Some Recent Developments in Stochastic Programming

A. Shapiro

School of Industrial and Systems Engineering,
Georgia Institute of Technology,
Atlanta, Georgia 30332-0205, USA

Presented at: 8-th SIAM Optimization Conference, Stockholm, Sweden, May 2005.

The concept of two-stage (linear) stochastic programming problem with recourse

$$\operatorname*{Min}_{x \in X} c \cdot x + \mathbb{E}[Q(x, \boldsymbol{\xi})], \tag{1}$$

where $X = \{x : Ax = b,\ x \geq 0\}$, $c \cdot x$ denotes the scalar product of vectors $c, x \in \mathbb{R}^n$, and $Q(x, \xi)$ is the optimal value of the second stage problem

$$\operatorname*{Min}_{y} q^T y \text{ s.t. } Tx + Wy = h,\ y \geq 0, \tag{2}$$

with $\xi = (q, T, W, h)$. (By bold script like $\boldsymbol{\xi}$ we sometimes denote random variables to distinguish them from a particular realization ξ.) The feasible set X can be finite, i.e., integer first stage problem. Both stages can be integer (mixed integer) problems.

Suppose that the probability distribution P of $\boldsymbol{\xi}$ has a finite support, i.e., $\boldsymbol{\xi}$ can take values $\xi_1, ..., \xi_K$ (called scenarios) with respective probabilities $p_1, ..., p_K$. In that case

$$\mathbb{E}_P[Q(x,\boldsymbol{\xi})] = \sum_{k=1}^{K} p_k Q(x,\xi_k),$$

where

$$Q(x,\xi_k) = \inf\left\{q_k \cdot y_k : T_k x + W_k y_k = h_k,\ y_k \geq 0\right\}.$$

It follows that we can write problem (1)-(2) as one large linear program:

$$\begin{array}{ll} \underset{x,y_1,\dots,y_K}{\text{Min}} & c\cdot x + \sum_{k=1}^{K} p_k(q_k \cdot y_k) \\ \text{subject to} & Ax = b, \\ & T_k x + W_k y_k = h_k,\ k = 1,...,K, \\ & x \geq 0,\ y_k \geq 0,\ k = 1,...,K. \end{array}$$

Multistage models Let ξ_t be a random process. Denote $\xi_{[1,t]} = (\xi_1,..,\xi_t)$ the history of the process ξ_t up to time t. The values of the decision vector x_t, chosen at stage t, may depend on the information $\xi_{[1,t]}$ available up to time t, but not on the future observations. The decision process has the form

decision(x_0) $\rightsquigarrow$ observation(ξ_1) $\rightsquigarrow$ decision(x_1) $\rightsquigarrow$ ... $\rightsquigarrow$ observation(ξ_T) $\rightsquigarrow$ decision(x_T).

There are several ways how this decision process can be made precise. Nested formulation of a linear multistage stochastic programming problem with recourse:

$$\operatorname*{Min}_{\substack{A_1 x_1 = b_1 \\ x_1 \geq 0}} \; c_1 \cdot x_1 + \mathbb{E}\Bigg\{ \operatorname*{Min}_{\substack{B_2 x_1 + A_2 x_2 = b_2 \\ x_2 \geq 0}} c_2 \cdot x_2 + \cdots + \mathbb{E}\Bigg[\operatorname*{Min}_{\substack{B_T x_{T-1} + A_T x_T = b_T \\ x_T \geq 0}} c_T \cdot x_T \Bigg] \Bigg\}.$$

Here $\xi_t = (c_t, B_t, A_t, b_t)$, $t = 2, ..., T$, is considered as a random process, $\xi_1 = (c_1, A_1, b_1)$ is supposed to be known, and $x_t = x_t(\xi_{[1,t]})$, $t = 2, ..., T$, are supposed to be functions of the history of the process up to time t. If the number of realizations (scenarios) of the process ξ_t is finite, then the above problem can be written as one large linear programming problem. In that respect it is convenient to represent the random process in a form of scenario tree.

Dynamic programming equations. Consider the "last stage" linear program of the multi-stage problem:

$$\operatorname*{Min}_{x_T} c_T \cdot x_T \ \text{ s.t. } \ B_T x_{T-1} + A_T x_T = b_T,\ x_T \geq 0.$$

The optimal value of this problem depends on x_{T-1} and data $\xi_T = (c_T, B_T, A_T, b_T)$, and is denoted $Q_T(x_{T-1}, \xi_T)$. At stage $t = T-1, ..., 2$ we consider the optimal value $Q_t(x_{t-1}, \xi_{[1,t]})$ of the problem

$$\begin{array}{ll} \operatorname*{Min}_{x_t} & c_t \cdot x_t + \mathbb{E}\left[Q_{t+1}(x_t, \boldsymbol{\xi}_{[1,t+1]}) \Big| \boldsymbol{\xi}_{[1,t]} \right] \\ \text{s.t.} & B_t x_{t-1} + A_t x_t = b_t,\ x_t \geq 0. \end{array} \qquad (3)$$

At the first stage we obtain the problem:

$$\begin{array}{ll} \operatorname*{Min}_{x_1} & c_1 \cdot x_1 + \mathbb{E}\left[Q_2(x_1, \boldsymbol{\xi}_2) \right] \\ \text{s.t.} & A_1 x_1 = b_1,\ x_1 \geq 0. \end{array} \qquad (4)$$

The expectation in (3) is conditional on the history of the process $\boldsymbol{\xi}_t$ up to time t. If the process is Markovian, then this expectation depends only on ξ_t, and hence the cost-to-go function $Q_t(x_{t-1}, \xi_t)$ is a function of x_{t-1} and ξ_t.

