Annals of Operations Research 56(1995)15–38

Continuous approximation schemes for stochastic programs

John R. Birge*
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109, USA

Liqun Qi**
School of Mathematics, The University of New South Wales, Kensington, N.S.W. 2033, Australia

One of the main methods for solving stochastic programs is approximation by discretizing the probability distribution. However, discretization may lose differentiability of expectational functionals. The complexity of discrete approximation schemes also increases exponentially as the dimension of the random vector increases. On the other hand, stochastic methods can solve stochastic programs with larger dimensions but their convergence is in the sense of probability one. In this paper, we study the differentiability property of stochastic two-stage programs and discuss continuous approximation methods for stochastic programs. We present several ways to calculate and estimate this derivative. We then design several continuous approximation schemes and study their convergence behavior and implementation. The methods include several types of truncation approximation, lower dimensional approximation and limited basis approximation.

Keywords: Approximation, derivative, continuous distribution, stochastic programming.

1. Introduction

Consider the following *stochastic program with recourse*:

$$\begin{aligned} &\text{minimize } z = c^{\mathrm{T}}x + \Psi(x) \\ &\text{subject to } Ax = b, \\ &\qquad x \geq 0, \end{aligned} \tag{1.1}$$

* His work is supported by Office of Naval Research Grant N0014-86-K-0628 and the National Science Foundation under Grant ECS-8815101 and DDM-9215921.
** His work is supported by the Australian Research Council.

where $\Psi(x) = E(\psi(Tx - \xi))$ and ψ is the *recourse function*, defined by

$$\psi(w) = \inf\{q^{\mathrm{T}}y : Wy = -w, y \geq 0\}. \tag{1.2}$$

Then

$$\Psi(x) = \int \psi(Tx - \xi)P(d\xi). \tag{1.3}$$

The dimensions in (1.1), (1.2) and (1.3) are: $x, c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $y, q \in \mathbb{R}^k$, $\xi \in \mathbb{R}^l$. The random l-vector ξ is defined on a probability space $(\Xi, \mathcal{A}, P)$. To ensure that Ψ is convex, real-valued and defined on $\mathbb{R}^n$, we assume throughout that (i) for each $t \in \mathbb{R}^k$, there exists $y \geq 0$, $y \in \mathbb{R}^k$ such that $Wy = t$, (ii) there exists $\pi \in \mathbb{R}^l$ such that $\pi^{\mathrm{T}}W \leq q^{\mathrm{T}}$, and (iii) $\int_\Xi \|\xi\| P(d\xi) < \infty$.

A popular approach to solve (1.1) is to discretize P to get a large-scale linear program approximation to (1.1) [2, 6, 7, 13, 15, 21, 24, 30]. The approximation function to Ψ in this case in general is *nondifferentiable*. However, if P is a *continuous distribution*, i.e.,

$$P(d\xi) = p(\xi)d\xi, \tag{1.4}$$

where p is a Lebesgue integrable function, then Ψ is *differentiable* [20, 28]. In fact, by (1.2), ψ is a piecewise linear function. Hence, ψ is piecewise constant except in a Lebesgue zero measure set. If P is discrete, then this Lebesgue zero measure set may have positive measure in P. Thus, Ψ may be nondifferentiable. If P is continuous, then we may omit this set and have

$$\nabla\Psi(x) = \int \nabla\psi(Tx - \xi)p(\xi)d\xi. \tag{1.5}$$

This observation leads to continuous distribution approximations to solve (1.1). Convergence properties and uses in algorithms are discussed in [4]. To make this approach feasible, one needs to approximate (1.4) efficiently. This paper considers such schemes. Section 2 presents approaches to estimate $\nabla\Psi(x)$ based on direct probability approximations. Section 3 discusses truncation approximations and gives convergence rates. Section 4 presents convergence results and approaches for lower dimensional approximations. Section 5 gives results on using combinations of the previous approaches and section 6 provides some examples and illustrative results.

2. Calculation and estimation of the derivative

Suppose that P is a continuous distribution defined by (1.4). The following proposition gives a formula for $\nabla\Psi(x)$ based upon basic optimal dual solutions of (1.2).

PROPOSITION 2.1

Let the basic dual optimal solutions of (1.2) be $\{\pi_j : j = 1, \ldots, N\}$. Let the basic matrix of (1.2), associated with π_j, be B_j. Then

$$\nabla\Psi(x) = \sum_{1 \le j \le N} \alpha_j \pi_j T, \tag{2.1}$$

where

$$\begin{aligned} \alpha_j &\equiv P(\xi \mid \pi_j \text{ optimal}) \\ &= P(\xi \mid B_j^{-1}(Tx - \xi) \ge 0) \\ &= P(\xi \mid B_j^{-1} Tx \ge B_j^{-1}\xi). \end{aligned} \tag{2.2}$$

Proof

By (1.4) and (1.5),

$$\begin{aligned} \nabla\Psi(x) &= \sum_{1 \le j \le N} \{\pi_j T P(\xi \mid \pi_j \text{ optimal})\} \\ &= \sum_{1 \le j \le N} \alpha_j \pi_j T. \end{aligned} \tag{2.3}$$

This formula holds for continuous distribution since the set of ξ in which the dual problem of (1.2) has no unique solution has zero measure in P in this case.

By the optimality condition of (1.2),

$$\begin{aligned} \alpha_j &\equiv P(\xi \mid \pi_j \text{ optimal}) \\ &= P(\xi \mid B_j^{-1}(Tx - \xi) \ge 0) \\ &= P(\xi \mid B_j^{-1} Tx \ge B_j^{-1}\xi). \end{aligned}$$

This proves (2.2). □

We now discuss some practical ways to calculate $\nabla\Psi(x)$. The first approach is based on approximating probabilities as in the Boole–Bonferroni approach of Prékopa [23].

THEOREM 2.2

In (2.1), we have

$$\alpha_j \equiv P(\xi \mid \pi_j \text{ optimal})$$
$$= 1 - \left(a_j - \frac{2b_j}{l}\right) + t_j\left[\frac{(c_j - 1)a_j}{c_j + 1} - \frac{2(l + c_j(c_j + 1))b_j}{l(c_j(c_j + 1))}\right], \tag{2.4}$$

where

$$a_j = \sum_{1 \le i \le l} P(\eta_{ji} > s_{ji}(x)),$$

$$b_j = \sum_{1 \le i < i' \le l} P(\eta_{ji} > s_{ji}(x), \eta_{ji'} > s_{ji'}(x)),$$

$$c_j - 1 = \left\lfloor \frac{2b_j}{a_j} \right\rfloor,$$

$$0 \le t_j \le 1,$$

$$\eta_{ji} = (B_j^{-1})_i \xi,$$

$$s_{ji}(x) = (B_j^{-1})_i Tx,$$

$(B_j^{-1})_i$ is the ith row of B_j^{-1}.

