Statistical inference of empirical estimates of stochastic programs

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

SAMSI Workshop on the Interface between Statistics and Optimization

February 2017

Consider stochastic optimization problem:

$$\operatorname*{Min}_{x \in \mathcal{X}} \Big\{ f(x) := \mathbb{E}_P[F(x,\xi)] \Big\}, \tag{1}$$

where ξ is a random vector having probability distribution P, and $F(x,\xi)$ is a real valued function. Assume that the expectation $f(x)$ is well defined and finite valued for every $x \in \mathcal{X}$.

Let $\widehat{P}_N$ be an empirical estimate of P, based on a sample of size N, and suppose that problem (1) is approximated by

$$\operatorname*{Min}_{x \in \mathcal{X}} \Big\{ \widehat{f}_N(x) := \mathbb{E}_{\widehat{P}_N}[F(x,\xi)] \Big\}. \tag{2}$$

Let ϑ_N and $\widehat{x}_N$ be the corresponding optimal value and an optimal solution of the approximating problem (2). What are statistical properties of ϑ_N and $\widehat{x}_N$?

Examples

Maximum Likelihood method.

Let $X_1, ..., X_N$ be an iid sample of realization of random vector $X \sim P$. Consider a parametric family with pdf $f(x, \theta)$, $\theta \in \Theta$, and the likelihood function $L_N(\theta) = \prod_{i=1}^N f(X_i, \theta)$. We have that

$$N^{-1} \log L_N(\theta) = N^{-1} \sum_{i=1}^N \log f(X_i, \theta) = \mathbb{E}_{\hat{P}_N}[\log f(X, \theta)],$$

where $\hat{P}_N = N^{-1} \sum_{i=1}^N \delta(X_i)$. By the Law of Large Numbers we have that $N^{-1} \log L_N(\theta)$ converges w.p.1 to $\mathbb{E}_P[\log f(X, \theta)]$. The ML problem

$$\operatorname*{Max}_{\theta \in \Theta} N^{-1} \log L_N(\theta)$$

can be considered as an empirical approximation of the problem

$$\operatorname*{Max}_{\theta \in \Theta} \mathbb{E}_P[\log f(X, \theta)].$$

Semidefinite Programming

Let Σ be an $m \times m$ (population) covariance matrix. Consider the problem

$$\operatorname*{Min}_{x \in \mathcal{X}} c^\top x \text{ s.t. } \Sigma + \sum_{i=1}^{n} x_i A_i \succeq 0, \tag{3}$$

where A_i, $i = 1, ..., n$, are $m \times m$ symmetric matrices, $\mathcal{X} \subset \mathbb{R}^n$. The true problem (3) is approximated by replacing (unknown) Σ with the sample covariance matrix

$$S = (N-1)^{-1} \sum_{j=1}^{N} (Y_j - \bar{Y})(Y_j - \bar{Y})^\top,$$

based on a sample $Y_1, ..., Y_N$ of size N. Classical example is the Minimum Trace Factor Analysis

$$\operatorname*{Min}_{\Psi \in \mathbb{D}^m} \text{Trace}(\Sigma - \Psi) \text{ s.t. } \Sigma - \Psi \succeq 0, \tag{4}$$

where $\mathbb{D}^m$ is the set of $m \times m$ diagonal positive semidefinite matrices.

Sample Average Approximation (SAA) method.

Consider 'true' stochastic programming problem (1). Let $\xi^1, ..., \xi^N$ be an iid sample of random vector ξ, say generated by Monte Carlo method. The corresponding problem

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

is considered as an approximation of the true problem.

Two-stage (linear) stochastic programming problem with recourse, $F(x, \xi) := c^{\mathsf{T}} x + Q(x, \xi)$, where $\mathcal{X} := \{x : Ax = b,\ x \geq 0\}$ and $Q(x, \xi)$ is the optimal value of the second stage problem

$$\operatorname{Min}_y q^{\mathsf{T}} y \ \text{ s.t. } \ Tx + Wy = h,\ y \geq 0,$$

with $\xi = (q, T, W, h)$.