- **How difficult is to solve two-stage problems?**

- **What about multistage problems?**

Consider stochastic optimization problem:

$$\operatorname*{Min}_{x\in X} \Big\{ f(x) := \mathbb{E}_P[F(x,\boldsymbol{\xi})] \Big\},$$

where $\boldsymbol{\xi}$ is a random vector having probability distribution P, $F(x,\xi)$ is a real valued function and $X \subset \mathbb{R}^n$. If P has a finite support (i.e., finite number $\xi_1, ..., \xi_K$ of scenarios), then

$$\mathbb{E}_P[F(x,\boldsymbol{\xi})] = \sum_{k=1}^{K} p_k F(x,\xi_k).$$

However, the number K of scenarios grows exponentially with dimension of the data $\boldsymbol{\xi}$. For example, if d random components of $\boldsymbol{\xi}$ are independent, each having just 3 possible realizations, then the total number of scenarios $K = 3^d$. No computer in a foreseeable future will be able to handle calculations involving 3^{100} scenarios.

Monte Carlo sampling approach

Let $\xi^1, ..., \xi^N$ be a generated (iid) random sample drawn from P and

$$\hat{f}_N(x) := N^{-1} \sum_{j=1}^{N} F(x, \xi^j)$$

be the corresponding sample average function. By the Law of Large Numbers, for a given $x \in X$, we have $\hat{f}_N(x) \to f(x) = \mathbb{E}_P[F(x, \boldsymbol{\xi})]$ w.p.1 as $N \to \infty$.

Notoriously slow convergence of order $O_p(N^{-1/2})$. By the Central Limit Theorem

$$N^{1/2}\left[\hat{f}_N(x) - f(x)\right] \Rightarrow N(0, \sigma^2(x)),$$

where $\sigma^2(x) := \mathrm{Var}[F(x, \boldsymbol{\xi})]$. In order to improve the accuracy by one digit the sample size should be increased 100 times. Good news: rate of convergence does not depend on the number of scenarios, only on the variance $\sigma^2(x)$. The accuracy can be improved by variance reduction techniques. However, the rate of the square root of N (of Monte Carlo sampling estimation) cannot be changed.

Monte Carlo sampling optimization

Two basic philosophies: interior and exterior Monte Carlo sampling. In interior sampling methods, sampling is performed inside a chosen algorithm with new (independent) samples generated in the process of iterations (e.g., Higle and Sen (stochastic decomposition), Infanger (statistical L-shape method), Norkin, Pflug and Ruszczynski (stochastic branch and bound method)).

In the exterior sampling approach the true problem is approximated by the sample average approximation problem:

$$\text{(SAA)} \qquad \operatorname*{Min}_{x \in X} \left\{ \hat{f}_N(x) := N^{-1} \sum_{j=1}^{N} F(x, \xi^j) \right\}.$$

Once the sample $\xi^1, ..., \xi^N \sim P$ is generated, the SAA problem becomes a deterministic optimization and can be solved by an appropriate algorithm.

Difficult to point out an exact origin of this method. Variants of this approach were suggested by a number of authors under different names.

Advantages of the SAA method:

• Ease of numerical implementation. Often one can use existing software.

• Good convergence properties.

• Well developed statistical inference: validation and error analysis, stopping rules.

• Easily amendable to variance reduction techniques.

• Ideal for parallel computations.

The idea of common random numbers generation. Suppose that $X = \{x_1, x_2\}$. Then the variance of $N^{1/2}\left[\hat{f}_N(x_1) - \hat{f}_N(x_2)\right]$ is:

$$\mathsf{Var}[F(x_1, \boldsymbol{\xi})] + \mathsf{Var}[F(x_2, \boldsymbol{\xi})] - 2\,\mathsf{Cov}[F(x_1, \boldsymbol{\xi}), F(x_2, \boldsymbol{\xi})]$$

It can be much smaller than

$$\mathsf{Var}[F(x_1, \boldsymbol{\xi})] + \mathsf{Var}[F(x_2, \boldsymbol{\xi})]$$

when the samples are independent.

Notation

v^0 is the optimal value of the true problem
S^0 is the optimal solutions set of the true problem
S^ε is the set of ε-optimal solutions of the true prob
$\hat{v}_N$ is the optimal value of the SAA problem
$\hat{S}_N^\varepsilon$ is the set of ε-optimal solutions of the SAA pro
$\hat{x}_N$ is an optimal solution of the SAA problem

Convergence properties

Vast literature on statistical properties of the SAA estimators $\hat{v}_N$ and $\hat{x}_N$:

Consistency. By the Law of Large Numbers, $\hat{f}_N(x)$ converge (pointwise) to $f(x)$ w.p.1. Under mild additional conditions, this implies that $\hat{v}_N \to v^0$ and $\mathrm{dist}(\hat{x}_N, S^0) \to 0$ w.p.1 as $N \to \infty$. In particular, $\hat{x}_N \to x^0$ w.p.1 if $S^0 = \{x^0\}$. (Consistency of Maximum Likelihood estimators, Wald (1949)).

Central Limit Theorem type results.

$$\hat{v}_N = \min_{x \in S^0} \hat{f}_N(x) + o_p(N^{-1/2}).$$

In particular, if $S^0 = \{x^0\}$, then

$$N^{1/2}[\hat{v}_N - v^0] \Rightarrow N(0, \sigma^2(x^0)).$$

These results suggest that the optimal value of the SAA problem converges at a rate of $\sqrt{N}$. In particular, if $S^0 = \{x^0\}$, then $\widehat{v}_N$ converges to v^0 at the same rate as $\widehat{f}_N(x^0)$ converges to $f(x^0)$.

If $S^0 = \{x^0\}$, then under certain regularity conditions, $N^{1/2}(\widehat{x}_N - x^0)$ converges in distribution. (Asymptotic normality of M-estimators, Huber (1967)).

