ASYMPTOTIC THEORY FOR SOLUTIONS IN STATISTICAL ESTIMATION AND STOCHASTIC PROGRAMMING

Alan J. King† and R. Tyrrell Rockafellar‡

Abstract. New techniques of local sensitivity analysis for nonsmooth generalized equations are applied to the study of sequences of statistical estimates and empirical approximations to solutions of stochastic programs. Consistency is shown to follow from a certain local invertibility property, and asymptotic distributions are derived from a generalized implicit function theorem that characterizes asymptotic behavior in situations where estimates are subjected to constraints and estimation functionals are nonsmooth.

Keywords: stochastic programs, generalized equations, consistency, central limits, contingent derivatives.

---

† IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, U.S.A. Research supported in part by a grant from the National Science Foundation while with the International Institute for Applied Systems Analysis, Laxenburg, Austria.

‡ Department of Mathematics, University of Washington, Seattle, WA 98195, U.S.A. Research supported in part by the Air Force Office of Scientific Research under grant 89-0081 and the National Science Foundation under grant DMS-8819586.

1. Introduction

Many problems in statistics and stochastic programming may be formulated as the selection of an element $\mathbf{x}^\nu$ from the solution set $J(\mathbf{z}^\nu)$ to a *generalized equation*

$$\text{choose } x \in I\!R^n \text{ such that } 0 \in f(\mathbf{z}^\nu, x) + N(x), \tag{1.1}$$

where f is a function taking values in $I\!R^m$, N is a multifunction, a mapping whose images are subsets of $I\!R^m$, and $\mathbf{z}^\nu$ is a data-dependent random process. To analyze the large sample behavior of these estimated solutions, one goes over from the asymptotics of the data process to those of the estimated solutions via some sort of local analysis of the generalized equation. We are interested here in establishing conditions on f and N that when holding at a pair (z^*, x^*) ensure that if for some sequence of positive numbers $\{\tau_\nu\}$ converging to 0 one has

$$\tau_\nu^{-1}[\mathbf{z}^\nu - z^*] \underset{\mathcal{D}}{\longrightarrow} \mathbf{w}, \tag{1.2}$$

then it follows that also

$$\tau_\nu^{-1}[\mathbf{x}^\nu - x^*] \underset{\mathcal{D}}{\longrightarrow} DJ(z^*)(\mathbf{w}),$$

where DJ is a certain contingent derivative of J to be further discussed below and the symbol $\mathcal{D}$ below the arrow denotes convergence in distribution. The chief applications we have in mind for the generalized equation (1.1) in the present study are that of constrained statistical estimation and the closely related subject of estimation in stochastic programming.

In statistics, the generalized equation (1.1) can represent the so-called "normal equations" of maximum likelihood estimation, and the multifunction N may be designed to allow the imposition of certain types of "hard" constraints—such as nonnegativity of variance estimates. Concrete examples of constrained estimation problems in statistics and a discussion of the connections between stochastic programming and statistical estimation can be found in Dupačová and Wets [5].

In stochastic programming, equation (1.1) can represent the first-order necessary conditions for solutions to the optimization problems

$$\text{minimize } \overline{\mathbf{h}}^\nu(x) \text{ over all } x \in C \subset I\!R^n, \tag{1.3}$$

where we denote

$$\overline{\mathbf{h}}^\nu(x) = \frac{1}{\nu}\sum_{i=1}^{\nu} h(x, \mathbf{s}_i).$$

When h is continuously differentiable in the first variable, then, under appropriate regularity conditions, the solutions to (1.3) are determined by the *first-order necessary conditions*

$$0 \in \nabla\overline{\mathbf{h}}^{\nu}(x) + N_C(x), \tag{1.4}$$

where the multifunction N_C is the normal cone operator of nonsmooth analysis. The translation of (1.4) to (1.1) is accomplished by setting f to be the evaluation functional $f(z,x) = z(x)$ and viewing the random variable $\mathbf{z}^{\nu} = \nabla\overline{\mathbf{h}}^{\nu}$ as an element of the space of $\mathbb{R}^n$-valued continuous functions on $\mathbb{R}^n$, denoted $\mathcal{C}_n(\mathbb{R}^n)$. Denote also by $E\nabla\mathbf{h}$ the expected value of the gradient of h. The main steps in computing the asymptotic distribution for $\mathbf{x}^{\nu} - x^*$ are first to compute the asymptotic distribution for $\nabla\overline{\mathbf{h}}^{\nu} - E\nabla\mathbf{h}$ from a central limit theorem in $\mathcal{C}_n(U)$, for some appropriate compact neighborhood U of x^*, and then apply the sensitivity analysis of the solution set J of (1.1) about the "point" $z^* = E\nabla\mathbf{h}$.

Developing the asymptotics of solutions to (1.4) or (1.1) from the limit behavior of random elements of a *function space* is a powerful idea that can be extended in many directions. One may study the dependence of the optimal value on the estimated objective function—this dependence is differentiable in the objective over the space of continuous functions, leading to an asymptotic result for the optimal values (Shapiro [17]). A deeper study of this dependence leads to confidence bounds for solutions (Ermoliev and Norkin [6]) under very general assumptions. Or, as in the present study, one may analyze the dependence of the optimal solutions on the estimated gradient mappings and apply a version of an implicit function theorem. In the rather special case of (unconstrained) maximum likelihood estimation in statistics, this latter program has been thoroughly worked out in, for example, Ibragimov and Has'minskii [9]. But complications naturally arising in optimization problems and their associated generalized equations require extended definitions of the concepts of consistency and central limits: the non-smoothness of the estimated functional $f(z,\cdot)$ may mean that there are more than one cluster point for a sequence of generalized M-estimates, or the domain of the multifunction N may constrain the support of the asymptotic distribution in certain ways.