Proof

Let $A_{ji} = A_{ji}(x) = \{\eta_{ji} | s_{ji}(x) \ge \eta_{ji}\}$. Then

$$P(\xi \mid B_j^{-1}Tx \ge B_j^{-1}\xi) = P(A_{j1} \cdots A_{jl}) = 1 - P(\bar{A}_{j1} + \cdots + \bar{A}_{jl}). \tag{2.5}$$

By the inequality of Dawson and Sankoff ((7) of [23]),

$$P(\bar{A}_{j1} + \cdots + \bar{A}_{jl}) \ge \frac{2}{c_j + 1} a_j - \frac{2}{c_j(c_j + 1)} b_j, \tag{2.6}$$

where

$$a_j = \sum_{1 \le i \le l} P(\bar{A}_{ji}) = \sum_{1 \le i \le l} P(\eta_{ji} > s_{ji}(x)),$$

$$b_j = \sum_{1 \le i < i' \le l} P(\bar{A}_{ji} \cdot \bar{A}_{ji'})$$

$$= \sum_{1 \le i < i' \le l} P(\eta_{ji} > s_{ji}(x), \eta_{ji'} > s_{ji'}(x)),$$

$$c_j - 1 = \left\lfloor \frac{2b_j}{a_j} \right\rfloor.$$

Similarly, by the inequality of Sathe et al. ((8) of [23]),

$$P(\bar{A}_{j1} + \cdots + \bar{A}_{jl}) \le a_j - \frac{2}{l} b_j. \tag{2.7}$$

Combining (2.5)–(2.7), we get the conclusions of this theorem. □

We may use (2.4) to approximate $\nabla\Psi(x)$ by assigning t_j a value in $[0, 1]$, e.g., 0.5. We may also use higher orders of the inclusion–exclusion formula for α_j (see section 6 for further discussion of this approach). The random variable η_{ji} is one-dimensional. If we know the marginal distribution of η_{ji} and the joint distribution of η_{ji} and $\eta_{ji'}$, where $1 \le i < i' \le l$, then we may calculate a_j and b_j. In general, η_{ji} and $\eta_{ji'}$ are linear combinations of $\{\xi_h, h = 1, \ldots, l\}$. Suppose that $\eta = \sum_{1 \le h \le l} \beta_h \xi_h$ and $\eta' = \sum_{1 \le h \le l} \beta'_h \xi_h$, where ξ_h, $h = 1, \ldots, l$, are one-dimensional random variables, β_h and β'_h, $h = 1, \ldots, l$, are real numbers. By probability theory,

$$E(\eta) = \sum_{1 \le h \le l} \beta_h E(\xi_h), \tag{2.8}$$

$$\mathrm{Var}(\eta) = \sum_{1 \le h \le l} \beta_h^2 \mathrm{Var}(\xi_h) + 2 \sum_{1 \le h < h' \le l} \beta_h \beta_{h'} \mathrm{Cov}(\xi_h, \xi_{h'}), \tag{2.9}$$

$$\mathrm{Cov}(\eta, \eta') = \sum_{1 \le h \le l} \beta_h \beta'_h \mathrm{Var}(\xi_h) + \sum_{\substack{1 \le h, h' \le l \\ h \ne h'}} \beta_h \beta'_h \mathrm{Cov}(\xi_h, \xi_{h'}). \tag{2.10}$$

Therefore, we may calculate the expectations, variances and covariances of η_{ji} and $\eta_{ji'}$ by the expectations, variances and covariances of $\{\xi_h, h = 1, \ldots, l\}$.

If ξ_h, $h = 1, \ldots, l$, are normally distributed, then η_{ji} and $(\eta_{ji}, \eta_{ji'})$ in (2.8)–(2.10) are also normally distributed and bivariately normally distributed respectively. Thus, we may calculate a_j and b_j in theorem 2.2 by some single and double integrals. Otherwise, we may use (2.8)–(2.10) to calculate expectations, variances and covariances of those random variables η_{ji} and $\eta_{ji'}$, and then use the method in [7] to approximate the relevant probabilities.

In fact, with gaussian variables, we may also take advantage of techniques designed specifically for these distributions as in Gassmann [17] and Deák [10]. Through the transformation, $\eta_j = B_j^{-1}\xi$, we obtain:

$$\alpha_j = \int_{\eta_{j1} \leq s_{j1}(x)} \cdots \int_{\eta_{jl} \leq s_{jl}(x)} p(\eta_{j1}, \ldots, \eta_{jl}) d\eta_{j1} \cdots d\eta_{jl},$$

where the variables have the same definitions as in theorem 2.2. Notice that $p(\eta_j)$ is a normal density so that the calculation of α_j reduces to evaluating the probabilities of the $s_j(x)$ translation of the lower orthant for the normally distributed vector, η_j.

We may also approximate other distributions using Gaussian random variables. By fixing the means, variances and covariances of the relevant random variables, we can use a gaussian random vector with the same characteristics. Any lack of precision may then come from higher order moments. Since it is often difficult to determine exact multivariate distributions, the use of normal distributions seems especially relevant even if they are not completely justified by the practical situation.

3. Truncation approximations

By *truncation approximations*, we mean any approximation of the form:

$$P^\nu(A) = P(A \cap \Xi^\nu)/P(\Xi^\nu), \tag{3.1}$$

for some $\Xi^\nu \subset \Xi$ for all $A \subset \mathscr{A}$ and where $P(\Xi \backslash \Xi^\nu) = \delta^\nu$. In other words, P^ν is the restriction of P to Ξ^ν or it results from truncating $\Xi \backslash \Xi^\nu$. These types of approximations have an advantage in that they only depend on that set Ξ^ν which can be found in many ways. In each case, however, we not only have the convergence cited above but we can also describe the rate of that convergence. We need only suppose the following condition on ψ.

(i) The set D and $\nabla_x \psi(Tx - \xi)$ are bounded so that, wherever it is defined, $\|(\partial\psi(Tx - \xi))/\partial x_i\| \leq \Delta$ for all i, $\xi \in \Xi$ and $x \in D$.

The result is that the convergence of any subgradient in $\partial\Psi^\nu$ to an element of $\partial\Psi$ can be bounded.

THEOREM 3.1

Suppose (i) and that P^ν is generated as in (3.1), then for any $\eta^\nu \in \partial\Psi^\nu(x)$, there exists $\eta \in \partial\Psi(x)$ such that,

$$\|\eta^\nu - \eta\| \leq 2\sqrt{n}\Delta\delta^\nu. \tag{3.2}$$

Proof

First note that $\|\eta^\nu - \eta\|$ can be written as

$$\left\{\sum_i \left\{\left|\int \eta_i(\xi)P^\nu(d\xi) - \int \eta_i(\xi)P(d\xi)\right|\right\}^2\right\}^{1/2},$$

where we choose the same $\eta(\xi)$ from $\partial_x\psi(Tx - \xi)$ for each ξ. For this, we observe that

$$\int \eta_i(\xi)P^\nu(d\xi) - \int \eta_i(\xi)P(d\xi) = \int_{\Xi^\nu} \eta_i(\xi)\left(\frac{\delta^\nu}{1-\delta^\nu}\right)P(d\xi) + \int_{\Xi\setminus\Xi^\nu} \eta_i(\xi)P(d\xi)$$

$$\leq 2\Delta\delta^\nu. \tag{3.3}$$

From (3.3) and the definition, we have $\|\eta^\nu - \eta\| \leq 2\sqrt{n}\Delta\delta^\nu$. □

Theorem 3.1 gives conditions under which distributions of the form in (3.1) may converge at a given rate. Other rates may be obtained using results about the difference between function values for different distributions as for example in Römisch and Schultz [26]. The rate can then depend on the some metric between the approximating and true measures. In the lower dimensional approximations, we again use this to obtain $\|\eta^\nu - \eta\| \leq \sigma^\nu$, where $\sigma^\nu \to 0$.

The choice of Ξ^ν is the crucial step in these types of approximations. In general, we would like fast convergence by having δ^ν converging to zero quickly but we would also like to have relatively easy computations. The following alternatives are considered for the truncation method:

(1) *Ball truncation*: Let $\Xi^\nu = \{\xi \mid \|\xi - \bar{\xi}\| \leq \gamma^\nu\}$ where $\gamma^\nu \to \infty$. This method is easy to describe but integrals over the ball may be difficult to evaluate.