Nested formulation the linear multistage stochastic program

$$\min_{\substack{A_1 x_1 = b_1 \\ x_1 \geq 0}} c_1^{\mathsf T} x_1 + \mathbb{E}_{|\xi_1}\left[\min_{\substack{B_2 x_1 + A_2 x_2 = b_2 \\ x_2 \geq 0}} c_2^{\mathsf T} x_2 + \cdots + \mathbb{E}_{|\xi_{[T-1]}}\left[\min_{\substack{B_T x_{T-1} + A_T x_T = b_T \\ x_T \geq 0}} c_T^{\mathsf T} x_T \right]\right].$$

Equivalent formulation

$$\begin{array}{cl} \min\limits_{x_1, x_2(\cdot), \dots, x_T(\cdot)} & \mathbb{E}\left[c_1^{\mathsf T} x_1 + c_2^{\mathsf T} x_2(\xi_{[2]}) \dots + c_T^{\mathsf T} x_T(\xi_{[T]}) \right] \\ \text{s.t.} & A_1 x_1 = b_1,\ x_1 \geq 0, \\ & B_t x_{t-1}(\xi_{[t-1]}) + A_t x_t(\xi_{[t]}) = b_t, \\ & x_t(\xi_{[t]}) \geq 0,\ t = 2, \dots, T. \end{array}$$

Here $\xi_t = (c_t, B_t, A_t, b_t)$, $t = 2, \dots, T$, is considered as a random process, $\xi_1 = (c_1, A_1, b_1)$ is supposed to be known, $\xi_{[t]} := (\xi_1, \dots, \xi_t)$ denotes history of the data process up to time t.

Optimization is performed over feasible policies (also called decision rules). A policy is a sequence of (measurable) functions $x_t = x_t(\xi_{[t]})$, $t = 1, ..., T$. Each $x_t(\xi_{[t]})$ is a function of the data process up to time t, this ensures the *nonanticipative* property of a considered policy. The constraints should be satisfied for almost every realization of the random data process.

Suppose that the random process ξ_t is stagewise independent, i.e., ξ_{t+1} is independent of $\xi_{[t]}$, $t = 1, ..., T-1$. An SAA problem is constructed by generating independent samples from marginal distributions of ξ_t of respective sample sizes N_t, $t = 2, ..., T$. Note that the total number of scenarios (sample paths) of the obtained scenario tree is $\prod_{t=2}^{T} N_t$.

Risk measures

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{Z} = L_p(\Omega, \mathcal{F}, P)$, $p \in [1, \infty)$, be the linear space of random variables (measurable functions) $Z : \Omega \to \mathbb{R}$ having finite p-th order moments. Functional $\rho : \mathcal{Z} \to \mathbb{R}$ is said to be a coherent risk measure if it satisfies the following axioms (Artzner, Delbaen, Eber, Heath (1999)):

(A1) Convexity: $\rho(\alpha Z_1 + (1-\alpha)Z_2) \leq \alpha\rho(Z_1) + (1-\alpha)\rho(Z_2)$ 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 $Z_2 \geq Z_1$, then $\rho(Z_2) \geq \rho(Z_1)$.

(A3) Translation Equivariance: If $a \in \mathbb{R}$ and $Z \in \mathcal{Z}$, then $\rho(Z + a) = \rho(Z) + a$.

(A4) Positive Homogeneity: $\rho(\alpha Z) = \alpha\rho(Z)$, $Z \in \mathcal{Z}$, $\alpha \geq 0$.