The key step that we make in this paper to accommodate such complications is the identification of the appropriate generalized differentiability and invertibility properties of the mapping

$$F(x) = f(z^*, x) + N(x)$$

that enable the determination of the behavior of the solutions to the perturbed generalized equation (1.1). Background material for this analysis appears in the two papers King [11], and King and Rockafellar [12]. The first generalizes the classical delta method to apply

to mappings, such as F, that are not Frechét differentiable or even single-valued, and the second studies generalized continuity and differentiability properties of solution functionals for perturbed nonsmooth generalized equations. In Section 2, we bring the perspective of these two papers to bear on our asymptotic analysis. Consistency will follow from a sort of local invertibility of F called *subinvertibility*, and the central limit will be given by the *contingent derivative* of F^{-1}, provided certain regularity conditions are satisfied. These results are applied in Section 3 to the asymptotic analysis for stochastic programs (of a more general form than (1.3)). Section 4 presents a brief study of a piecewise linear-quadratic tracking problem, in which there arises the possibility of a nonsmooth expected gradient $E\nabla\mathbf{h}$ to which standard implicit function theorems cannot be applied. Earlier versions of these results appeared in King [10].

The first study of asymptotic theory for constrained maximum likelihood estimates was by Aitchison and Silvey [1], who proved asymptotic normality under conditions of second-order sufficiency and linear independence for equality constraints only. Huber [8] established asymptotic behavior under "non-standard" assumptions that could be applied to the sort of nonsmooth mappings that arise in inequality-constrained optimization, but the domain of the estimates was assumed to be an open set. Recently, Dupačová and Wets [5], and especially, Shapiro [16] have applied Huber's theory to the problem of determining the central limit behavior of the solution estimates to stochastic programs: [5] gives conditions under which asymptotic normality may occur, and [16] gives conditions under which non-normal behavior may arise from deterministic inequality constraints. These papers rely on a certain smoothness of the expected gradient $E\nabla\mathbf{h}$ near x^*, an assumption that is helpful but not necessary in the theory we offer.

2. General Theory

The underlying topology on which our analysis is based is that of the convergence of closed sets in $\mathbb{R}^n$. For $\{A_\nu\}$ a sequence of closed subsets of $\mathbb{R}^n$, define the (closed) sets

$$\liminf_\nu A_\nu = \{x = \lim x_\nu \,\big|\, x_\nu \in A_\nu \text{ for all but finitely many } \nu\}$$

$$\limsup_\nu A_\nu = \{x = \lim x_\nu \,\big|\, x_\nu \in A_\nu \text{ for infinitely many } \nu\}.$$

The sequence $\{A_\nu\}$ *set-converges to* $A = \lim_\nu A_\nu$, if $A = \liminf A_\nu = \limsup A_\nu$. Let $(Z, \mathcal{A})$ be an arbitrary measurable space. A multifunction $F : Z \rightrightarrows \mathbb{R}^n$ is *closed-valued* (or convex, etc.) if F has closed (or convex, etc.) images. A closed-valued multifunction is *measurable* if it is Borel-measurable when considered as a map into the space of closed subsets topologized by set-convergence, or equivalently, if for all closed subsets C of $\mathbb{R}^n$

the set $F^{-1}(C) := \{z \in Z \,\big|\, F(z) \cap C \neq \emptyset\}$ belongs to the sigma-algebra $\mathcal{A}$. If the measurable space is a probability space, we shall sometimes refer to a closed-valued measurable multifunction F as a *random closed set* and denote it $\mathbf{F}$. The *domain* of a multifunction F, $\operatorname{dom} F$, is the set of points where its image is nonempty; its *graph* is the set of pairs $\operatorname{gph} F := \{(z, x) \in Z \times I\!R^n \,\big|\, x \in F(z)\}$. If Z is a topological space then we say that F is *closed* (or upper semicontinuous) if $\operatorname{gph} F$ is a closed subset of $Z \times I\!R^n$. It is well-known that a closed multifunction is closed-valued and measurable; the basic background on these topics is covered, for example, in Rockafellar [15].

Our first result shows that the solution mapping J for (1.1) has this property, and it is an easy corollary to show that this implies a certain form of consistency.

Proposition 2.1. *Suppose that the function $f : I\!R^n \times Z \to I\!R^m$ is jointly continuous, and that the multifunction $N : I\!R^n \rightrightarrows I\!R^m$ is closed. Then the solution mapping $J : Z \rightrightarrows I\!R^n$ defined by*

$$J(z) = \{x \in I\!R^n \,\big|\, 0 \in f(z, x) + N(x)\} \tag{2.1}$$

is closed-valued and measurable.

Proof. Consider a sequence of pairs $\{(z^\nu, x^\nu)\}$, each an element of $\operatorname{gph} J$, that converges to a pair (z^*, x^*). By continuity, $f(x^\nu, z^\nu)) \to f(x^*, z^*)$. Since N is closed, it follows that $-f(x^*, z^*) \in N(x^*)$. This implies $x^* \in J(z^*)$, so J is closed and therefore closed-valued and measurable. □

Corollary 2.2. *(Consistency.) Under the conditions of Theorem 2.1, if*

$$\mathbf{z}^\nu \to z^* \quad \text{a.s.}$$

and $\{\mathbf{x}^\nu\}$ is a sequence of solutions to (1.1) with a cluster point $\mathbf{x}$, then $\mathbf{x} \in J(z^*)$ with probability one.

Remark. The corollary can be strengthened if there are natural conditions that imply (or if one does not mind imposing conditions that require) that solutions of (1.1) belong to some compact set. In this case, almost all solution sequences will have cluster points.

It is not at all guaranteed at this stage of the game that there exist any solutions $\mathbf{x}^\nu$ to the generalized equation (1.1) as $\nu \to \infty$. To simplify the verification of the existence of such solutions, we introduced in [12] the following notion: a multifunction $F : I\!R^n \rightrightarrows I\!R^m$ is *subinvertible* at $(x^*, 0)$ if one has $0 \in F(x^*)$ and there exist a positive number ε, a compact convex neighborhood U of x^*, and a nonempty convex-valued mapping $G : \varepsilon B \rightrightarrows U$ such

that: $\operatorname{gph} G$ is closed, the point x^* belongs to $G(0)$, and $G(y) \subset F^{-1}(y)$ for all $y \in \varepsilon B$, where B is the unit ball in $\mathbb{R}^m$. Under this assumption, it can be shown that for all $y \in \varepsilon B$ there exists at least one solution $x(y)$ to the perturbed generalized equations

$$0 \in F(x) - y.$$

Natural conditions implying such subinvertibility may be found in Sections 5 and 6 of [12]. The reader may easily verify, for instance, that multifunctions whose inverse F^{-1} admits a selection $x(\cdot)$ that is continuous on a neighborhood of 0 are subinvertible at $(x(0), 0)$.