(2) *Box truncation*: Let $\Xi^\nu = \{\xi \| \xi(i) - \bar{\xi}(i)| \leq \gamma^\nu(i)$ where $\gamma^\nu(i) \to \infty$ for all $i\}$. The only difference from the ball approximation is that we use boxes of possibly varying shapes. Depending on the distribution, these integrals may be easier than those in ball truncation.

(3) *Scaling truncation*: Suppose Ξ is bounded. Let $\Xi^\nu - \bar{\xi} = \alpha^\nu(\Xi - \bar{\xi})$ where $\alpha^\nu \to 1$. In this case, we use the shape of Ξ to determine the form of Ξ^ν. This may present an advantage over the previous approaches in that each component of ξ is treated similarly relative to Ξ.

(4) *Level set truncation*: We assume that, similar to the scaling truncation, ξ is distributed so that for some Ξ^0, $P(\alpha(\Xi^0 - \bar{\xi}) + \bar{\xi}) = 1 - \delta(\alpha)$, where $\delta \to 0$ as $\alpha \to \infty$. We then let $\Xi^\nu = \alpha(\nu)(\Xi^0 - \bar{\xi}) + \bar{\xi}$ for some sequence of $a(\nu) \to \infty$ as $\nu \to \infty$. These values may for example correspond to choices of ν such that $\delta(\alpha(\nu)) = 1/\nu$. For unimodal continuous distributions with mode $\bar{\xi}$, the Ξ^ν correspond to level sets of the density function. This approach may be especially useful for gaussian distributions with ellipsoidal level regions.

(5) *Limited basis truncation*: In this approach, we suppose that $\Pi^\nu = \{\pi_i : i = 1, \ldots, \nu\}$ where each π_i is optimal in the dual of (1.2) for some $w = Tx - \xi$. We then let $\Xi^\nu(x) = \{\xi | \phi(Tx - \xi) = \max_{i=1,\ldots,\nu} \pi_i(Tx - \xi)\}$, i.e., such that the dual of (1.2) is optimized by some $\pi \in \Pi^\nu$. Note that this approximation also depends on x but that it is finitely convergent in ν for any fixed x. Many strategies may be used to determine the Π^ν set. By choosing only π with high probability of optimality in (1.2) we may be able to limit the ν and still produce accurate results. To alleviate the difficulty, we may combine this procedure with the integration approximation mentioned in section 2 or with the lower dimensional approximations in section 4 below. The combined approximations will still converge as given in section 5.

Each of these procedures has the convergence rate in theorem 3.1. They also are continuous and retain the differentiability properties mentioned above. The form of the gradient also allows the use of an algorithm that only evaluates the probability of individual bases or dual vectors π_i being optimal in (1.2) and its dual. Since probabilities of these convex polyhedral regions may be relatively easier to evaluate than the conditional expectations required for function evaluations, these procedures are especially attractive.

4. Lower dimensional approximation

In addition to the methods considered above that use full dimensional but hopefully simpler integration, we may also approximate the distributions using lower dimensional distributions that still retain differentiability properties and avoid higher dimensional integrals. These procedures are generalizations of discrete point approximations and, as we show below, obtain improved convergence rates over discrete approximations.

We may begin by replacing ξ by $H^\nu \eta^\nu$, where $H^\nu \in \mathbb{R}^{N \times N_\nu}$, or, in other words, we may let P^ν be:

$$P^\nu(A) = Q^\nu\{\eta^\nu | H^\nu \eta^\nu \in A\}, \tag{4.1}$$

where Q^ν is a probability measure on η^ν. In some examples, H^ν might be a single vector $\bar{\xi}$ so that η^ν approximates ξ using a distribution on the ray through $\bar{\xi}$ or an identity on certain critical components of ξ with other components of ξ expressed as linear transformations of the critical components. The choice of H^ν would depend on specific problem structure, however. General rules would probably not prove too effective.

Building on this approximation, we can construct generalizations of discrete approximations. We may in fact create several lower dimensional H^ν with different probabilities and some translation, i.e., $H_i^\nu \eta_i^\nu + \beta_i$ would be used in (4.1) and given a probability p_i so that (4.1) becomes:

$$P^\nu(A) = \sum_{i=1}^{K(\nu)} p_i Q_i^\nu \{\eta_i^\nu | H_i^\nu \eta_i^\nu + \beta_i \in A\}, \tag{4.2}$$

where Q_i^ν is the probability measure on η_i^ν. In this way, we obtain direct generalizations from the discrete distributions.

We obtain convergence of these distributions by allowing the number of lower dimensional approximations $K(\nu)$ or the dimension N_ν to increase sufficiently. We concentrate on increasing the number of the approximations since increases in the dimension N_ν lead eventually to integrations as difficult as the original. We establish weak convergence of the measure P^ν to P and, moreover, convergence of moments, using Wasserstein metrics. These results build on Römisch and Schultz's work in [27].

In our context, we let $\mathscr{P}^l$ be the class of Borel probability measures on $\mathbb{R}^l$. This is further restricted to:

$$\mathscr{M}_p = \left\{ Q \in \mathscr{P} \middle| \int \|\xi\|^p Q(d\xi) < \infty \right\}.$$

We then let

$$\mathscr{D}(Q_1, Q_2) = \{Q \in \mathscr{P}^{2l} | Q \circ \Pi_1^{-1} = Q_1, Q \circ \Pi_2^{-1} = Q_2\},$$

where $\Pi_1(\xi)$ is projection on the first l coordinates of ξ and $\Pi_2(\xi)$ is projection on the second l coordinates. For $p \geq 1$ the L_p-Wasserstein metric is defined as:

$$W_p(Q_1, Q_2) = \left[\inf \left\{ \int_{\mathbb{R}^l \times \mathbb{R}^l} \|\xi_1 - \xi_2\|^p Q(d\xi_1, d\xi_2) | Q \in \mathscr{D}(Q_1, Q_2) \right\} \right]^{1/p}.$$

In the following, we will use several results given previously. We use $\to_w$ to denote weak convergence.

THEOREM 4.1

For $\mathcal{M}_p$ defined above, $(\mathcal{M}_p, W_p)$ is a metric space.

Proof

See Givens and Shortt [18]. □

THEOREM 4.2

If $Q \in \mathcal{M}_p$ and $Q_n \in \mathcal{M}_p$ $(n = 1, \ldots)$, then $W_p(Q_n, Q) \to 0$ as $n \to \infty$ if and only if $Q_n \to_w Q$ and $\lim_{n\to\infty} \int_{\mathbb{R}^l} \|\xi\|^p Q_n(dz) = \int_{\mathbb{R}^l} \|z\|^p Q(d\xi)$.

Proof

See Rachev [25]. □

THEOREM 4.3

If $h : \mathbb{R}^l \to \mathbb{R}$ is Lipschitzian on bounded subsets of $\mathbb{R}^l$, then, for all $P, Q \in \mathcal{M}_p$, $|\int h(x)P(dx) - \int h(x)Q(dx)| \leq \{M_q(P) + M_q(Q)\}W_p(P, Q)$, where $(1/p) + (1/q) = 1$, $p > 1$, and $M_q(P) = \{\int L_h(\|x\|)^q P(dx)\}^{1/q}$, where $L_h(r) = \sup_{x\neq x', \|x\| \leq r, \|x'\| \leq r}\{(|h(x) - h(x')|)/\|x - x'\|\}$.