The required regularity conditions are that the expected value function $f(x)$ is smooth (twice differentiable) at x^0 and the Hessian matrix $\nabla^2 f(x^0)$ is positive definite. This typically happens if the probability distribution P is continuous. In such cases $\widehat{x}_N$ converges to x^0 at the same rate as the stochastic approximation iterates calculated with the optimal step sizes (Shapiro, 1996).

Large Deviations type bounds.

Let $Y_1, ..., Y_N$ be an iid sequence of random variables, and $Z_N := \frac{1}{N}\sum_{j=1}^N Y_j$. Then (the upper bound of Cramér's LD theorem)

$$\mathbb{P}(Z_N \geq z) \leq \exp\Big\{-NI(z)\Big\}$$

for $z > \mu$, where $\mu := \mathbb{E}[Y]$ and

$$I(z) := \sup_{t\in\mathbb{R}} \Big\{tz - \Lambda(t)\Big\},$$

$\Lambda(t) := \log M(t)$ and $M(t) := \mathbb{E}[e^{tY}]$ is the moment generating function. Assume that $M(t)$ is finite valued for all t in a neighborhood of zero (if $M(t) = +\infty$ for any $t \neq 0$, then $I(z) \equiv 0$ and hence the above upper bound trivially holds). Note that $\Lambda(t)$ is convex, $\Lambda'(0) = \mu$ and $\Lambda''(0) = \sigma^2$, where $\sigma^2 := \text{Var}[Y]$. The rate function $I(z)$ is convex, $I'(\mu) = 0$, $I''(\mu) = \sigma^{-2}$, and $I(z)$ attains its minimum at $z = \mu$. Consequently,

$$I(z) = \frac{(z-\mu)^2}{2\sigma^2} + o(|z-\mu|^2),$$

and $I(z) > 0$ for any $z \neq \mu$. Note that if $Y \sim N(\mu, \sigma^2)$, then

$$M(t) = \exp\{\mu t + \sigma^2 t^2/2\}$$

and hence $I(z) = \frac{(z-\mu)^2}{2\sigma^2}$.

If

$$M(t) \leq \exp\{\mu t + \sigma^2 t^2/2\},$$

then $I(z) \geq \frac{(z-\mu)^2}{2\sigma^2}$ and hence for $z \geq \mu$,

$$\mathbb{P}(Z_N \geq z) \leq \exp\left\{\frac{-N(z-\mu)^2}{2\sigma^2}\right\}.$$

If distribution of Y is supported on a bounded interval $[a, b]$, then for $z \geq \mu$ (Hoeffding's inequality):

$$\mathbb{P}(Z_N \geq z) \leq \exp\left\{\frac{-N(z-\mu)^2}{(b-a)^2/2}\right\}.$$

Uniform LD bounds.

Let X be a subset of $\mathbb{R}^n$ of finite diameter $D := \sup_{x',x\in X} \|x' - x\|$, $\xi^1, ..., \xi^N$ be an iid sample from a distribution supported on $\Xi \subset \mathbb{R}^d$. Suppose that: (i) there is a constant $\sigma > 0$ such that

$$M_x(t) \leq \{\sigma^2 t^2/2\}, \;\; \forall t \in \mathbb{R}, \; \forall x \in X,$$

where $M_x(t)$ is the moment generating function of the random variable $F(x, \boldsymbol{\xi}) - f(x)$, (ii) there is a constant $L > 0$ such that

$$\Big|F(x', \xi) - F(x, \xi)\Big| \leq L\|x' - x\|, \;\; \forall \xi \in \Xi, \; \forall x', x \in X.$$

Then for any $\gamma > 0$,

$$\mathbb{P}\left\{\sup_{x\in X}\left|\hat{f}_N(x) - f(x)\right| \geq \gamma\right\} \\ \leq \left(\frac{O(1)DL}{\gamma}\right)^n \exp\left\{-\frac{N\gamma^2}{16\sigma^2}\right\}.$$

Note that if $\hat{x}_N$ is a δ-optimal solution of the SAA problem and $\sup_{x\in X}\left|\hat{f}_N(x) - f(x)\right| < \gamma$, then $\hat{x}_N$ is an ε-optimal solution of the true problem with $\varepsilon = \gamma + \delta$. Now choose $\varepsilon > 0$, $\delta \in [0, \varepsilon)$ and $\alpha \in (0, 1)$. Then, for $\gamma := \varepsilon - \delta$, by solving

$$\left(\frac{O(1)DL}{\gamma}\right)^n \exp\left\{-\frac{N\gamma^2}{16\sigma^2}\right\} \leq \alpha,$$

we obtain that for sample size

$$N \geq \frac{O(1)\sigma^2}{(\varepsilon - \delta)^2}\left[n \log\left(\frac{O(1)DL}{(\varepsilon - \delta)^2}\right) + \log\left(\frac{1}{\alpha}\right)\right]$$

we are guaranteed that with probability at least $1-\alpha$ any δ-optimal solution of the SAA problem is an ε-optimal solution of the true problem.