Theorem 2.3. *Suppose, in addition to the assumptions of Corollary 2.2, that the multifunction $F(\cdot) := f(\cdot, z^*) + N(\cdot)$ is subinvertible at $(x^*, 0)$, for some $x^* \in J(z^*)$, and that*

$$\mathbf{z}^\nu \to z^* \quad \text{a.s.}$$

Then there exists a compact convex neighborhood U of x^* such that, with probability one,

$$\emptyset \neq \limsup_{\nu \to \infty} J(\mathbf{z}^\nu) \cap U \subset J(z^*) \cap U.$$

Proof. Let U be the compact neighborhood of x^* in the definition of subinvertibility. In the event of the convergence $\mathbf{z}^\nu \to z^*$, the subinvertibility of F implies, by Proposition 3.1 of [12], that $U \cap J(\mathbf{z}^\nu)$ is eventually nonempty; this and the compactness of U prove that $\limsup J(\mathbf{z}^\nu) \cap U \neq \emptyset$. Since J is closed, by Proposition 2.1, it follows that $J \cap U$ is closed, from which we obtain the right-hand inclusion. □

We next consider the possible limiting behavior of solutions to (1.1). The key step is to apply an appropriately generalized implicit function theorem that follows from an analysis of certain generalized derivatives of the multifunction $F(\cdot) = f(z^*, \cdot) + N(\cdot)$, which we now briefly review. (For more details, consult [12] and the references therein.) The *contingent derivative* of a multivalued mapping $G : Z \rightrightarrows \mathbb{R}^n$ at a point $z \in \operatorname{dom} G$ and $x \in G(z)$ is the mapping $DG(z|x)$ whose graph is the *contingent cone* to the graph of G at $(z, x) \in Z \times \mathbb{R}^n$, that is,

$$\limsup_{t \downarrow 0} t^{-1}[\operatorname{gph} G - (z, x)] = \operatorname{gph} DG(z|x). \tag{2.2}$$

The contingent derivative always exists, because the lim sup of a net of sets always exists; and it is closed because the lim sup is always a closed set. The contingent derivative of the inverse of G is just the inverse of the contingent derivative, and is denoted $DG^{-1}(x|z)$. This definition may be specialized in two directions. If one has lim sup = lim inf in (2.2), then G

is called *proto-differentiable* at (z, x). A stronger property that is related to differentiability for functions is *semi-differentiability*, which requires the existence of the set limit

$$\lim_{\substack{t\downarrow 0\\ w'\to w}} t^{-1}[G(z+tw')-x] = DG(z|x)(w) \tag{2.3}$$

for all directions w in Z. These definitions can be applied to functions, of course. If $g : Z \to \mathbb{R}^n$ has a contingent derivative $Dg(z)$ that is everywhere single-valued, then g is *B-differentiable* at z and we can show that

$$\lim_{\substack{t\downarrow 0\\ w'\to w}} t^{-1}[g(z+tw')-g(z)] = Dg(z)(w) \tag{2.4}.$$

In the case of a function of two variables, as we have in (1.1), we shall employ the partial B-derivatives, $D_z f(z, x)$ and $D_x f(z, x)$. It will be necessary to suppose a certain uniformity of the partial B-derivative in z, namely that for every $\varepsilon > 0$ there exist neighborhoods Ω of z^* in Z and U of x^* in $\mathbb{R}^n$ such that for every $x \in U$ the function

$$z \mapsto f(z, x) - f(z^*, x) - D_z f(z^*, x^*)(z - z^*)$$

is Lipschitz continuous with Lipschitz constant ε on Ω. When this holds, we say that f has a *strong* partial B-derivative in z at (z^*, x^*).

An immediate application of the contingent derivative may be seen in the following theorem that establishes the existence of bounds in probability on the solution sequences.

Theorem 2.4. *Assume that the space Z is a separable Banach space, and that the function $f : Z \times \mathbb{R}^n \to \mathbb{R}^m$ satisfies a Lipschitz condition*

$$|f(z^*, x) - f(z, x)| \le \alpha \|z^* - z\|$$

uniformly for all x in a neighborhood of x^*. Define the multifunction $F(\cdot) = f(z^*, \cdot) + N(\cdot)$, and suppose that

$$DF^{-1}(0|x^*)(0) = \{0\},$$

i.e. the inverse of the contingent derivative of F contains at most the single element 0. Then there exist a neighboorhood U of x^* and a constant $\lambda \ge 0$ such that if $\mathbf{x}^\nu$ is a solution to (1.1) that belongs also to U one has

$$P\{|\mathbf{x}^\nu - x^*| > \delta\} \le P\{\alpha\lambda\|\mathbf{z}^\nu - z^*\| > \delta\}$$

for all sufficiently small $\delta > 0$.

Proof. The assumption on the inverse of the contingent derivative at 0 implies, by proposition 2.1 of [12], the existence of a neighborhood U of x^* and a constant $\lambda \geq 0$ such that

$$U \cap F^{-1}(y) \subset x^* + \lambda |y| B$$

for all y sufficiently close to 0 in $\mathbb{R}^m$. The conclusion follows from this and the uniform Lipschitz condition stated in the theorem. □

The set U in this theorem may be thought of as a sort of basin in which the asymptotics, if any, will apply and whose existence is guaranteed by the single-valuedness at 0 of the inverse of the contingent derivative. In the classical case this assumption is equivalent to the invertibility of the Jacobian of F and would therefore be sufficient to apply the implicit function theorem—from which would flow not only the bounds in probability but also an explicit description of the limiting behavior of the solutions $\mathbf{x}^\nu$. But such single-valuedness is not sufficient for the situations discussed in the introduction, and conditions must be made that compensate for the absence of the good local behavior that follows from differentiability. In [12] the following list of assumptions were shown to be sufficient for a certain implicit function theorem for the generalized equation (1.1).

Analytical Assumptions.

M.1 The space Z is a separable Banach space, and the function $f : Z \times \mathbb{R}^n \to \mathbb{R}^m$ is jointly continuous and has partial B-derivatives in both variables at (z^*, x^*), with $D_z f(z^*, x^*)$ strong.

M.2 The multifunction $N : \mathbb{R}^n \to \mathbb{R}^m$ is closed and proto-differentiable at $(x^*, -f(z^*, x^*))$.