Proof

See Römisch and Schultz [27]. □

We now construct the lower dimensional approximations to employ the results above and obtain convergence results. We make the following assumptions relating the distributions of P and the Q_i^ν. Let $\Pi_i^\nu = (H^{\nu\mathrm{T}}H^\nu)^{-1}H^{\nu\mathrm{T}}$ be projection into the affine space spanned by H^ν and let $\Pi_i^{-\nu} = \{\xi | \Pi_i^\nu(\xi - \beta_i) = \eta\}$.

(i) The distribution P is continuous, $P \in \mathcal{M}_p$ for some $p > 1$, and $P(A) = \int_{\xi\in A} p(\xi)d\xi$.

(ii) The distribution P^ν is constructed as in (4.2) such that

(a) there exists a partition of $\mathbb{R}^l$ into $A_1 \cup A_2 \cup \cdots \cup A_{K(\nu)}$, where $A_i \cap A_j = \emptyset$ for $i \neq j$;

(b) each A_i is partitioned into disjoint subsets A_i^1 and A_i^2 such that $\sup_{\xi,\xi'\in A_i^1 \cap \Pi_i^{-\nu}(\eta)} \|(\xi - \xi'\|^p \leq \epsilon(\nu)$ for all i and η, and $\int_{A_i^2} \|\xi\|^p P(d\xi) \leq \delta(\nu)$ with $\epsilon(\nu) \to 0$ and $\delta(\nu) \to 0$ as $\nu \to \infty$;

(c) the weights $p_i = P(A_i)$ and $Q_i^\nu(B) = \int_B q_i^\nu(\eta)d\eta$ where $q_i^\nu(\eta)d\eta = (1/p_i)\int_{\xi \in A_i \cap \Pi_i^\nu(\eta)} p(\xi)d\xi$.

The conditions in (i) and (ii) may appear difficult to satisfy in general. We, however, wish to consider distributions that can be constructed as linear transformations of independent random variables.

This is always possible, for example, with multivariate normal distributions. Other multivariate distributions may be explicitly constructed this way.

In this case, if $\xi = H^\nu\eta + H^{\nu'}\zeta$ where η and ζ are independently distributed and $H^{\nu'}$ spans the null space of H^ν, then

$$q_i^\nu(\eta)d\eta = (1/p_i) \int_{H^\nu\eta + H^{\nu'}\zeta + \beta_i \in A_i} p_1(\eta)p_2(\zeta)d\eta d\zeta,$$

which is $p_1(\eta)d\eta$ multiplied by the relative probability that $H^\nu\eta + H^{\nu'}\zeta + \beta_i \in A_i$ given η. By choosing the $A_i = \{\xi | \xi = H^\nu\eta + H^{\nu'}\zeta + \beta_i$ for some η and $\|H^{\nu'}\zeta - \beta\| \leq \gamma_i\}$, the relative probability can be calculated easily. In fact, we can also just let the β_i result from the product of $H^{\nu'}$ and some random ζ as a generalization of the empirical measure.

The next theorem shows that this distribution satisfies the conditions for weak convergence and convergence of the p moments.

THEOREM 4.4

If P and P^ν satisfy (i) and (ii), then $W_p(P, P^\nu) \leq \epsilon(\nu) + \delta(\nu)$.

Proof

We will construct a distribution $Q \in \mathscr{D}(P, P^\nu)$ that leads to the conclusion. For $\xi \in \mathbb{R}^{2l}$, let $\xi_1 = \Pi_1(\xi)$ and let $\xi_2 = \Pi_2(\xi)$. We let $A_i^\nu = \{\xi \in A_i | \xi - H_i^\nu \Pi_i^\nu(\xi - \beta_i) = 0\}$ and $A_i(\eta) = \{\xi | \xi - \beta_i = H_i^\nu\eta\}$. We then define

$$Q(\xi | \xi \in A) = \sum_{i=1}^{K(\nu)} \int_{A_i(\eta) = \xi_1; \xi \in A} \int_{\xi_2 \in A_i \cap \Pi_i^{-\nu}(\eta)} p(\xi_2)d\xi_2.$$

First, we wish to show that $Q \in \mathscr{D}(P, P^\nu)$. For this, we write $p(\xi)d\xi$ as $p'(\eta, \zeta)d\eta d\zeta$ (where $p'(\eta, \zeta) = |[HH']|p(H\eta + H'\zeta)$), then

$$Q \circ \Pi_1^{-1}(A) = \sum_{i=1}^{K(\nu)} \int_{A_i(\eta) \in A} \frac{1}{p_i} q_i^\nu(\eta)d\eta = P^\nu(A), \tag{4.3}$$

and

$$Q \circ \Pi_2^{-1}(A) = \sum_{i=1}^{K(\nu)} \int_{A_i(\eta) \in A_i} \int_{\xi \in A_i \cap \Pi_i^{-\nu}(\eta)} p(\xi) d\xi = P(A). \tag{4.4}$$

We next use the assumptions to obtain bounds on the Wasserstein metric. We have

$$\begin{aligned}
&\int_{\mathbb{R}^{2l}} \|\xi_1 - \xi_2\|^p Q(d\xi_1, d\xi_2) \\
&\quad = \sum_{i=1}^{K(\nu)} \int_{A_i(\eta)=\xi_1;\xi \in A} \left[\int_{\xi_2 \in A_i \cap \Pi_i^{-\nu}(\eta)} \|\xi_2 - H_i^\nu \Pi_i^\nu (\xi_2 - \beta_i)\|^p p(\xi_2) d\xi_2 \right] \\
&\quad \le \sum_{i=1}^{K(\nu)} \int_{A_i(\eta)=\xi_1;\xi \in A} \left[\int_{\xi_2 \in A_i^! \cap \Pi_i^{-\nu}(\eta)} \epsilon(\nu) + \int_{\xi_2 \in A_i^! \cap \Pi_i^{-\nu}(\eta)} \|\xi_2 - \beta_i)\|^p \right] p(\xi_2) d\xi_2 \\
&\quad \le \sum_{i=1}^{K(\nu)} \int_{A_i(\eta)=\xi_1;\xi \in A} [\epsilon_\nu(p_i) + \delta(\nu)(p_i)] q_i^\nu(\eta) d\eta \\
&\quad = \epsilon(\nu) + \delta(\nu).
\end{aligned} \tag{4.5}$$

This completes the proof. □

Using theorems 4.4 and 4.3 plus the Lipschitz continuity of the stochastic programming recourse function (see Wang [29]), we obtain convergence of Ψ values. With theorems 4.4 and 4.2, we also see that weak convergence is established and hence convergence of subdifferentials as in [3, theorem 3.1]. Moreover, depending on the properties of the distributions, we also obtain Lipschitz continuity of directional derivatives and hence convergence rates for gradient convergence as in theorem 3.1 for truncation approximations. As mentioned above, the advantage of this approximation concerns the maintenance of differentiability and improved convergence over discrete approximations. If we, for example, suppose the strong convexity conditions used in Römisch and Schultz [27], then we can use the random generation of β_i in the null space of some matrix H such that η is independent of ζ as above to obtain an expected Hausdorff distance between minimizers with P and P^ν that is $O(n^{-1/(2K)})$ where $K = \text{Nullity}(H)$ instead of l as with the discrete empirical measure.

To realize the differentiability result, we suppose the simplest case of a single H^ν and generation of β_i in the null space of H^ν. In this case, we have the following result.

THEOREM 4.5

If P^ν is generated as in (4.2) and satisfies (i) and (ii) where $H^\nu = H$ for all ν, $H \in \mathbb{R}^{l \times r}$, rank $H = r$, $H\beta_i = 0$ for all i, and no row of $B_j^{-1}H$ is a zero vector for any optimal basis in (1.2), B_j, $j = 1, \ldots, N$, then $\Psi^\nu(x)$ is differentiable for all ν and $x \in \text{int}(\text{dom}\, \Psi)$.

