$ converges to z^*, and there exists some unique Lagrange multipliers vector associated to z^* satisfying the KKT conditions, then $\lambda_\ell^*(\overline{\xi})$ converges to λ^* as $\ell \to \infty$.

Lagrange multipliers convergence (cont'd)

Proof.
From the unicity of λ^*, the Mangasarian-Fromowitz constraint qualification (MFCQ) holds at z^*, and therefore in a neighborhood of z^*. A consequence is that the Lagrange multipliers are uniformly bounded for ϵ close to zero. It is therefore sufficient to show that every limit point of the sequence $\{\lambda_\ell^*(\overline{\xi})\}$, $\ell = 1, \ldots, \infty$, is equal to λ^*.

Let λ' be such a limit point. By continuity, (z^*, λ') satisfies the KKT conditions, so λ' is equal to λ^*. □

The unicity of λ^* can be ensured with some appropriate constraints qualification condition, as the linear independence constraint qualification (LICQ).

On the LICQ

This constraints qualification will be especially useful for our discusssion.

Consider the original program. Recall that the active set $\mathcal{A}(z)$ at any feasible point z is the union of the index set of equality constraints and active inequality constraints:

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

Definition

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

Strict complementarity

Another useful concept for our needs is the strict complementarity condition.

Definition

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

Consider again a particular sampling process $\overline{\xi}$ dans $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, such that $\epsilon_n(z, \overline{\xi}) \to 0$ as $n \to \infty$. If the assumptions of the above lemma hold for some subsequence $\{z_\ell^*\}_{\ell=1}^\infty \to z^*$, the gradient of the Lagrangian $\nabla L\left(z_\ell^*(\overline{\xi}), \lambda_\ell^*(\overline{\xi})\right)$ converges to $\nabla \mathcal{L}(z^*, \lambda^*)$, as ℓ tends to infinity, with $\lambda_\ell^*(\overline{\xi}) \to \lambda^*$.

Strict complementarity (cont'd)

Assume that the strict complementarity condition holds at z^* for the approximate problem. We obtain that for ℓ large enough, $[\lambda_\ell^*]_j(\overline{\xi})$, $j \in \{1, \ldots, k\} \cap \mathcal{A}(z^*)$, are strictly positive and thus the corresponding constraints are active at $z_\ell^*(\overline{\xi})$.

Moreover, as $\epsilon_n(z, \overline{\xi}) \to 0$, $\hat{c}_j\left(z_\ell^*(\overline{\xi}), \epsilon_\ell(z_\ell^*(\overline{\xi}), \overline{\xi})\right) \to c_j(z^*)$ and, for ℓ large enough, $\mathcal{A}\left(z_\ell^*(\overline{\xi})\right) = \mathcal{A}(z^*)$, so the strict complementarity condition holds at $\left(z_\ell^*(\overline{\xi}), \lambda_\ell^*(\overline{\xi})\right)$ for the approximate problem.

This allows us to state a second-order convergence result.

A second-order convergence result

Theorem (Second-order convergence)

Assume that, for almost every sampling process $\overline{\xi}$ *in* $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$, *there exists some* $z^* \in \mathcal{Z}\left(\{z_n^*(\overline{\xi})\}\right) \cap \overset{o}{V}$ *associated to an unique vector of Lagrange multipliers* λ^* *and some subsequence* $\{z_\ell^*\}_{\ell=1}^{\infty} \subseteq \{z_n^*\}_{n=1}^{\infty}$, *such that* $z_\ell^*(\overline{\xi}) \to z^*$, *and*

(a) $\epsilon_n\left(z_n^*(\overline{\xi}), \overline{\xi}\right) \to 0$ *uniformly on* V, *as* $n \to \infty$,

(b) $z_n^*(\overline{\xi}) \in \overset{o}{V}$, $n = 1, \ldots,$

(c) $\nabla^2_{zz}\hat{g}\left(z_\ell^*(\overline{\xi}), \epsilon_\ell(z_\ell^*(\overline{\xi}), \overline{\xi})\right) \to \nabla^2_{zz} g(z^*)$ *as* $\ell \to \infty$,

(d) $\nabla^2_{zz}\hat{c}_j\left(z_\ell^*(\overline{\xi}), \epsilon_\ell(z_\ell^*(\overline{\xi}), \overline{\xi})\right) \to \nabla^2_{zz} c_j(z^*)$ $(j = 1, \ldots, M)$ *as* $\ell \to \infty$.