If the set X is finite, then the corresponding estimate of the sample size becomes (under the assumption (i)):

$$N \geq \frac{2\sigma^2}{(\varepsilon - \delta)^2} \log \left(\frac{|X|}{\alpha} \right).$$

Example Let $F(x, \xi) := \|x\|^{2k} - 2k \left(\xi^T x \right)$, where $k \in \mathbb{N}$ and

$$X := \{x \in \mathbb{R}^n : \|x\| \leq 1\}.$$

Suppose, that $\xi \sim N(0, \sigma^2 I_n)$. Then $f(x) = \|x\|^{2k}$, and for $\varepsilon \in [0, 1]$, the set of ε-optimal solutions of the true problem is

$$\{x : \|x\|^{2k} \leq \varepsilon\}.$$

Let $\bar{\xi}_N := (\xi^1 + ... + \xi^N)/N$. The corresponding sample average function is

$$\hat{f}_N(x) = \|x\|^{2k} - 2k \left(\bar{\xi}_N^T x \right),$$

and $\hat{x}_N = \|\bar{\xi}_N\|^{-\gamma} \bar{\xi}_N$, where $\gamma := \frac{2k-2}{2k-1}$ if $\|\bar{\xi}_N\| \leq 1$, and $\gamma = 1$ if $\|\bar{\xi}_N\| > 1$. Therefore, for $\varepsilon \in (0, 1)$, the optimal solution of the SAA problem is an ε-optimal solution of the true problem iff $\|\bar{\xi}_N\|^\nu \leq \varepsilon$, where $\nu := \frac{2k}{2k-1}$.

We have that $\bar{\xi}_N \sim N(0, \sigma^2 N^{-1} I_n)$, and hence $N\|\bar{\xi}_N\|^2/\sigma^2$ has the chi-square distribution with n degrees of freedom. Consequently, the probability that $\|\bar{\xi}_N\|^\nu > \varepsilon$ is equal to the probability

$$\mathbb{P}\left(\chi^2_n > N\varepsilon^{2/\nu}/\sigma^2\right).$$

Moreover, $\mathbb{E}[\chi^2_n] = n$ and the probability $\mathbb{P}(\chi^2_n > n)$ increases and tends to 1/2 as n increases. Consequently, for $\alpha \in (0, 0.3)$ and $\varepsilon \in (0, 1)$, for example, the sample size N should satisfy

$$N > \frac{n\sigma^2}{\varepsilon^{2/\nu}} \tag{5}$$

in order to have the property: "with probability $1 - \alpha$ an (exact) optimal solution of the SAA problem is an ε-optimal solution of the true problem". Note that $\nu \to 1$ as $k \to \infty$.

Suppose that the true problem is convex piecewise linear. That is:
(i) the distribution P has a finite support,
(ii) for almost every ξ the function $F(\cdot,\xi)$ is piecewise linear and convex,
(iii) the feasible set X is polyhedral (i.e., is defined by a finite number of linear constraints).

Suppose also that the optimal solutions set S^0, of the true problem, is nonempty and bounded.

Then :
(1) W.p.1 for N large enough, $\hat{x}_N$ is an *exact* optimal solution of the true problem. More precisely, w.p.1 for N large enough, the set $\widehat{S}_N$ of optimal solutions of the SAA problem is nonempty and forms a face of the (polyhedral) set S^0.

(2) Probability of the event $\{\widehat{S}_N \subset S^0\}$ tends to one *exponentially fast*. That is, there exists a constant $\gamma > 0$ such that

$$\lim_{N\to\infty} \frac{1}{N} \log\left[1 - P(\widehat{S}_N \subset S^0)\right] = -\gamma.$$

(Shapiro & Homem-de-Mello, 2000)

The idea of repeated solutions.

Solve the SAA problem M times using M independent samples each of size N. Let $\hat{v}_N^{(1)}, ..., \hat{v}_N^{(M)}$ be the optimal values and $\hat{x}_N^{(1)}, ..., \hat{x}_N^{(M)}$ be optimal solutions of the corresponding SAA problems. Probability that at least one of $\hat{x}_N^{(i)}$, $i = 1, ..., M$ is an optimal solution of the true problem is $1 - p_N^M$ where

$$p_N := P(\hat{x}_N \neq x^0) \approx C N^{-1/2} e^{-N\gamma}.$$

and hence

$$p_N^M \approx (C N^{-1/2})^M e^{-NM\gamma}.$$

Cutting plane (Benders cuts, L-shape) type algorithms. Empirical observation: on average the number of iterations (cuts) does not grow, or grows slowly, with increase of the sample size N. From theoretical point of view it converges to the respective number of the true problem.

Validation analysis

How one can evaluate quality of a given solution $\hat{x} \in S$? Two basic approaches:

(1) Evaluate the gap $f(\hat{x}) - v^0$.

(2) Verify the KKT optimality conditions at $\hat{x}$.

Statistical test based on estimation of $f(\hat{x}) - v^0$ (Norkin, Pflug & Ruszczynski 98, Mak, Morton & Wood 99):

(i) Estimate $f(\hat{x})$ by the sample average $\hat{f}_{N'}(\hat{x})$, using sample of a large size N'.

(ii) Solve the SAA problem M times using M independent samples each of size N. Let $\hat{v}_N^{(1)}, ..., \hat{v}_N^{(M)}$ be the optimal values of the corresponding SAA problems. Estimate $\mathbb{E}[\hat{v}_N]$ by the average $M^{-1} \sum_{j=1}^M \hat{v}_N^{(j)}$. Note that

$$\mathbb{E}\left[\hat{f}_{N'}(\hat{x}) - M^{-1} \sum_{j=1}^M \hat{v}_N^{(j)}\right] = \left(f(\hat{x}) - v^0\right) + \left(v^0 - \mathbb{E}[\hat{v}_N]\right),$$

and that $v^0 - \mathbb{E}[\hat{v}_N] > 0$. For ill-conditioned problems the bias $v^0 - \mathbb{E}[\hat{v}_N]$ can be large.

The bias $v^0 - \mathbb{E}[\hat{v}_N]$ is positive and (under mild regularity conditions)

$$\lim_{N\to\infty} N^{1/2}\left(v^0 - \mathbb{E}[\hat{v}_N]\right) = \mathbb{E}\left[\max_{x\in S^0} Y(x)\right],$$

where $(Y(x_1),...,Y(x_k))$ has a multivariate normal distribution with zero mean vector and covariance matrix given by the covariance matrix of the random vector $(F(x_1,\boldsymbol{\xi}),...,F(x_k,\boldsymbol{\xi}))$. For ill-conditioned problems this bias is of order $O(N^{-1/2})$ and can be large if the ε-optimal solution set S^ε is large for some small $\varepsilon \geq 0$.