It is said that risk measure $\rho : \mathcal{Z} \to \mathbb{R}$ is *law invariant* if $Z_1, Z_2 \in \mathcal{Z}$ and $Z_1 \overset{D}{\sim} Z_2$ implies that $\rho(Z_1) = \rho(Z_2)$. The notation $Z_1 \overset{D}{\sim} Z_2$ means that Z_1 and Z_2 are distributionally equivalent, i.e., $P(Z_1 \le z) = P(Z_2 \le z)$ for all $z \in \mathbb{R}$. If ρ is law invariant, then it can be considered as a function of the cdf $G(z) = P(Z \le z)$. Let $\widehat{G}_N(z) = N^{-1} \sum_{i=1}^N \mathbb{1}(Z_i \le z)$ be the empirical cdf based on sample $Z_1, ..., Z_N$. Then $\rho(G)$ can be estimated by $\rho(\widehat{G}_N)$.

Average Value-at-Risk measure, for $\alpha \in (0, 1)$,

$$\mathsf{AVaR}_\alpha(G) = \frac{1}{1-\alpha} \int_\alpha^1 G^{-1}(t)dt = \inf_{t \in \mathbb{R}} \left\{ t + (1-\alpha)^{-1} \mathbb{E}_G[Z - t]_+ \right\}.$$

Its empirical estimate

$$\mathsf{AVaR}_\alpha(\widehat{G}_N) = \inf_{t \in \mathbb{R}} \left\{ t + \frac{1}{N(1-\alpha)} \sum_{i=1}^N [Z_i - t]_+ \right\}.$$

Any law invariant coherent risk measure $\rho : \mathcal{Z} \to \mathbb{R}$ can be written in the following form (the so called Kusuoka representation)

$$\rho(G) = \sup_{\mu \in \mathfrak{M}} \int_0^1 \mathsf{AVaR}_\alpha(G) d\mu(\alpha),$$

where $\mathfrak{M}$ is a set of probability measures on the interval $[0,1)$. Consequently ρ can be represented in the following minimax form

$$\begin{aligned} \rho(G) &= \sup_{\mu \in \mathfrak{M}} \int_0^1 \inf_{t \in \mathbb{R}} \left\{ \int_{-\infty}^{+\infty} h_\alpha(z,t) dG(z) \right\} d\mu(\alpha) \\ &= \sup_{\mu \in \mathfrak{M}} \inf_{\tau(\cdot) \in L_p} \int_0^1 \int_{-\infty}^{+\infty} h_\alpha(z, \tau(\alpha)) dG(z) d\mu(\alpha), \end{aligned}$$

where

$$h_\alpha(z,t) := t + (1-\alpha)^{-1}[z-t]_+.$$

Consistency

Under what conditions the empirical estimates ϑ_N and $\widehat{x}_N$ converge as $N \to \infty$ (say w.p.1) to their true counterparts? That is, whether $\vartheta_N \to \vartheta_0$ and $\text{dist}(\widehat{x}_N, \mathcal{S}_0) \to 0$ w.p.1, where ϑ_0 is the optimal value and $\mathcal{S}_0$ is the set of optimal solutions of the true problem.

Uniform Laws of Large Numbers

Theorem 1 (convex case) *Suppose that the set $\mathcal{X} \subset \mathbb{R}^n$ is compact, the function $F(x, \xi)$ is convex in $x \in \mathbb{R}^n$, the expected value function $f(x)$ is finite valued on a neighborhood $\mathcal{V}$ of $\mathcal{X}$ and $\widehat{f}_N(x)$ converges w.p.1 to $f(x)$ for every $x \in \mathcal{V}$. Then*

$$\sup_{x \in \mathcal{X}} \left| \widehat{f}_N(x) - f(x) \right| \to 0 \text{ w.p.1 as } N \to \infty.$$

Theorem 2 *Suppose that: (i) $\mathcal{X}$ is a compact metric space, (ii) for any $x \in \mathcal{X}$ the function $F(\cdot, \xi)$ is continuous at x for a.e. ξ, (iii)* $\sup_{x \in \mathcal{X}} |F(x, \xi)| \leq g(\xi)$ *with $\mathbb{E}|g(\xi)| < \infty$, (iv) the sample is iid. Then the expected value function $f(x)$ is finite valued and continuous on $\mathcal{X}$, and* $\sup_{x \in \mathcal{X}} \left| \widehat{f}_N(x) - f(x) \right| \to 0$ w.p.1.