Proof

Suppose that $\Psi^\nu(x)$ is not differentiable for some $x \in \text{int}(\text{dom}\, \Psi)$. Then, there exists some i, j, k such that $Q_i^\nu(\eta | \eta \in C) > 0$ where $C = \{\eta | B_k^{-1}(Tx - H\eta - \beta_i) \geq 0, B_j^{-1}(Tx - H\eta - \beta_i) \geq 0$ for optimal bases B_k and B_j with $k \neq j\}$. Since Q_i^ν is a continuous distribution, this is only possible if the dimension of this set is r. For B_k and B_j both to be optimal, for all $\eta \in C$, there must exist some row t such that $(B_k^{-1})_t.(Tx - H\eta - \beta_i) = 0$. Since the dimension of C is r,

$$\dim(\eta | (B_k^{-1})_t.(H\eta) = (B_k^{-1})_t.(Tx - \beta_i)) = r. \tag{4.6}$$

It follows from (4.6) that $(B_k^{-1})_t.H = 0$, completing the proof. □

The result in theorem 4.5 gives a simple way to check whether differentiability is maintained. A limited basis approximation can also be applied in conjunction with this approach to guarantee differentiability for that approximation without requiring that all optimal bases be available.

We can see further advantages of differentiability by examining the structure of the approximate objective function gradient, $\nabla\Psi^\nu$. We can write this as:

$$\nabla\Psi^\nu(x) = \sum_{1 \leq i \leq K(\nu)} \sum_{i \leq j \leq N} p_i \alpha_j(\beta_i) \pi_j T, \tag{4.7}$$

where $\alpha_j(\beta_i) = Q_i^\nu(\eta | \pi_j$ is optimal in the dual of (1.2) for $w = Tx - H_i^\nu \eta - \beta_i)$. We write the gradient in this manner to show that it is an expectation of α_j values taken at different observations, β_i. The true value is the expectation over all possible β_i. This observation can allow the use of numerical integration ideas in the choice of the β_i. With differentiability established, we may also ask for higher order derivatives based now on the properties of the q^ν density functions. With appropriate distributions, we may then use the Peâno kernel to bound our approximation and, in the event of a polynomial density, to guide the choice of β_i to establish rapid convergence.

The lower dimensional approximations can also be combined with the separable function approaches as in Birge and Wets [6] and Birge and Wallace [5]. These approaches create an upper bounding function on ψ that is separable in its components. For example, suppose that ψ can be decomposed into $\psi_1(\eta)$ and $\psi_2(\zeta)$. The separable function approach is to create an upper bounding function ϕ such that $\phi(\eta)$ is for example $\sum_i \phi(\eta_i)$. Computing gradients of ϕ then involves only finding the relative probabilities of different intervals of η_i individually. In this way, we can quickly calculate $\alpha_j(\beta_i)$ in (4.7).

The goal in lower dimensional approximations may then be to find transformations of independent random variables η and ζ that yield ξ and to choose η so that the recourse function is most nearly separable in its components. This ability requires some knowledge about the specific problem but may be possible in certain instances. Note that if the recourse function ψ is also separable in ζ then we should observe that only a single $\alpha_j(\beta_i)$ needs to be calculated and then only updated with the relevant probability p_i.

5. Combined approximation

In the last two sections, we discussed approaches to approximate a continuous distribution by a simpler continuous distribution. In this section, we discuss the approach to approximate a continuous distribution by a combination of several simple distributions. The combined approximation is of the form

$$P^\nu(d\xi) = \sum_{1 \le j \le N_\nu} \lambda_j^\nu P_j^\nu(d\xi), \tag{5.1}$$

where $\lambda_j^\nu \ge 0$, $\sum \lambda_j^\nu = 1$. For brevity of symbols, we use λ_j and P_j instead of λ_j^ν and P_j^ν though they actually depend upon ν. By (1.3), (1.5) and (5.1), $\Psi(x)$ and $\nabla\Psi(x)$ are now approximated by

$$\Psi^\nu(x) = \sum_{1 \le j \le N_\nu} \lambda_j \Psi_j(x) \tag{5.2}$$

and

$$\nabla\Psi^\nu(x) = \sum_{1 \le j \le N_\nu} \lambda_j \nabla\Psi_j(x), \tag{5.3}$$

where

$$\Psi_j(x) = \int \psi(Tx - \xi) P_j(d\xi) \tag{5.4}$$

and

$$\nabla \Psi_j(x) = \int \nabla \psi(Tx - \xi) P_j(d\xi). \tag{5.5}$$

P_j should be simple such that the right-hand sides of (5.4) and (5.5) are easy to calculate. Depending upon the choices of P_j, we have different variants of combined approximations:

(1) *Discrete approximation*: Let

$$P_j(d\xi) = \begin{cases} 1, & \text{if } \xi = \xi^j, \\ 0, & \text{otherwise,} \end{cases}$$

where $\xi^1, \ldots, \xi^{N_\nu}$ are points in $\mathbb{R}^l$. Then we have a discrete approximation, which has been discussed intensively in the literature of stochastic programming.

(2) *Combined box approximation*: Let

$$P_j(d\xi) = \begin{cases} 1/(2\epsilon_\nu)^l, & \text{if } |\xi - \xi^j| \leq \epsilon_\nu, \\ 0, & \text{elsewhere.} \end{cases}$$

Then

$$\nabla \Psi_j(x) = \int_{|\xi - \xi^j| \leq \epsilon_\nu} \nabla \psi(Tx - \xi) d\xi.$$

Here, $\nabla \psi(Tx - \xi)$ only assumes a small number of values of $\{\pi_1, \ldots, \pi_N\}$, yet $\nabla \Psi_j$ is continuous.

(3) *Combined normal distribution approximation*: Let P_j be a multi-variable normal distribution such that the expectations, variances and covariances of $\{\xi_1, \ldots, \xi_l\}$ are specified. Then we may use theorem 2.2 to estimate $\nabla \Psi_j(x)$.

6. Examples

In this section, we consider some specific examples and demonstrate how the basic procedures can be implemented. The two examples will be a simple two-dimensional problem to demonstrate the geometry and a slightly larger problem with three random variables that comes from a power planning problem (see Louveaux and Smeers [22]). The methods are chosen to be representative and to give some idea of the efficiency of the continuous approximations presented

above. In each case, we consider the approximation of the recourse (second-stage) problem.

The first example has the following form:

$$\psi(w) = \left\{ \begin{array}{llllll} \min & 5y_1 & +10|y_2| & +10|y_3| & & \\ \text{s.t.} & y_1 & +y_2 & & = \xi_1 - x_1, & \\ & y_1 & & +y_3 & = \xi_2 - x_2, & \\ & y_1 & & & \geq 0, & \end{array} \right\} \tag{6.1}$$

where $-w = \xi - Tx$, $T = (1,1)^{\mathrm{T}}$. In the computational tests below, we assume that $x_1 = x_2 = 0$ and the random variables ξ_i are independently uniformly distributed on $[-0.5, 1.5]$.