Common random numbers variant: generate a sample (of size N) and calculate the gap

$$\hat{f}_N(\hat{x}) - \inf_{x\in X} \hat{f}_N(x).$$

Repeat this procedure M times (with independent samples), and calculate the average of the above gaps. This procedure works well for well conditioned problems, does not improve the bias problem.

KKT statistical test Let

$$X := \{x \in \mathbb{R}^n : c_i(x) = 0,\ i \in I,\ c_i(x) \le 0,\ i \in J\}.$$

Suppose that the probability distribution is continuous. Then $F(\cdot, \boldsymbol{\xi})$ is differentiable at $\hat{x}$ w.p.1 and

$$\nabla f(\hat{x}) = \mathbb{E}_P \left[\nabla_x F(\hat{x}, \boldsymbol{\xi})\right].$$

KKT-optimality conditions at an optimal solution $x^0 \in S^0$ can be written as follows:

$$-\nabla f(x^0) \in C(x^0),$$

where

$$C(x) := \left\{ y = \sum_{i \in I \cup J(x)} \lambda_i \nabla c_i(x),\ \lambda_i \ge 0,\ i \in J(x) \right\},$$

and $J(x) := \{i : c_i(x) = 0,\ i \in J\}$. The idea of the KKT test is to estimate the distance

$$\delta(\hat{x}) := \operatorname{dist}\left(-\nabla f(\hat{x}), C(\hat{x})\right),$$

by using the sample estimator

$$\hat{\delta}_N(\hat{x}) := \operatorname{dist}\left(-\nabla \hat{f}_N(\hat{x}), C(\hat{x})\right).$$

The covariance matrix of $\nabla \hat{f}_N(\hat{x})$ can be estimated (from the same sample), and hence a confidence region for $\nabla f(\hat{x})$ can be constructed. This allows a statistical validation of the KKT conditions. (Shapiro & Homem-de-Mello 98).

Complexity of multistage stochastic programming

Consider the following T–stage stochastic programming problem

$$\operatorname*{Min}_{x_1 \in \mathcal{G}_1} F_1(x_1) + \mathbb{E}\Big[\inf_{x_2 \in \mathcal{G}_2(x_1,\xi_2)} F_2(x_2,\xi_2) + \mathbb{E}\Big[\cdots + \mathbb{E}\left[\inf_{x_T \in \mathcal{G}_T(x_{T-1},\xi_T)} F_T(x_T,\xi_T) \right]\Big]\Big]$$

driven by the random data process $\xi_2, ..., \xi_T$. Here $x_t \in \mathbb{R}^{n_t}$, $t = 1, ..., T$, are decision variables, $F_t : \mathbb{R}^{n_t} \times \mathbb{R}^{d_t} \to \mathbb{R}$ are continuous functions and $\mathcal{G}_t : \mathbb{R}^{n_{t-1}} \times \mathbb{R}^{d_t} \rightrightarrows \mathbb{R}^{n_t}$, $t = 2, ..., T$, are measurable multifunctions, the function $F_1 : \mathbb{R}^{n_1} \to \mathbb{R}$ and the set $\mathcal{G}_1 \subset \mathbb{R}^{n_1}$ are deterministic. We assume that the set $\mathcal{G}_1$ is nonempty. For example, in linear case $F_t(x_t, \xi_t) := \langle c_t, x_t \rangle$, $\mathcal{G}_1 := \{x_1 : A_1 x_1 = b_1,\ x_1 \geq 0\}$,

$$\mathcal{G}_t(x_{t-1}, \xi_t) := \{x_t : B_t x_{t-1} + A_t x_t = b_t,\ x_t \geq 0\},$$

$\xi_1 := (c_1, A_1, b_1)$ is known at the first stage (and hence is nonrandom), and $\xi_t := (c_t, B_t, A_t, b_t)$, $t = 2, ..., T$, are data vectors some (all) elements of which can be random.

Conditional sampling: let ξ_2^i, $i = 1, ..., N_1$, be an iid random sample of $\boldsymbol{\xi}_2$. Conditional on $\boldsymbol{\xi}_2 = \xi_2^i$, a random sample ξ_3^{ij}, $j = 1, ..., N_2$, is generated and etc. The obtained scenario tree is considered as a sample average approximation of the true problem. Note that the total number of scenarios $N = \prod_{t=1}^{T-1} N_t$.

For $T = 3$, under certain regularity conditions, for $\varepsilon > 0$ and $\alpha \in (0, 1)$, and the sample sizes N_1 and N_2 satisfying

$$O(1)\Bigg[\left(\frac{D_1 L_1}{\varepsilon}\right)^{n_1} \exp\left\{-\frac{O(1)N_1\varepsilon^2}{\sigma_1^2}\right\} + \left(\frac{D_1 L_3}{\varepsilon}\right)^{n_1} \left(\frac{D_2 L_2}{\varepsilon}\right)^{n_2} \exp\left\{-\frac{O(1)N_2\varepsilon^2}{\sigma_2^2}\right\}\Bigg] \le \alpha,$$

we have that any $\varepsilon/2$-optimal solution of the SAA problem is an ε-optimal solution of the true problem with probability at least $1 - \alpha$.

In particular, suppose that $N_1 = N_2$. Then the required sample size $N_1 = N_2$:

$$N_1 \ge \frac{O(1)\sigma^2}{\varepsilon^2}\left[(n_1 + n_2)\log\left(\frac{DL}{\varepsilon}\right) + \log\left(\frac{O(1)}{\alpha}\right)\right].$$

Risk Analysis

Min-max approach to stochastic programming:

$$\operatorname*{Min}_{x \in X} \left\{ f(x) := \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[F(x,\omega)] \right\},$$

where $\mathcal{A}$ is a set of probability measures (distributions). This approach was already considered in Žáčková (1966).