M.3 The multifunction $F(\cdot) = f(z^*, \cdot) + N(\cdot)$ is subinvertible at $(x^*, 0)$.

M.4 The inverse contingent derivative

$$DF^{-1}(0|x^*)(y) = \{u \,\big|\, y \in D_x f(z^*, x^*)(u) + DN(x^*| - f(z^*, x^*))(u)\}$$

is at most a singleton for every $y \in \mathbb{R}^m$.

Remark 2.5. In [12], Section 5, it was shown that when the multifunction F is maximal monotone, then assumption M.4 implies M.3.

The essential step of any asymptotic argument is to introduce "local coordinates" around z^* and x^*: let the sequence of positive numbers $\{\tau_\nu\}$ tend to 0, let

$$\mathbf{z}^\nu = z^* + \tau_\nu \mathbf{w}^\nu,$$

and define

$$J_\nu(\mathbf{w}^\nu) = \{u \,\big|\, 0 \in f(z^* + \tau_\nu \mathbf{w}^\nu, x^* + \tau_\nu u) + N(x^* + \tau_\nu u)\}.$$

It is clear from the definition of the contingent derivative that

$$\limsup_{\nu\to\infty} \operatorname{gph} J_\nu = \operatorname{gph} DJ(z^*|x^*).$$

The standard argument now would run as follows: if $u_\nu(w)$ is a selection of $J_\nu(w)$ and it could be shown that $u_\nu \to u$ in sup-norm on compacts, where $\{u(w)\} = DJ(z^*|x^*)(w)$, then it would follow that $\mathbf{w}^\nu \underset{\mathcal{D}}{\longrightarrow} \mathbf{w}$ implies $u_\nu(\mathbf{w}^\nu) \underset{\mathcal{D}}{\longrightarrow} u(\mathbf{w})$ (cf. Billingsley [3] Theorem 5.5). One way to establish the convergence of the $u_\nu(\mathbf{w}^\nu)$ may be found in King [11]: first demonstrate the convergence in distribution of the $J_\nu(\mathbf{w}^\nu)$ as *random sets*, and then show that the single-valuedness assumption implies the corresponding convergence of any sequence of random selections.

Theorem 2.6. *(Asymptotic Distributions.) Assume M.1–4 and suppose that $\{\mathbf{z}^\nu\}$ is a sequence of random variables in the separable Banach space Z such that*

$$\tau_\nu^{-1}[\mathbf{z}^\nu - z^*] \underset{\mathcal{D}}{\longrightarrow} \mathbf{w},$$

Then, if a sequence $\{\mathbf{x}^\nu\}$ of measurable selections from the solution sets to (1.1) converges almost surely, it converges to the point x^*, and moreover,

$$\tau_\nu^{-1}[\mathbf{x}^\nu - x^*] \underset{\mathcal{D}}{\longrightarrow} DF^{-1}(0|x^*)(-D_z f(z^*, x^*)(\mathbf{w})).$$

Proof. The analytical assumptions M.1–4 imply that there exists a compact neighborhood U of x^* such that the solution multifunction J is semi-differentiable as a mapping from Z into $\mathbb{R}^n$ at the pair (z^*, x^*), with derivative

$$DJ(z^*|x^*)(\mathbf{w}) = DF^{-1}(0|x^*)(-D_z f(z^*, x^*)(\mathbf{w})).$$

Cf. Theorem 4.1 and Remark 4.3 of [12]. Now observe that eventually

$$\tau_\nu^{-1}(\mathbf{x}^\nu - x^*) \in \tau_\nu^{-1}[U \cap J(\mathbf{z}^\nu) - x^*].$$

The semi-differentiability of J implies, by Theorem 3.2 of [11], that the sequence of sets on the right-hand side converges in distribution to $DJ(z^*|x^*)(\mathbf{w})$. To obtain from this the convergence in distribution of the selections on the left side, we can apply Theorem 2.3 of [11] provided this sequence is *tight*. But by Theorem 4.1 of [12],

$$\tau_\nu^{-1}|\mathbf{x}^\nu - x^*| \le \lambda \tau_\nu^{-1}\|\mathbf{z}^\nu - z^*\|,$$

where λ is the Lipschitz constant for J at z^* and $\|\cdot\|$ is the norm in Z. The sequence on the right is *a fortiori* tight, and the proof is complete. □

We now apply Theorem 2.6 to treat M-estimates of the form

$$0 \in \overline{\mathbf{f}}^{\nu}(x) + N(x), \tag{2.5}$$

where for convenience we define $\overline{\mathbf{f}}^{\nu}(x) = \frac{1}{\nu}\sum_{i=1}^{\nu} f(x, \mathbf{s}_i)$ and $E\mathbf{f}(x) = Ef(x, \mathbf{s}_i)$. In the Appendix we show that the following assumptions imply the asymptotic normality of $\overline{\mathbf{f}}^{\nu}$ in the space of continuous functions $\mathcal{C}_m(U)$, for a given compact set U.

Probabilistic Assumptions.

P.1 The function $f : U \times S \to I\!R^m$ is continuous in the first variable and measurable in the second.

P.2 The sequence of random variables $\{\mathbf{s}_i\}$ is independent and identically distributed.

P.3 There is a point $x \in U$ with $E|f(x, \mathbf{s}_1)|^2 < \infty$.

P.4 There is a function $a : S \to I\!R$ with $E|a(\mathbf{s}_1)|^2 < \infty$ satisfying

$$|f(x_1, s) - f(x_2, s)| \le a(s)|x_1 - x_2| \quad \forall x_1, x_2 \in U.$$

Theorem 2.7. *Suppose that the assumptions P.1–4 hold for f with respect to a compact neighborhood U of x^*, that the function $E\mathbf{f}$ is B-differentiable at x^*, and that the assumptions M.2–4 are satisfied for the generalized equation (2.5) with $F = E\mathbf{f} + N$. Then, if a sequence $\{\mathbf{x}^{\nu}\}$ of measurable selections from the solution sets to (2.5) converges almost surely, it converges to the point x^*, and moreover,*

$$\sqrt{\nu}[\mathbf{x}^{\nu} - x^*] \underset{\mathcal{D}}{\longrightarrow} DF^{-1}(0|x^*)(-\mathbf{w}^*),$$

where $\mathbf{w}^*$ is normally distributed in $I\!R^m$ with mean 0 and covariance $\operatorname{cov} f(x^*, \mathbf{s}_1)$.