The second example is the recourse function described in Louveaux and Smeers [22]. The first period variables x correspond to investments made into capacity available in the second period. We assume the second period demands in all three different sectors are random. (The model in [22] has only one random demand.) This example has the following form:

$$\psi(w) = \left\{ \begin{array}{llllll} \min & 40y_1 & +45y_2 & +32y_3 & +55y_4 & \\ & +24y_5 & +27y_6 & +19.2y_7 & +33y_8 & \\ & +4y_9 & +2.5y_{10} & +3.2y_{11} & +5.5y_{12} & \\ \text{s.t.} & y_1 & +y_2 & +y_3 & +y_4 & \geq \xi_1, \\ & y_5 & +y_6 & +y_7 & +y_8 & \geq \xi_2, \\ & y_9 & +y_{10} & +y_{11} & +y_{12} & \geq \xi_3, \\ & y_1 & +y_5 & +y_9 & & \leq x_1, \\ & y_2 & +y_6 & +y_{10} & & \leq x_2, \\ & y_3 & +y_7 & +y_{11} & & \leq x_3, \\ & y_4 & +y_8 & +y_{12} & & \leq x_4, \\ & y_i & & & \geq 0, & i = 1, \ldots, 12, \end{array} \right\} \tag{6.2}$$

where $-w$ again represents the right-hand side coefficient vector with

$$T = \begin{pmatrix} 0 \\ -I \end{pmatrix},$$

where 0 is a 3×4 matrix and I is a four dimensional identity matrix. The x_i correspond to input capacities while the ξ_i random variables correspond to demands in the three operating modes. In the computational tests below, we assume that $x_1 = 2$,

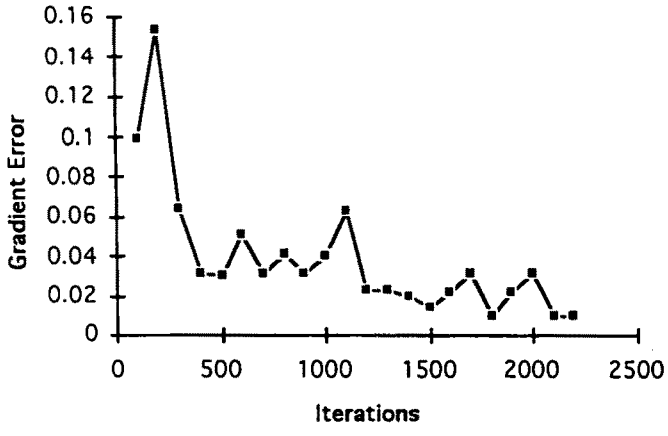

Fig. 1. Simulation errors for example 6.1.

$x_2 = 5$, $x_3 = 5$, $x_4 = 6$, and the random variables ξ_i are independently uniformly distributed on $[3, 7]$ for $i = 1$, on $[2, 6]$ for $i = 2$, and on $[1, 5]$ for $i = 3$. For $i = 4$, 5, 6, 7, we can interpret ξ_i as having zero values with probability one.

We will compare the following techniques:

(i) *Sampling distribution*;

(ii) *Refinements of Jensen and Edmundson–Madansky bounds*;

(iii) *Boole–Bonferroni probability approximation*;

(iv) *Box truncation*;

(v) *Limited basis truncation*;

(vi) *Lower-dimensional approximation.*

In each example, for an approximation represented by Ψ^ν, we consider the gradient error, $\|\nabla\Psi^\nu - \nabla\Psi\|$. This gradient error was used as a metric because our main goal in this exercise is in evaluating gradient-based methods.

(i) *Sampling distribution*: Here we sample the random vector. In our tests, since we had low dimensions, we used a low discrepancy quasi-random sequence to choose the samples. We followed the Hammersley sequence (see Fox [14], Deák [11]) which has been shown to produce small errors in low dimensions.

The errors from using the quasi-random sample appear in figure 1 for example 6.1 and in figure 2 for example 6.2 as functions of the number of observations. Figure 2 also includes the error function from using a lower-dimensional sample described below. Note that the errors fluctuate, indicating some difficulty in providing coverage with a sampling procedure. An error of 0.01 is approximately one percent of $\|\nabla\Psi\|$ in example 6.1. An error of 0.1 is approximately one percent of $\|\nabla\Psi\|$ in example 6.2. Note that this level of accuracy is achieved in about 1000 iterations for the example 6.2 although the lower dimensional example 6.1 requires

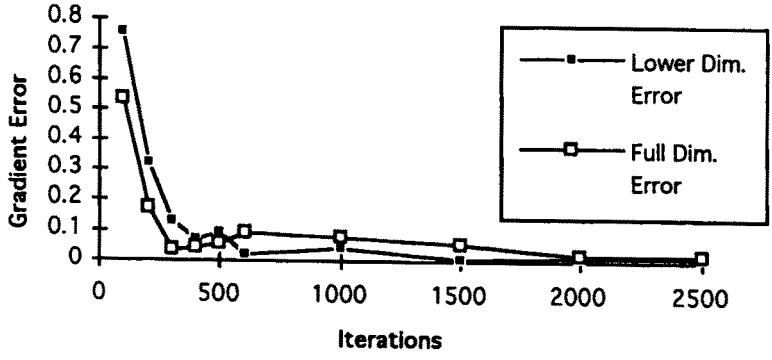

Fig. 2. Simulation errors using full and lower dimension for example 6.2.

more than 2000 iterations to achieve a similar percentage error. The difference is attributable to higher overall variance in the sample gradients in example 6.1.

(ii) *Refinements of Jensen and Edmundson–Madansky bounds*: These bounds result from extremal measures in the space of measures satisfying the same first moment condition as the true distribution (see [6]). We consider the refinements suggested in Frauendorfer and Kall [16] to observe their convergence to the optimal objective value and their effectiveness in predicting the gradient values. These bounds will be taken as examples of bounds using discrete measures although improvements exists (see [8]).

The gradient errors from the E–M and Jensen bound approximations appear in figure 3 for example 6.1 and figure 4 for example 6.2. Here, iterations refer to the number of partitions used for these approximations. The errors in function value in each case are less than 0.5% of $\Psi(x)$ after 40 iterations, but gradient errors may still

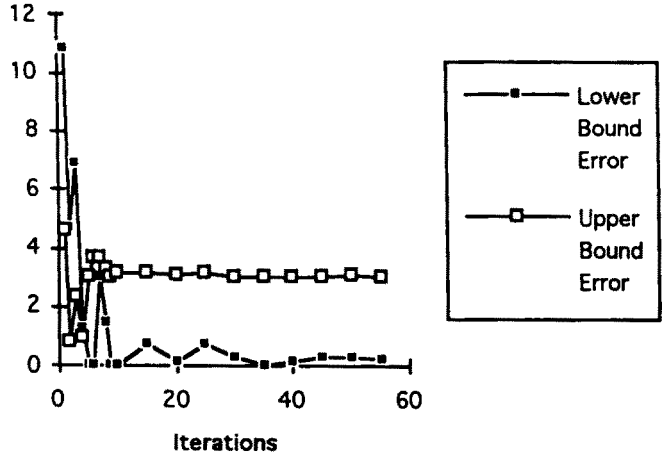

Fig. 3. Bound errors for example 6.1.

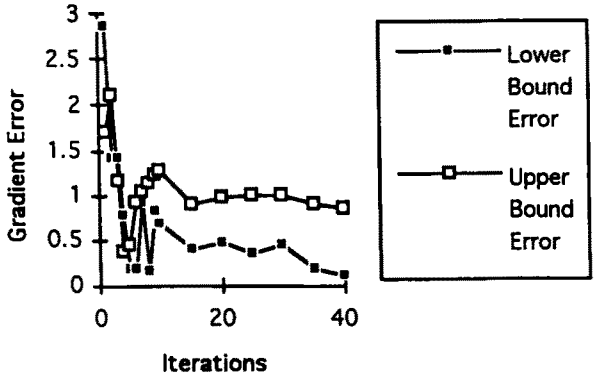

Fig. 4. Bound errors for example 6.2.

remain high. The main reason for this, in terms of the upper bound in particular, is that degenerate bases are chosen at extreme points. Their incorporation into the bound may always lead to some error. The lower bound estimate does not suffer this effect so dramatically (because indeed these estimates converge to true values, see [4]), but evaluations at nondifferentiable points still affect the bound. The overall observation here is that the bounding approximations are not indicated for gradient-based methods although they still appear quite useful in cutting plane methods (see, e.g., [1]).