Optimization of mean-risk models:

$$\operatorname*{Min}_{x \in X} \rho[F_x(\omega)],$$

where $\rho : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is a risk function, $\mathcal{Z}$ is a (linear) space of "allowable" functions $Z(\omega)$ and $F_x(\cdot) \in \mathcal{Z}$ for all $x \in X$

Markowitz's approach:

$$\rho(Z) := \mathbb{E}[Z] + c\,\mathsf{Var}[Z], \ \ Z \in \mathcal{Z},$$

where $c > 0$ is a weight constant.

Axiomatic approach (coherent measures of risk), by Artzner, Delbaen, Eber, Heath (1999):
(A1) Convexity:

$$\rho(\alpha Z_1 + (1-\alpha)Z_2) \le \alpha\rho(Z_1) + (1-\alpha)\rho(Z_2)$$

for all $Z_1, Z_2 \in \mathcal{Z}$ and $\alpha \in [0,1]$.
(A2) Monotonicity: If $Z_1, Z_2 \in \mathcal{Z}$ and $Z_2 \ge Z_1$, then $\rho(Z_2) \ge \rho(Z_1)$.
(A3) Translation Equivariance: If $a \in \mathbb{R}$ and $Z \in \mathcal{Z}$, then $\rho(Z + a) = \rho(Z) + a$.
(A4) Positive Homogeneity:

$$\rho(\alpha Z) = \alpha\rho(Z), \;\; Z \in \mathcal{Z}, \; \alpha > 0.$$

Space $\mathcal{Z}$ is paired with a linear space $\mathcal{Y}$ of finite signed measures on $(\Omega, \mathcal{F})$ such that the scalar product (bilinear form)

$$\langle \mu, Z \rangle := \int_\Omega Z(\omega) d\mu(\omega)$$

is well defined for all $Z \in \mathcal{Z}$ and $\mu \in \mathcal{Y}$. Typical examples $\mathcal{Z} := L_p(\Omega, \mathcal{F}, P)$ and $\mathcal{Y} := L_q(\Omega, \mathcal{F}, P)$, where $p, q \in [1, +\infty]$ such that $1/p + 1/q = 1$, and P is a probability (reference) measure on $(\Omega, \mathcal{F})$.

Dual Representation of Risk Functions

By Fenchel-Moreau theorem if ρ is convex (assumption (A1)) and lower semicontinuous, then

$$\rho(Z) = \sup_{\mu \in \mathcal{A}} \{\langle \mu, Z \rangle - \rho^*(\mu)\} \ ,$$

where

$$\rho^*(\mu) = \sup_{Z \in \mathcal{Z}} \{\langle \mu, Z \rangle - \rho(Z)\} \, ,$$
$$\mathcal{A} := \mathrm{dom}(\rho^*) = \{\mu \in \mathcal{Y} : \rho^*(\mu) < +\infty\} \, .$$

Now, condition (A2) (monotonicity) holds iff $\mu \succeq 0$ for every $\mu \in \mathcal{A}$. Condition (A3) (translation equivariance) holds iff $\mu(\Omega) = 1$ for every $\mu \in \mathcal{A}$. If ρ is positively homogeneous, then $\rho^*(\mu) = 0$ for every $\mu \in \mathcal{A}$. If conditions (A1)–(A4) hold, then $\mathcal{A}$ is a set of probability measures and

$$\rho(Z) = \sup_{\mu \in \mathcal{A}} \mathbb{E}_\mu[Z].$$

Consequently, problem $\mathrm{Min}_{x \in X} \, \rho[F(x, \omega)]$ is equivalent to the min-max problem.

In various forms the dual representation was derived in Artzner, Delbaen, Eber, Heath (1999), Föllmer and Schied (2002), Rockafellar, Uryasev and Zabarankin (2003), Ruszczyński and Shapiro (2004).

Example Mean-variance risk function ($c > 0$):

$\rho(Z) := \mathbb{E}[Z] + c\,\mathsf{Var}[Z], \;\; Z \in \mathcal{Z} := L_2(\Omega, \mathcal{F}, P)$.

Dual representation:

$$\rho(Z) = \sup_{\zeta \in \mathcal{Z},\, \mathbb{E}[\zeta]=1} \left\{ \langle \zeta, Z\rangle - (4c)^{-1}\mathsf{Var}[\zeta] \right\}.$$

Satisfies conditions (A1) and (A3), does not satisfy (A2) and (A4).

Example Mean-upper-semideviation risk function of order $p \in [1, +\infty)$:

$$\rho(Z) := \mathbb{E}[Z] + c\,\psi_p(Z),\; Z \in \mathcal{Z} := L_p(\Omega, \mathcal{F}, P),$$

$c \geq 0$ and

$$\psi_p(Z) := \left(\mathbb{E}_P \left\{ \left[Z - \mathbb{E}_P[Z] \right]_+^p \right\} \right)^{1/p}.$$

Then ρ satisfies (A1),(A3),(A4), and also (A2) (monotonicity) if $c \leq 1$. The max-representation

$$\rho(Z) = \sup_{\zeta \in \mathcal{A}} \int_\Omega Z(\omega)\zeta(\omega)dP(\omega)$$

holds with

$$\mathcal{A} = \left\{ \zeta : \zeta = 1 + h - \int_\Omega h dP,\; \|h\|_q \leq c,\; h \geq 0 \right\}.$$

Recall that the (primal) optimizatiom problem can be written in the form

$$\operatorname*{Min}_{x\in X}\sup_{\mu\in\mathcal{A}}\mathbb{E}_\mu[F(x,\omega)].$$

A point $(\bar{x},\bar{\mu})\in X\times\mathcal{A}$ is a saddle point of the above problem if

$$\bar{x}\in\arg\min_{x\in X}\mathbb{E}_{\bar{\mu}}[F(x,\omega)]$$

and

$$\bar{\mu}\in\arg\max_{\mu\in\mathcal{A}}\mathbb{E}_\mu[F(\bar{x},\omega)].$$

Note that $\arg\max_{\mu\in\mathcal{A}}\mathbb{E}_\mu[F(\bar{x},\omega)] = \partial\rho(\bar{Z})$, where $\bar{Z}(\cdot):=F(\bar{x},\cdot)$.