Proof. This fits the pattern of Theorem 2.6 if we observe that the evaluation map $e : \mathcal{C}_m(U) \times U \to I\!R^m$ has a strong partial B-derivative at any point $x \in U$ with B-derivative $D_z e(z, x)(w) = w(x)$. (See Remark 4.2 of [12].) Thus the generalized equation (2.5) is equivalent to

$$0 \in e(\overline{\mathbf{f}}^{\nu}, x) + N(x),$$

and it is easy to verify that M.1–4 hold here. The result now follows from this observation, Theorem 2.6, and Theorem A3 of the Appendix. □

3. Asymptotics for Stochastic Programs

As an application of the theory of the previous section, we consider the asymptotic behavior of sequences of solutions to a slightly more general version of a stochastic program than mentioned in the introduction, namely

$$\begin{aligned} &\text{minimize} && \overline{\mathbf{h}}^{\nu}(x) \\ &\text{subject to} && \overline{\mathbf{g}}^{\nu}(x) \in Q^o \\ &\text{and} && x \in C, \end{aligned} \tag{3.1}$$

where the set C is a convex polyhedral subset of $\mathbb{R}^n$, the set Q^o is the polar of a convex polyhedral cone in $\mathbb{R}^m$, and for all $s \in S$ the functions $h(\cdot, s) : \mathbb{R}^n \to \mathbb{R}$ and $g(\cdot, s) : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable. This form is a mathematically convenient generalization of the usual statement of a nonlinear program with equality and inequality constraints (which can be obtained by setting $Q = \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}_+$); it was introduced and studied by Robinson [14]. The problems (3.1) are to be regarded as perturbations of the "true" problem

$$\begin{aligned} &\text{minimize} && E\mathbf{h}(x) \\ &\text{subject to} && E\mathbf{g}(x) \in Q^o \\ &\text{and} && x \in C, \end{aligned} \tag{3.2}$$

In [12] we provided a second-order sensitivity analysis of this type of nonlinear program. The results of this section are direct consequences of that analysis, together with our results from the preceeding section.

In nonlinear programming, the sensitivity analysis of solutions cannot be separated from the sensitivity analysis of the Lagrange multipliers for the constraints. This study is no exception. Since in (3.1) we wish to cover the case of estimated constraints $\overline{\mathbf{g}}^{\nu}(x) \in Q^o$, we are forced to consider sequences of Kuhn-Tucker pairs $(\mathbf{x}^{\nu}, \mathbf{y}^{\nu})$ for (3.1) and not only sequences of solutions. Define the Lagrangian $k(x, y, s) = h(x, s) + y^T g(x, s)$, and let (x^*, y^*) be a Kuhn-Tucker pair for the problem (3.2), i.e. a solution to the Kuhn-Tucker equations

$$\begin{aligned} 0 \in &\nabla E\mathbf{h}(x) + y^T \nabla E\mathbf{g}(x) + N_C(x) \\ 0 \in &- E\mathbf{g}(x) + N_Q(y) \end{aligned} \tag{3.3}$$

By $N_C(x)$ and $T_C(x)$ we denote the normal and tangent cones, respectively, to a set C at a point x. The following analytical assumptions are assumed to hold at the given Kuhn-Tucker pair (x^*, y^*).

Analytical Assumptions for Stochastic Programs.

S.1 The Lagrangian $E\mathbf{k}(x,y)$ is twice continuously differentiable, and the *second-order sufficient condition* holds at (x^*,y^*):

$$u^T\nabla^2 E\mathbf{k}(x^*,y^*)u > 0$$

for every nonzero vector $u \in T_C(x^*)$ satisfying

$$\nabla E\mathbf{g}(x^*)u \in T_{Q^o}(E\mathbf{g}(x^*)) \quad \text{and} \quad \nabla E\mathbf{h}(x^*) = 0.$$

S.2 The constraint set $\{x \in C \,\big|\, E\mathbf{g}(x) \in Q^o\}$ is *regular* at x^* in the sense of [14], i.e.,

$$0 \in \text{ int}[E\mathbf{g}(x^*) + \nabla E\mathbf{g}(x^*)(C - x^*) - Q^o].$$

S.3 The *linear independence condition* holds at x^*, that is, the Jacobian matrix $\nabla E\mathbf{g}(x^*)$ has full rank.

The reader will recall that S.2 is the counterpart of the Mangasarian-Fromovitz constraint qualification for nonlinear programs. The linear independence assumption S.3 does not explicitly exclude inactive constraints, as in the usual statement of this condition, so we simply suppose these are dropped from the problem statement.

Let us rewrite the optimization problem (3.1) as one of generalized M-estimation by defining a function $f : I\!R^{n+m} \times S \to I\!R^{n+m}$ as

$$f(x,y,s) = (\nabla k(x,y,s), -g(x,s)),$$

and note that the Kuhn-Tucker conditions for the problem (3.1) correspond to the generalized equation

$$0 \in \overline{\mathbf{f}}^{\nu}(x,y) + N_{C\times Q}(x,y). \tag{3.4}$$

Theorem 3.1. *(Consistency.) Suppose that for the function f as above there exists a compact neighborhood U of x^* such that*

$$E\{\sup_{x\in U} |f(x,\mathbf{s}_1)|\} < +\infty, \tag{3.5}$$

and that the analytical assumptions S.1–2 hold. If $\{(\mathbf{x}^\nu,\mathbf{y}^\nu)\}$ is a sequence of Kuhn-Tucker pairs for (3.1) and $(\mathbf{x},\mathbf{y})$ is a cluster point of this sequence, then $(\mathbf{x},\mathbf{y})$ is a Kuhn-Tucker pair for (3.2) with probability one.