Epiconvergence

Consider a sequence $f_k : \mathbb{R}^n \to \overline{\mathbb{R}}$ of extended real valued functions. It is said that f_k *epiconverge* to a function f, written $f_k \overset{e}{\to} f$, if for any $x \in \mathbb{R}^n$ the following two conditions hold:
(i) for any sequence x_k converging to x one has

$$\liminf_{k \to \infty} f_k(x_k) \geq f(x),$$

(ii) there exists a sequence x_k converging to x such that

$$\limsup_{k \to \infty} f_k(x_k) \leq f(x).$$

Let $\rho : \mathcal{Z} \to \mathbb{R}$, $\mathcal{Z} = L_p(\Omega, \mathcal{F}, P)$, $p \in [1, \infty)$, be a law invariant, convex risk measure (i.e., satisfies axioms of Convexity, Monotonicity and Translation Equivariance), and $F_x(\omega) = F(x, \xi(\omega))$. Suppose that for every $x \in \mathbb{R}^n$ the random variable $F_x \in \mathcal{Z}$. Then we can consider the composite function $\phi(x) := \rho(F_x)$. Let $\widehat{G}_{xN}$ be empirical cdf associated with $Y_x^j = F(x, \xi^j)$, $j = 1, ..., N$, and iid sample $\xi^1, ..., \xi^N$. Then $\widehat{\phi}_N(x) := \rho(\widehat{G}_{xN})$ gives an estimate of $\phi(x)$.

Theorem 3 *Suppose that:(i) $F_x \in \mathcal{Z}$ for every $x \in \mathbb{R}^n$, (ii) the function $F(x, \omega) = F_x(\omega)$ is random lower semicontinuous, (iii) for every $\bar{x} \in \mathbb{R}^n$ there is a neighborhood $\mathcal{V}_{\bar{x}}$ of $\bar{x}$ and a function $h \in \mathcal{Z}$ such that $F(x, \cdot) \geq h(\cdot)$ for all $x \in \mathcal{V}_{\bar{x}}$. Then the functions ϕ is lower semicontinuous and $\widehat{\phi}_N \stackrel{\text{e}}{\to} \phi$ w.p.1.*

Central Limit Theorem type results. Notoriously slow convergence of order $O_p(N^{-1/2})$. By the CLT, for a given $x \in \mathcal{X}$,

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

where $\sigma^2(x) := \mathsf{Var}[F(x,\xi)]$.

Delta method

Let $Y_N \in \mathbb{R}^d$ be a sequence of random vectors, converging in probability to a vector $\mu \in \mathbb{R}^d$. Suppose that there exists a sequence $\tau_N \to +\infty$ such that $\tau_N(Y_N - \mu) \xrightarrow{\mathcal{D}} Y$. Let $G : \mathbb{R}^d \to \mathbb{R}^m$ be a vector valued function, differentiable at μ, and $M := \nabla G(\mu)$ be the $m \times d$ Jacobian matrix. Then $\tau_N \left[G(Y_N) - G(\mu)\right] \xrightarrow{\mathcal{D}} MY$.

In particular, suppose that $N^{1/2}(Y_N - \mu)$ converges in distribution to a (multivariate) normal distribution $\mathcal{N}(0, \Sigma)$ with zero mean vector and covariance matrix Σ. Then it follows that $N^{1/2}\left[G(Y_N) - G(\mu)\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, M\Sigma M^{\mathsf{T}})$.