Theorem. Suppose that $F(\cdot,\omega)$ is convex (for all $\omega\in\Omega$), $\bar{x}$ is an optimal solution of the primal problem and ρ is subdifferentiable at $\bar{Z}$. Then there exists a saddle point $(\bar{x},\bar{\mu})$. Consequently, the optimal value of the primal problem is equal to the optimal value, and $\bar{x}$ is an optimal solution, of the problem

$$\operatorname*{Min}_{x\in X}\mathbb{E}_{\bar{\mu}}[F(x,\omega)].$$

Two-stage stochastic programs

Suppose that the function $F(x,\omega)$ is given by the optimal value of the second stage problem:

$$\operatorname*{Min}_{y\in\mathcal{G}(x,\omega)} g(x,y,\omega),$$

where $g:\mathbb{R}^n\times\mathbb{R}^m\times\Omega\to\mathbb{R}$ is a random lower semicontinuous function and $\mathcal{G}:\mathbb{R}^n\times\Omega\rightrightarrows\mathbb{R}^m$ is a closed valued measurable multifunction. For example (linear two stage program), $\mathcal{G}(x,\omega):=\{y: T(\omega)x+W(\omega)y=h(\omega),\ y\geq 0\}$ and $g(x,y,\omega):=c^Tx+q(\omega)^Ty$.

Then, for a fixed $x\in X$,

$$\rho(F(x,\omega))=\inf_{y(\cdot)\in\mathcal{G}(x,\cdot)}\rho[g(x,y(\omega),\omega)],$$

and hence the the first stage problem is equivalent to the problem

$$\begin{array}{ll}\operatorname{Min}_{x\in X,\, y(\cdot)\in\mathcal{M}} & \rho[g(x,y(\omega),\omega)]\\ \text{subject to} & y(\omega)\in\mathcal{G}(x,\omega) \text{ a.e. } \omega\in\Omega.\end{array}$$

How this can be extended to a dynamic process (multistage programming)?

Conditional Risk Mappings

(Ruszczyński and Shapiro (2004))

Let $\mathcal{F}_1 \subset \mathcal{F}_2$ be sigma algebras of subsets of a set Ω, and $\mathcal{Z}_1 \subset \mathcal{Z}_2$ be linear spaces of real valued functions $\phi(\omega)$, $\omega \in \Omega$, measurable with respect to $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. We say that a mapping $\rho : \mathcal{Z}_2 \to \mathcal{Z}_1$ is a *conditional risk mapping* if the following properties hold:
(A1) Convexity: if $t \in [0,1]$ and $Z_1, Z_2 \in \mathcal{Z}_2$, then

$$t\rho(Z_1) + (1-t)\rho(Z_2) \succeq \rho[tZ_1 + (1-t)Z_2].$$

(A2) Monotonicity: if $Z_2 \succeq Z_1$, then $\rho(Z_2) \succeq \rho(Z_1)$.
(A3) Translation Equivariance: if $Z_1 \in \mathcal{Z}_1$ and $Z_2 \in \mathcal{Z}_2$, then

$$\rho(Z_2 + Z_1) = \rho(Z_2) + Z_1.$$

The inequalities in (A1) and (A2) are understood componentwise, i.e., $Z_2 \succeq Z_1$ means that $Z_2(\omega) \ge Z_1(\omega)$ for every $\omega \in \Omega$.

Example: $\rho(Z) := \mathbb{E}[Z|\mathcal{F}_1]$, $Z \in \mathcal{Z}_2$.

Dual Representation

Let ρ be a lower semicontinuous, positively homogeneous conditional risk mapping. Then

$$[\rho(Z)](\omega) = \sup_{\mu \in \mathcal{A}(\omega)} \mathbb{E}_\mu[Z], \;\; Z \in \mathcal{Z}_2,$$

where $\mathcal{A} : \Omega \rightrightarrows \mathcal{P}_{\mathcal{Y}_2|\mathcal{F}_1}$ and $\mathcal{P}_{\mathcal{Y}_2|\mathcal{F}_1}$ is the set of probability measures $\nu \in \mathcal{Y}_2$ such that for every $B \in \mathcal{F}_1$ it holds that $\nu(B) = 1$ if $\omega \in B$, and $\nu(B) = 0$ if $\omega \notin B$.

Example Let $\mathcal{Z}_i := L_1(\Omega, \mathcal{F}_i, P)$ and $\mathcal{Y}_i := L_\infty(\Omega, \mathcal{F}_i, P)$, $i = 1, 2$. For $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$,

$$\rho(Z) := \mathbb{E}[Z|\mathcal{F}_1] + \Phi(Z|\mathcal{F}_1), \;\; Z \in \mathcal{Z}_2,$$

where $\Big[\Phi(Z|\mathcal{F}_1)\Big](\omega)$ is equal to

$$\inf_{A \in \mathcal{Z}_1} \mathbb{E}\Big\{\varepsilon_1[A - Z]_+ + \varepsilon_2[Z - A]_+ \Big| \mathcal{F}_1\Big\}(\omega).$$

For $\varepsilon_1 \leq 1$, ρ is a continuous, positively homogeneous conditional risk mapping with

$$\mathcal{A}(\omega) = \left\{ \begin{array}{ll} (1+h)f_\omega : & -\varepsilon_1 \leq h(\omega) \leq \varepsilon_2, \; a.e.\, \omega, \\ & \mathbb{E}[hf_\omega|\mathcal{F}_1] = 0 \end{array} \right\},$$

where $f_\omega(\cdot)$ is the conditional density of P with respect to $\mathcal{F}_1$. For $\varepsilon_1 = 1$, $\rho(\cdot)$ can be viewed as an extension of the Conditional Value at Risk function.