Proof. Under the assumptions S.1–2, it was shown in [12], Proposition 7.1, that the multifunction $E\mathbf{f} + N_{C\times Q}$ is subinvertible at 0. Without loss of generality suppose that

the compact set U is that given by the definition of subinvertibility (take the intersection, for example). Assumption (3.5) implies, by the strong law of large numbers, that $\overline{\mathbf{f}}^\nu \to E\mathbf{f}$ with probability one; cf. Etemadi [7]. Now apply Corollary 2.3 to the generalized equation (3.4), recalling that the function $(z, x) \mapsto z(x)$ is jointly continuous on $\mathcal{C}_n(U) \times U$. □

To obtain an expression for the central limit behavior, we saw in the previous section that it was necessary to consider an associated random generalized equation involving the normal random vector $\mathbf{w}(x^*)$ and the derivatives of $E\mathbf{f} + N$. For stochastic programs the corresponding object is a certain random quadratic program, which we now describe. If the probabilistic assumptions P.1–4 are satisfied for $f = (\nabla k, -g)$, then there exist Gaussian random functions $\mathbf{w}_1$ and $\mathbf{w}_2$ such that

$$\sqrt{\nu}[\nabla\overline{\mathbf{k}}^\nu - E\nabla\mathbf{k}] \underset{\mathcal{D}}{\longrightarrow} \mathbf{w}_1$$

and

$$\sqrt{\nu}[\overline{\mathbf{g}}^\nu - E\mathbf{g}] \underset{\mathcal{D}}{\longrightarrow} \mathbf{w}_2$$

Let $\mathbf{c}_1^* = \mathbf{w}_1(x^*, y^*)$ and $\mathbf{c}_2^* = \mathbf{w}_2(x^*)$. The *random quadratic program* giving the asymptotic distribution is

$$\begin{aligned} &\text{minimize} && \mathbf{c}_1^* u + \tfrac{1}{2} u^T \nabla^2 E\mathbf{k}(x^*, y^*) u \\ &\text{subject to} && \nabla E\mathbf{g}(x^*) u + \mathbf{c}_2^* \in [Q']^o \\ &\text{and} && u \in C' \end{aligned} \tag{3.6}$$

where

$$Q' = \{v \in T_Q(y^*) \,\big|\, v^T E\mathbf{g}(x^*) = 0\}$$

and

$$C' = \{u \in T_C(x^*) \,\big|\, u^T \nabla E\mathbf{k}(x^*, y^*) = 0\}.$$

Theorem 3.2. *Suppose that the probabilistic assumptions P.1–4 are satisfied for $f = (\nabla k, -g)$ and the analytical assumptions S.1–3 hold. If a sequence of Kuhn-Tucker pairs $\{(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$ for the problems (3.1) converges almost surely, then it converges to (x^*, y^*), and moreover,*

$$\sqrt{\nu}[(\mathbf{x}^\nu, \mathbf{y}^\nu)) - (x^*, y^*)] \underset{\mathcal{D}}{\longrightarrow} (\mathbf{u}, \mathbf{v}),$$

where $(\mathbf{u}, \mathbf{v})$ is the Kuhn-Tucker pair for the random quadratic program (3.6).

Proof. The multifunction $N = N_{C \times Q}$ is closed and proto-differentiable at every pair in its graph, because it is polyhedral; thus M.2 holds. In [12], Theorem 7.2, we showed that

assumptions S.1–3 imply M.4, and in the proof of Theorem 3.1 we observed that S.1–2 imply M.3. Assumption S.1 implies in particular that $E\mathbf{f}$ is B-differentiable at x^*. An application of Theorem 2.7 finishes the proof. □

4. Estimation for Linear-Quadratic Tracking Problems.

The following is a brief discussion of asymptotics for a class of linear-quadratic functions used for tracking stochastic objectives. A full treatment of this subject cannot properly be done in the confines of the present paper; our intention here is to expose the reader to an application of the preceding theory where twice differentiability does not smooth the way to an asymptotic theory.

The tracking problem we will consider here is

$$\begin{aligned} &\text{minimize} && E\rho(x) := E\{\rho(\mathbf{r}^T x)\} \\ &\text{subject to} && x \in C \end{aligned} \tag{4.1}$$

where the set X is a polyhedral subset of $\mathbb{R}^n$, and the function $\rho : \mathbb{R} \to \mathbb{R}$ is the one-sided piecewise linear-quadratic function

$$\rho(t) = \begin{cases} 0 & \text{if } t > 0 \\ \dfrac{1}{2}t^2 & \text{if } t \le 0 \end{cases}$$

penalizing deviations of $\mathbf{r}^T x$ below zero. We are after a theorem that establishes the asymptotic behavior of solutions $\mathbf{x}^\nu$ solving the estimated tracking problem

$$\begin{aligned} &\text{minimize} && \frac{1}{\nu}\sum_{i=1}^{\nu}\{\rho(\mathbf{r}_i^T x)\} \\ &\text{subject to} && x \in C \end{aligned} \tag{4.2}$$

Let us first go over some elementary facts.

Proposition 4.1. *Assume that $E|\mathbf{r}|^2 < \infty$. Then $E\rho(x)$ is finite and continuously differentiable, with*

$$\nabla E\rho(x) = E\{\mathbf{r}\rho'(\mathbf{r}^T x)\}$$

where

$$\rho'(t) = \begin{cases} 0 & \text{if } t > 0 \\ t & \text{if } t \le 0 \end{cases}$$

Proof. This merely asserts that the exchange of expectation and differentiation is permitted under our assumptions on $\mathbf{r}$. □

We now place this problem in the setting of Theorem 2.7, observing that a solution x^* is optimal for (4.1) if and only if it solves the generalized equation

$$0 \in E\{\mathbf{r}\rho'(\mathbf{r}^T x)\} + N_C(x) \tag{4.3}$$

and $\mathbf{x}^\nu$ is optimal for (4.2) if and only if it solves

$$0 \in \frac{1}{\nu}\sum_{i=1}^{\nu} \mathbf{r}_i \rho'(\mathbf{r}_i^T x) + N_C(x) \tag{4.4}$$

For notational convenience, denote the gradient mapping $r\rho'(r^T x)$ as $f(x, r)$, and as in the introduction, denote $E\{f(x, \mathbf{r})\}$ by $E\mathbf{f}$. The next fact to be proved is the B-differentiability of this gradient mapping.