We also assume that the strict complementarity condition and the LICQ hold at (z^*, λ^*) *for the true problem.*

A second-order convergence result (cont'd)

Theorem (Second-order convergence (cont'd))

Then, for amost every sampling process $\bar{\xi}$,

(i) *the LICQ holds at* $z_\ell^*(\bar{\xi})$, *for* ℓ *large enough,*

(ii) (z^*, λ^*) *satisfies the second-order necessary condition for the true program:*

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

If moreover $\exists$ *some constant* $\alpha_{\bar{\xi}} > 0$ *such that,* $\forall$ ℓ *large enough,*

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

then (z^*, λ^*) *almost surely satisfies the sufficient conditions for the true problem.*

A second-order convergence result (cont'd)

Theorem (Second-order convergence (cont'd))

In other terms

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

This is equivalents to say that z^* is an isolated minimizer for the original problem.

Proof.

See Bastin, Cirillo and Toint, 2006. □

Discussion

Remark that the LICQ and the strict complementarity condition imply that the minimizer is isolated while sufficient second-order condition is usually used to characterize strict local minimizers.

Reminder: an isolated local minimizer is also strict, but the converse is not necessarily true.

If $z^* \in \mathcal{Z}\left(\{z_n^*(\bar{\xi})\}\right) \cap \overset{o}{V}$ is a strict but not isolated local minimizer, every neighborhood of z^* contains other local minimizers that are candidates to be limit points of sequences of SAA solutions, as n grows to infinity, and z^* can therefore be difficult to identify.

Non-degeneracy

The non-degeneracy assumption

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

can also be replaced by requiring that the Jacobian of equality equations involved in the KKT conditions associated to the original program

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

is non-singular at (z^*, λ^*).

Second-order convergence: corollary (cont'd)

Corollary

Assume that, for almost every sampling process $\overline{\xi}$ *in* $(\Xi_\Pi, \mathcal{F}_\Pi, P_\Pi)$*, there exist some* $z^* \in \mathcal{Z}\left(\{z_n^*(\overline{\xi})\}\right) \cap \overset{o}{V}$ *associated to an unique Lagrange multipliers vector* λ^* *and some subsequence* $\{z_\ell^*\}_{\ell=1}^{\infty} \subseteq \{z_n^*\}_{n=1}^{\infty}$*, such that* $z_\ell^*(\overline{\xi}) \to z^*$*, that assumptions (a)–(d) of the second-order convergence theorem hold and that the strict complementarity condition hold at* (z^*, λ^*) *for the original program.*

Assume moreover that the Jacobian of the equality relations of the KKT conditions is not-singular at (z^*, λ^*)*. Then there exists almost surely some* z^* *in* $\mathcal{Z}\left(\{z_N^*(\overline{\xi})\}\right) \cap \overset{o}{V}$*, associated to some vector* λ^**, that satisfies the sufficient second-order conditions of the original program.*

Second-order convergence: corollary (cont'd)

Proof.
See Bastin, Cirillo, and Toint, 2006. □

The inverse implication of the second-order convergence theorem can be obtained from classical results in perturbations analysis (A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, USA (1983), Theorem 3.2.2). More developments if the context of stochastic programming can be found in Rubinstein and Shapiro (2003) and A. Shapiro, *Probabilistic constrained optimization: Methodology and applications*, in S. Uryasev, editor, *Statistical inference of stochastic optimization problems*, pages 282–304. Kluwer Academic Publishers, 2000.