(iii) *Probability approximation*: For this approach, we use the Boole–Bonferroni method to approximate the probability of a basis's optimality. This approximation is not useful for the first problem, with only two variables, but example 6.2 contains 7 constraints so that evaluating the probability of each basis involves evaluating the probability of a region corresponding to the intersection of 7 half-spaces. Some preprocessing was required to find each optimal basis for the range of possible demand outcomes given. To obtain these, we use the following extreme point generation approach.

Step 1. Solve $\psi(Tx - \bar{\xi})$ to obtain an optimal basis, B_1 (also optimal as long as $(B_1)^{-1}(\xi - Tx) \geq 0$). Let $k = 1$, $r = 1$, $i = 1$.

Step 2. If $(B_1)_{r\cdot}^{-1}(\xi - Tx) < 0$ for any $\xi \in \Xi$, perform a dual simplex pivot step to drop the variable which is basic in row r for B_i (maintaining dual feasibility). Let the new basis be B^*. Otherwise, let $r = r + 1$, go to step 4.

Step 3. If $B^* \in \{B_1, \ldots, B_k\}$, let $r = r + 1$, go to step 4. Otherwise, go to step 5.

Step 4. If $r < l$, then let $i = i + 1$. If $i < k$, let $r = 1$, go to 2. If $i = k$, STOP – all optimal bases have been identified.

Step 5. If $\{\xi \in \Xi | (B^*)_{r\cdot}^{-1}(\xi \pm e_i\epsilon - Tx) \geq 0\} \neq \emptyset$ for $i = 1, \ldots, l$ and some $\epsilon > 0$, then let $B_{k+1} = B^*$, $k = k + 1$, $r = r + 1$, and go to step 4. Otherwise, let $r = r + 1$, and go to step 4.

The use of variations of ϵ in each component of ξ_i is used to ensure that B^* is feasible with some positive probability (assuming Ξ has full dimension) so that zero probability bases are not included. The above procedure identifies all such optimal bases for the linear program in (1.2) with $-w = \xi - Tx$ by ensuring that any infeasible component leads an additional iteration with the branch only ending when all adjacent dual feasible bases are either infeasible or have already been identified. Under degeneracy, several bases B^* may be adjacent to a given basis. We assume that a lexicographic ordering is used to avoid this difficulty. The result is then an unambiguous listing of the bases.

Given identification of the bases, the Boole–Bonferroni approach calculates bounds using the formula in proposition 2.1. In example 6.2, the α_j probabilities were calculated exactly using $t_j = 1$ for five of the six alternative optimal bases. For the other basis, the optimal $t_j = 0.75$. The result for the five bases with $t_j = 1$ is that it is sufficient to calculate α_j only using a_j and b_j, or there is no probability that three regions of the form, $\eta_{ji} > s_{ji}(x)$, intersect. It appears that performing calculations with combinations of up to three intersecting $\eta_{ji} > s_{ji}(x)$ regions yields good results. In example 6.2, all basis probabilities were calculated exactly using three (of a possible seven) terms of the inclusion–exclusion approach, i.e., $\alpha_j = 1 - a_j + b_j - \sum_{1 \le i < i' < i'' \le l} P(\eta_{ji} > s_{ji}(x), \eta_{ji'} > s_{ji'}(x), \eta_{ji''} > s_{ji''}(x))$.

If $t_j = 1$ was used for all bases, the gradient error was 0.66 or approximately 5% of $\|\nabla\Psi\|$. For $t_j = 0.5$, some basis probability estimates became negative. In general, it has been observed (see [23]) that the bound with $t_j = 1$ (due to Dawson and Sankoff) is much sharper than with $t_j = 0$, so that one might generally bias toward higher t_j values and use $t_j = 1$ in most instances.

Given the difficulty in choosing t_j universally, it appears best to consider the values of $\sum_{1 \le i < i' < i'' \le l} P(\eta_{ji} > s_{ji}(x), \eta_{ji'} > s_{ji'}(x), \eta_{ji''} > s_{ji''}(x))$ if this additional computation is not difficult and the distributions of the η_{ji} are known (as in the normal case). It appears that relatively small numbers of intersecting regions need to be considered to obtain fairly accurate bounds on a basis's probability of optimality. This is again consistent with Prékopa's observations.

The remaining difficulty in this probability approximation choice is the identification of optimal bases. This number may be too great for easy computation. For this reason, we consider methods for reducing the number of optimal bases considered. The next approaches do this explicitly.

(iv) *Box truncation*: We use the box approximation method by considering boxes centered at the mean value of each random vector and considering progressively larger fractions of the full range of each component of the random vector. The gradient errors, $\|\nabla\Psi^\nu - \nabla\Psi\|$, for example 6.1 as a function of the fraction of range in each component covered appear in figure 5. The corresponding errors for example 6.2 appear in figure 6. The range is covered symmetrically starting at the mean of each component.

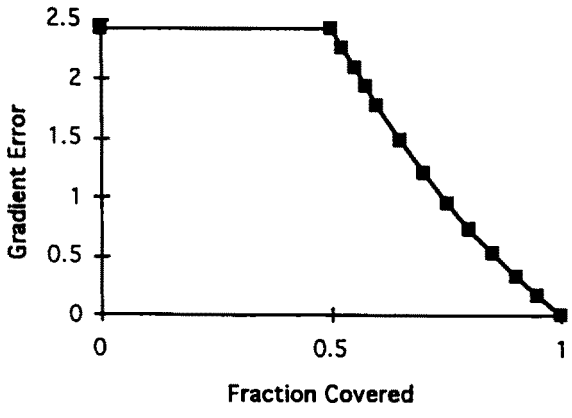

Fig. 5. Box truncation errors for example 6.1.

Note that the bounds do not improve until new bases beyond those immediately adjacent to the mean point are added. In each case, this improvement only begins when half of each component's region is considered. The error then improves almost linearly to zero.

The advantage of box truncation would be in conjunction with a basis probability approximation. For smaller regions, fewer bases are required so that less computation is necessary. In example 6.1, two bases are optimal for component fractions less than 0.5, but all five bases are optimal for fractions above 0.5. In example 6.2, four bases are optimal until the fraction covered in each component exceeds 0.5. At this point, five bases are optimal until the fraction exceeds 0.75, when all six optimal bases are included. The relatively small range of these numbers of bases indicates that box truncation may not have significant advantages over including all bases. In larger problems, however, the total number of optimal bases may make it practically impossible to include all bases. In that case, box truncation may offer some advantages.

(v) *Limited basis truncation*: We considered the sets of bases identified near the mean values as in the box truncation step. For example 6.1, using the two

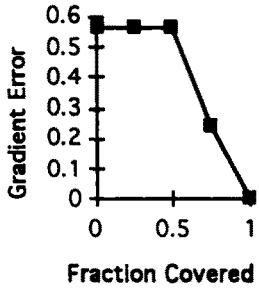

Fig. 6. Box truncation errors for example 6.2.

bases optimal at the mean $\bar{\xi} = (0.5, 0.5)^T$ yields the same error, 2.43, as in figure 5 for fractions less than 0.5. The only other consistent choice would be to include all optimal bases which results in zero error. In this case, limited bases do not appear too useful.