Proposition 4.2. *Assume $E|\mathbf{r}|^2 < \infty$. Then the gradient mapping $E\mathbf{f}$ is B-differentiable, with*

$$D[E\mathbf{f}](x)(u) = \int_{[r^T x<0]} rr^T uP(dr) + \int_{[r^T x=0]\cap[r^T u<0]} rr^T uP(dr) \tag{4.5}$$

Proof. To verify B-differentiability in this case we examine the limit of difference quotients:

$$\lim_{t\downarrow 0, u'\to u} t^{-1}[E\mathbf{f}(x + tu') - E\mathbf{f}(x)].$$

We split the integration over the three subsets defined by the sign of $r^T x$, namely

$$R_- = \{r : r^T x < 0\},\ R_0 = \{r : r^T x = 0\},\ R_+ = \{r : r^T x > 0\}.$$

Since $r \in R_+$ implies eventually $r^T(x + tu') > 0$, we have

$$\lim_{t\downarrow 0, u'\to u} t^{-1} \int_{r\in R_+} [r\rho'(r^T(x + tu')) - r\rho'(r^T x)]P(dr) = 0.$$

In the case of R_-, we make a similar observation and then apply the Lebesgue dominated convergence theorem to yield

$$\lim_{t\downarrow 0, u'\to u} t^{-1} \int_{r\in R_-} [r\rho'(r^T(x + tu')) - r\rho'(r^T x)]P(dr) = \int_{r\in R_-} rr^T uP(dr).$$

(The absolute values of the integrands on the left are dominated by $|r|^2|u'|$.) This gives the first term in (4.5). For the remaining case, R_0, we must find the limit

$$\lim_{t\downarrow 0, u'\to u} t^{-1} \int_{r\in R_0} r\rho'(r^T tu')P(dr).$$

We apply a similar argument to this limit, splitting the integration according to the sign of $r^T u$. This gives the second term in (4.5). □

The dependence on u in the domain of integration in the second term of (4.5) can mean that $D[E\mathbf{f}](x)(u)$ is not linear in u and thus, in general, the tracking objective $E\rho$ is not twice differentiable. But twice differentiability is not required in Theorem 2.7; one needs only to establish M.2–4 and P.1–4. As we shall see, we have already done all the hard work and the only point left is to set reasonable conditions on the B-derivative (4.5) so that the analytical assumption M.4 holds.

Theorem 4.3. *Assume $E|\mathbf{r}|^2 < \infty$, and suppose that at a solution x^* to (4.1) the matrix $\int_{[r^T x^* < 0]} rr^T P(dr)$ is positive definite. Then, with f and $E\mathbf{f}$ defined as above, there exist unique solutions $u(w)$ to the second-order generalized equation*

$$w \in D[E\mathbf{f}](x)(u) + DN_C(x^* | - E\mathbf{f}(x^*))(u).$$

Furthermore, if a sequence $\{\mathbf{x}^\nu\}$ of solutions to (4.2) converges almost surely, it converges to the true solution x^*, and moreover,

$$\sqrt{\nu}[\mathbf{x}^\nu - x^*] \underset{\mathcal{D}}{\longrightarrow} u(\mathbf{w}),$$

where $\mathbf{w}^*$ is normally distributed with mean 0 and covariance cov $\mathbf{f}(x^*)$.

Proof. Proposition 4.2 shows that the B-derivative $D[E\mathbf{f}](x^*)$ is the sum of the two operators in (4.5): one is by assumption a positive definite linear operator with domain $\mathbb{R}^n$, and the other is a maximal monotone operator with domain $\mathbb{R}^n$. The first claim now is a consequence of Minty's Theorem [13]. The final claims follow from an application of Theorem 2.7. Condition M.4 has just been established. Conditions P.1–4 are given by the setting of our problem and the finiteness of $E|r|^2$. The normal cone operator is polyhedral, hence proto-differentiable—as already mentioned in Section 3—which establishes assumption M.2. The operator $F := E\mathbf{f} + N_C$ is maximal monotone, so by Remark 2.5 condition M.3 follows from M.4. This completes the proof. □

Appendix

In this appendix we briefly discuss central limit theory for random variables in $\mathcal{C}_m(U)$, the space of continuous $\mathbb{R}^m$-valued functions on a compact subset $U \subset \mathbb{R}^n$. Further details may be found in Araujo and Giné [2], on which this presentation has been based. The main result (A3) is a "well-known" theorem that does not seem to have been published for $m \geq 2$. The argument presented here was suggested by Professor R. Pyke.

For now, let Z be a separable Banach space equipped with its Borel sets $\mathcal{A}$, and let Z^* be the dual space of continuous linear functionals on Z. If $\mathbf{z}$ is a random variable taking values in Z, we say that $\mathbf{z}$ is (Pettis) *integrable* if there is an element $E\mathbf{z} \in Z$ for which $\ell(E\mathbf{z}) = E\{\ell(\mathbf{z})\}$ for all $\ell \in Z^*$, where $E\{\cdot\}$ denotes ordinary expected value. The *covariance* of $\mathbf{z}$, denoted $\operatorname{cov}\mathbf{z}$ is defined to be the mapping from $Z^* \times Z^*$ into $\mathbb{R}$ given by

$$(\operatorname{cov}\mathbf{z})(\ell_1, \ell_2) = E\{[(\ell_1(\mathbf{z}) - \ell_1(E\mathbf{z})][\ell_2(\mathbf{z}) - \ell_2(E\mathbf{z})]\}.$$

A random variable $\mathbf{z}$ taking values in Z will be called *Gaussian* with mean $E\mathbf{z}$ and covariance $\operatorname{cov}\mathbf{z}$ provided that for all $\ell \in Z^*$ the real-valued random variable $\ell(\mathbf{z})$ is normally distributed with mean $\ell(E\mathbf{z})$ and covariance $(\operatorname{cov}\mathbf{z})(\ell, \ell)$.

Let us now turn to the specific case at hand, that of the Banach space $\mathcal{C}_m(U)$. Let $(S, \mathcal{S})$ be a measurable space.