Conditional Expectation Representation

Let $\mathcal{I}$ be a family of probability measures on $(\Omega, \mathcal{F}_2)$. Then

$$[\rho(Z)](\omega) := \sup_{\nu \in \mathcal{I}} \mathbb{E}_\nu[Z|\mathcal{F}_1](\omega)$$

is a positively homogeneous conditional risk mapping, provided it is well defined. Conversely, under certain regularity conditions, a positively homogeneous conditional risk mapping can be represented in the above form.

Dynamic Programming Equations

Suppose that there are: a sequence of sigma algebras $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_T$, with $\mathcal{F}_1 = \{\emptyset, \Omega\}$, $\mathcal{Z}_1 \subset \cdots \subset \mathcal{Z}_T$ a corresponding sequence of linear spaces of $\mathcal{F}_t$-measurable functions, $\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}} : \mathcal{Z}_t \to \mathcal{Z}_{t-1}$ a sequence of conditional risk mappings, a sequence $\mathcal{G}_t : \mathbb{R}^{n_{t-1}} \times \Omega \rightrightarrows \mathbb{R}^{n_t}$ of $\mathcal{F}_t$-measurable multifunctions, $t = 1, ..., T$. Note that since the sigma algebra $\mathcal{F}_1$ is trivial, $\mathcal{Z}_1 = \mathbb{R}$, $\mathcal{G}_1 \subset \mathbb{R}^{n_1}$ is constant, and the composite mapping

$$\rho_T := \rho_{\mathcal{Z}_2|\mathcal{Z}_1} \circ \cdots \circ \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}} : \mathcal{Z}_T \to \mathcal{Z}_1$$

is a risk function.

Consider the following multi-stage problem:

$$\begin{array}{cl} \operatorname*{Min}_{X \in \mathcal{M}} & \rho_T\Big[F_1(X_1) + F_2(X_2) + \cdots + F_T(X_T)\Big] \\ \text{s.t.} & X_t(\omega) \in \mathcal{G}_t(X_{t-1}(\omega), \omega),\ t = 1, ..., T, \end{array}$$

where $\Big[F_t(X_t)\Big](\omega) := F_t(X_t(\omega), \omega)$, $\mathcal{M} := \mathcal{M}_1 \times \cdots \times \mathcal{M}_T$, and $\mathcal{M}_t$, $t = 1, ..., T$, are linear spaces of $\mathcal{F}_t$-measurable functions $X_t : \Omega \to \mathbb{R}^{n_t}$. Note that the decision process X_t is adapted to the filtration $\mathcal{F}_t$, i.e., $X_t(\omega)$ is $\mathcal{F}_t$-measurable, $t = 1, ..., T$.

Dynamic programming equations:

$$[V_T(x_{T-1})](\omega) = \inf_{x_T \in \mathcal{G}_T(x_{T-1}, \omega)} F_T(x_T, \omega),$$

and for $t = T - 1, ..., 2$,

$$[V_t(x_{t-1})](\omega) = \inf_{x_t \in \mathcal{G}_t(x_{t-1}, \omega)} \Big\{F_t(x_t, \omega) + [\rho_{\mathcal{Z}_{t+1}|\mathcal{Z}_t}(V_{t+1}(x_t))](\omega)\Big\}.$$

The first-stage problem:

$$\operatorname*{Min}_{x_1 \in \mathcal{G}_1} \Big\{Q_2(x_1) := \rho_{\mathcal{Z}_2|\mathcal{Z}_1}(V_2(x_1))\Big\}.$$

Example of financial planning

We want to invest an amount of W_0 in n assets, x_i, $i = 1, ..., n$, in each. That is,

$$W_0 = \sum_{i=1}^n x_i. \tag{6}$$

Suppose that we can rebalance our portfolio at several, say T, periods of time. That is, at the beginning we choose values x_{i0} of our assets subject to the budget constraint . At the period $t = 1, ..., T$, our wealth is

$$W_t = \sum_{i=1}^n \xi_{it} x_{i,t-1},$$

where $\xi_{it} = (1 + R_{it})$ and R_{it} is the return of the i-th asset at the period t. Our objective is to maximize $\rho_T(W_1 + ... + W_T)$ subject to the balance constraints

$$\begin{array}{l} \sum_{i=1}^n x_{it} = W_t, \;\; x_t \geq 0, \\ W_{t+1} = \sum_{i=1}^n \xi_{i,t+1} x_{it}, \end{array} \quad t = 0, ..., T-1.$$

Suppose that the random process $\boldsymbol{\xi}_t$, $t = 1, ..., T$, is between stages independent, i.e., $\boldsymbol{\xi}_t$ is independent of $(\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_{t-1})$. Then the problem can be solved in a myopic way by solving at each stage the problem

$$\begin{array}{ll} \underset{W, x_{t-1}}{\text{Max}} & \rho_t(W) \\ \text{s.t.} & W = \sum_{i=1}^n \xi_{it} x_{i,t-1}, \\ & \sum_{i=1}^n x_{i,t-1} = 1, \\ & x_{i,t-1} \geq 0, \; i = 1, ..., n. \end{array}$$