In example 6.2, the error using the optimal basis found at the mean $\bar{\xi}$ is 4.89 or more than 40% of $\|\nabla\Psi\|$. The error with the four neighboring bases is 1.09 or about 10% of $\|\nabla\Psi\|$. Note that this error is higher than the box truncation errors for any coverage fraction. This difference results from the lack of symmetry in including the complete region where each of these four bases is optimal. It appears from this observation that symmetry makes box truncation more effective than using full regions where a basis is optimal.

(vi) *Lower-dimensional approximation*: We suppose that the η vector from section 4 is one-dimensional so that all computations of $Q_i^\nu(B)$ only involve single integrals. In both examples, 6.1 and 6.2, we used the last random component for η and used the quasi-random sampling procedure to determine the A_i areas (implicitly defined as equal probability regions around each point, β_i).

For example 6.1, the use of this lower dimensional computation led to rapid convergence. Gradient errors were 0.78 after ten iterations, but became less than 0.001 after thirty iterations. This represents a considerable improvement over the results in figure 1. Some additional calculation was necessary for the calculation of each $Q_i^\nu(B)$ but this integration simply involved a single parametric linear program solution with little additional work (at most two pivot steps) over the linear program required for each function evaluation.

For example 6.2, the results of the lower dimensional approximation appear in figure 2 as mentioned earlier. The results here are less conclusive. The lower dimensional approximation does have somewhat lower error after 600 iterations (by approximately 50% over iterations from 600 to 2500), but improvement is not as dramatic as in the smaller example.

For larger problems, the modest improvements of the lower dimensional approach for a single gradient evaluation may still hold. One key to its effectiveness is in choosing the most critical component as the η variable. The other advantage of the lower dimensional approximation is that it is possible to maintain differentiability and to achieve better algorithm performance based on this attribute. This aspect of the approximations is a subject for future study.

Overall, our computational study indicates that lower dimensional sampling distributions, lower bounding and Boole–Bonferroni probability approximations, and some form of distribution support truncation may all be useful in providing efficient and reasonably accurate gradient estimates for stochastic programming algorithms. The results here indicate basic forms for these approximations and some characteristics on small problems. Each approach may be preferred in certain examples, but these results do indicate that degeneracy issues with upper bounding

approximations and symmetry issues with limited basis truncation seem to make these methods the least preferred for gradient estimation among the approaches given here.

7. Conclusions

This paper presented several results on using continuous distribution functions in approximations for stochastic programs. The main motivation is in providing differentiable approximate value functions that will lead to more stable computational implementations. The results are based on abilities to approximate probabilities accurately in higher dimensions and on using low dimension integration combined with discrete approximation to achieve differentiable but computable estimates. Some example results indicate that probability and gradient approximations can be accurate with separate low dimensional integrals. Future research will investigate these procedures in the context of optimization methods.

References

[1] J. Birge, Decomposition and partitioning methods for multi-stage stochastic linear programs, Oper. Res. 33 (1985) 989–1007.
[2] J.R. Birge and L. Qi, Computing block-angular Karmarkar projections with applications to stochastic programming, Manag. Sci. 34 (1988) 1472–1479.
[3] J.R. Birge and L. Qi, Semiregularity and generalized subdifferentials with applications for optimization, Math. Oper. Res. 18 (1993) 982–1005.
[4] J.R. Birge and L. Qi, Subdifferentials in approximation for stochastic programs, SIAM J. Optim., to appear.
[5] J.R. Birge and S.W. Wallace, PA separable piecewise linear upper bound for stochastic linear programs, SIAM J. Contr. Optim. 26 (1988) 725–739.
[6] J.R. Birge and R.J-B. Wets, Designing approximation schemes for stochastic optimization problems, in particular, for stochastic programs with recourse, Math. Progr. Study 27 (1986) 54–102.
[7] J.R. Birge and R.J-B. Wets, Computing bounds for stochastic programming problems by means of a generalized moment problem, Math. Oper. Res. 12 (1987) 149–162.
[8] J.R. Birge and R.J-B. Wets, Sublinear upper bounds for stochastic programs with recourse, Math. Progr. 43 (1989) 131–149.
[9] F.H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley, New York, 1983).
[10] I. Deák, Three-digit accurate multiple normal probabilities, Numer. Math. 35 (1980) 369–380.
[11] I. Deák, *Random Number Generators and Simulation* (Adadémiai Kiadó, Budapest, 1990).
[12] Y. Ermoliev, Stochastic quasigradient methods and their application to system optimization, Stochastics 9 (1983) 1–36.
[13] Y. Ermoliev and R. Wets, *Numerical Techniques in Stochastic Programming* (Springer, Berlin, 1988).
[14] B.L. Fox, Implementation and relative efficiency of quasirandom sequence generators, ACM Trans. Math. Software 12 (1986) 362–376.
[15] K. Frauendorfer, *Stochastic Two-Stage Programming*, Lecture Note Series (Springer, Berlin, 1992).

[16] K. Frauendorfer and P. Kall, A solution method for SLP recourse problems with arbitrary multivariate distributions – The independent case, Prob. Contr. Inf. Theory 17 (1988) 177–205.
[17] H.I. Gassmann, Conditional probability and conditional expectation of a random vector, in: *Numerical Techniques for Stochastic Optimization*, eds. Y. Ermoliev and R. Wets (Springer, Berlin, 1988) pp. 237–254.
[18] C.R. Givens and R.M. Shortt, A class of Wasserstein metrics for probability distributions, Michigan Math. J. 31 (1984) 231–240.
[19] J.L. Higle and S. Sen, Stochastic decomposition: An algorithm for two-stage stochastic linear programs with recourse, Math. Oper. Res. 16 (1991) 650–669.
[20] P. Kall, *Stochastic Linear Programming* (Springer, Berlin, 1976).
[21] P. Kall, Stochastic programming – An introduction, *6th Int. Conf. Stochastic Programming*, Udine, Italy (September, 1992).
[22] F.V. Louveaux and Y. Smeers, Optimal investments for electricity generation: A stochastic model and a test-problem, in: *Numerical Techniques for Stochastic Optimization*, eds. Y. Ermoliev and R. Wets (Springer, Berlin, 1988).
[23] A. Prékopa, Boole–Bonferroni inequalities and linear programming, Oper. Res. 36 (1988) 145–162.
[24] A. Prékopa and R.J-B. Wets, *Stochastic Programming 84*, Math. Progr. Study 27 & 28 (1986).
[25] S.T. Rachev, The Monge–Kantorovich mass transference problem and its stochastic applications, Theory Prob. Appl. 29 (1984) 647–676.
[26] W. Römisch and R. Schultz, Distribution sensitivity in stochastic programming, Math. Progr. 50 (1991) 197–226.
[27] W. Römisch and R. Schultz, Stability analysis for stochastic programs, Ann. Oper. Res. 31 (1991) 241–266.
[28] J. Wang, Distribution sensitivity analysis for stochastic programs with simple recourse, Math. Progr. 31 (1985) 286–297.
[29] J. Wang, Lipschitz continuity of objective functions in stochastic programs with fixed recourse and its applications, Math. Progr. Study 27 (1986) 145–152.
[30] R.J-B. Wets, Stochastic programming: Solution techniques and approximation schemes, in: *Mathematical Programming: The State of the Art – Bonn 1982* eds. A. Bachem, M. Grötschel and B. Korte (Springer, Berlin, 1983) pp. 566–603.