Proposition A1. *Assume P.1–4. Then the mapping $s \mapsto f(\cdot, s)$ is Borel measurable as a mapping from S into $\mathcal{C}_m(U)$.*

Proof. It suffices to show that for every $\alpha > 0$, the set

$$\{s \mid \sup_{x \in U} |f(s, x)| \leq \alpha\}$$

is a measurable subset of S. This follows easily from standard results in the theory of measurable multifunctions; see, for example, Rockafellar [15], Theorem 2K. □

Corollary A2. *$\bar{\mathbf{f}}^\nu$ is a $\mathcal{C}_m(U)$-valued random variable for every ν.* □

Theorem A3. *Suppose that $f : U \times S \to \mathbb{R}^m$ satisfies the probabilistic assumptions P.1–4. Then there exists a Gaussian random variable $\mathbf{w}$ taking values in $\mathcal{C}_m(U)$ such that*

$$\sqrt{\nu}(\bar{\mathbf{f}}^\nu - E\mathbf{f}) \xrightarrow[\mathcal{D}]{} \mathbf{w}.$$

Proof. Each $\bar{\mathbf{f}}^\nu$ is a vector of continuous functions $(\bar{\mathbf{f}}_1^\nu, \ldots, \bar{\mathbf{f}}_m^\nu)$. The conditions of the theorem imply that for each $j = 1, \ldots, m$ there is a Gaussian random variable in $\mathcal{C}_m(U)$ with zero mean and covariance equal to $\operatorname{cov}\mathbf{f}_j$, which we suggestively call $\mathbf{w}_j$, such that

$$\sqrt{\nu}(\bar{\mathbf{f}}_j^\nu - E\mathbf{f}_j) \xrightarrow[\mathcal{D}]{} \mathbf{w}_j;$$

cf. [2], Theorem 7.17. It follows that the finite-dimensional distributions of $\mathbf{w}^\nu := \sqrt{\nu}(\bar{\mathbf{f}}^\nu - E\mathbf{f})$ converge to those of $\mathbf{w}$, i.e. for all finite subsets $\{x_1, \ldots, x_k\} \subset U$ one has

$$(\mathbf{w}^\nu(x_1), \ldots, \mathbf{w}^\nu(x_k)) \xrightarrow[\mathcal{D}]{} (\mathbf{w}(x_1), \ldots, \mathbf{w}(x_k)).$$

This determines the limit $\mathbf{w}$, if it exists, uniquely as that in the statement of the theorem. Thus by Prohorov's Theorem (Billingsley [3] Theorem 6.1) it remains only to show that the sequence $\{\mathbf{w}^\nu\}$ is *tight* in $\mathcal{C}_m(U)$, i.e. for each $\varepsilon > 0$ there is a compact set $A \subset \mathcal{C}_m(U)$ such that $\Pr\{\mathbf{w}^\nu \in A\} > 1 - \varepsilon$ for all sufficiently large ν. By adapting the argument of [3], Theorem 8.2, for $\mathcal{C}_m(U)$ we find that the tightness of $\{\mathbf{w}^\nu\}$ is equivalent to the simultaneous satisfaction of the following two conditions:

(i) There exists $x \in U$ such that for each $\eta > 0$ there is $\alpha \geq 0$ with

$$\Pr\{|\mathbf{w}^\nu(x)| > \alpha\} \geq \eta, \quad \forall \nu \geq 1.$$

(ii) For each positive ε and η there exist $\delta > 0$ and an integer ν_0 such that

$$\Pr\{\sup_{(x-y)<\delta} |\mathbf{w}^\nu(x) - \mathbf{w}^\nu(y)| \geq \varepsilon\} \leq \eta, \quad \forall \nu \geq \nu_0.$$

These conditions follow easily from the tightness of the coordinate sequences $\{\mathbf{w}_j^\nu\}$ for $j = 1, \ldots, m$ since

$$\Pr\{|\mathbf{w}^\nu(x)| > \alpha\} \leq \sum_{j=1}^{m} \Pr\left\{|\mathbf{w}_j^\nu(x)| > \frac{\alpha}{\sqrt{m}}\right\},$$

and similarly for the probability in condition (ii), and hence these can be made as small as one pleases by application of conditions (i) and (ii) to the co-ordinate sequences. Thus $\{\mathbf{w}^\nu\}$ is tight, and the proof is complete. □

References

1. J. Aitchison and S.D. Silvey, "Maximum likelihood estimation of parameters subject to restraints", *Annals of Mathematical Statistics* **29**(1948), 813–828.
2. A. Araujo and E. Giné, *The Central Limit Theorem for Real and Banach Valued Random Variables*, (Wiley, New York, 1980).
3. P. Billingsley, *Convergence of Probability Measures*, (Wiley, New York, 1968).
4. B.R. Clarke, "Nonsmooth analysis and Fréchet differentiability of M-functionals", *Probability Theory and Related Fields* **73** (1986) 197–209.
5. J. Dupačová and R.J-B Wets, "Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems", *Annals of Statistics* (1988) 1517–1549.
6. Yu. M. Ermoliev and V. I. Norkin, "Normalized convergence in stochastic optimization", (Working Paper WP-89-091, International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria, 1989).
7. N. Etemadi, "An elementary proof of the strong law of large numbers", *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandet Gebeite* **55** (1981) 119–122.
8. P.J. Huber, "The behavior of maximum likelihood estimates under non-standard conditions", *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics* (1967), 221–233.
9. I.A. Ibragimov and R.Z. Has'minskii, *Statistical Estimation: Asymptotic Theory* (Springer, New York, 1981).
10. A.J. King, *Asymptotic Behaviour of Solutions in Stochastic Optimization: Nonsmooth Analysis and the Derivation of Non-normal Limit Distributions*, (Ph.D. Dissertation, University of Washington, 1986).
11. A.J. King, "Generalized delta theorems for multivalued mappings and measurable selections", *Mathematics of Operations Research* (1989) 720–736.
12. A.J. King and R.T. Rockafellar, "Sensitivity analysis for nonsmooth generalized equations", *Mathematical Programming* (to appear, 1991).
13. G.J. Minty, "Monotone (nonlinear) operators in Hilbert space", *Duke Mathematics Journal* **29** (1962) 341–346.
14. S.M. Robinson, "Generalized equations and their solutions, part II: applications to nonlinear programming", *Mathematical Programming Study* **19** (1982) 200–221.
15. R.T. Rockafellar, "Integral functionals, normal integrands and measurable selections", in *Nonlinear Operators and the Calculus of Variations*, (Lecture Notes in Mathematics 543, Springer-Verlag, Berlin, 1976) 157–207.
16. A. Shapiro, "Asymptotic properties of statistical estimators in stochastic programming", *Annals of Statistics* **17** (1989) 841–858.
17. A. Shapiro, "Asymptotic analysis of stochastic programs", (Manuscript, University of South Africa, Pretoria, South Africa, 1990).