Infinite dimensional Delta Theorem

Let B_1 and B_2 be two Banach spaces, K be a closed subset of B_1 and $G : K \to B_2$. It is said that G is Hadamard directionally differentiable at $\mu \in K$ tangentially to K if for any $d \in T_K(\mu)$ the following limit exists

$$G'_\mu(d) = \lim_{\substack{t \downarrow 0 \\ d' \to_K d}} \frac{G(\mu + td') - G(\mu)}{t}.$$

The notation $d' \to_K d$ means that $K \ni d' \to d$. The contingent cone $T_K(\mu)$ consists of the limits of sequences $(y_n - \mu)/t_n$, where $y_n \in K$ and $t_n \downarrow 0$. In particular if the set K is convex, then $T_K(\mu)$ is the tangent cone to K at μ.

Theorem 4 (Delta Theorem) *Let B_1 and B_2 be Banach spaces, equipped with their Borel σ-algebras, K be a closed subset of B_1, $G : K \to B_2$, τ_N be a sequence of positive numbers tending to infinity as $N \to \infty$, and Y_N be a sequence of random elements of B_1 such that $Y_N \in K$ w.p.1. Suppose that the set K is separable, the mapping G is Hadamard directionally differentiable at a point $\mu \in K$ tangentially to K, and the sequence $X_N := \tau_N(Y_N - \mu)$ converges in distribution to a random element Y of B_1. Then*

$$\tau_N \left[G(Y_N) - G(\mu) \right] \stackrel{\mathcal{D}}{\to} G'_\mu(Y).$$

Moreover if the set K is convex, then

$$\tau_N \left[G(Y_N) - G(\mu) \right] = G'_\mu(X_N) + o_p(1),$$

or equivalently

$$G(Y_N) = G(\mu) + G'_\mu(Y_N - \mu) + o_p(\tau_N^{-1}).$$

Second order Delta Theorem

Suppose further that the set K is convex and second order Hadamard directional derivative tangentially to K exists for all $d \in T_K(\mu)$:

$$G''_\mu(d) = \lim_{\substack{t \downarrow 0 \\ d' \to_K d}} \frac{G(\mu + td') - G(\mu) - tG'_\mu(d')}{\frac{1}{2}t^2}.$$

Then

$$\tau_N^2 \left[G(Y_N) - G(\mu) - G'_\mu(Y_N - \mu)\right] \xrightarrow{\mathcal{D}} \tfrac{1}{2}G''_\mu(Y).$$

and

$$G(Y_N) = G(\mu) + G'_\mu(Y_N - \mu) + \tfrac{1}{2}G''_\mu(Y_N - \mu) + o_p(\tau_N^{-2}).$$

Asymptotics of the SAA problems. Let $\mathcal{X}$ be a nonempty compact subset of $\mathbb{R}^n$ and consider the space $B = C(\mathcal{X})$ of continuous functions $\psi : \mathcal{X} \to \mathbb{R}$. Assume that:

(A1) For some point $x \in \mathcal{X}$ the expectation $\mathbb{E}[F(x,\xi)^2]$ is finite.

(A2) There exists a measurable function $C(\xi)$ such that $\mathbb{E}[C(\xi)^2]$ is finite and

$$\left| F(x,\xi) - F(x',\xi) \right| \le C(\xi)\|x - x'\|,$$

for all $x, x' \in \mathcal{X}$ and a.e. ξ.

We can view $Y_N := \hat{f}_N$ as a random element of $C(\mathcal{X})$. Consider the min-function $V : B \to \mathbb{R}$ defined as $V(Y) := \inf_{x\in\mathcal{X}} Y(x)$. Clearly $\hat{\vartheta}_N = V(Y_N)$. It is possible to show that for any $\mu \in C(\mathcal{X})$ and $\mathcal{X}^*(\mu) := \arg\min_{x\in\mathcal{X}} \mu(x)$,

$$V'_\mu(\delta) = \inf_{x\in\mathcal{X}^*(\mu)} \delta(x), \;\; \forall \delta \in C(\mathcal{X}),$$

and the above directional derivative holds in the Hadamard sense.

By a functional CLT, under assumptions (A1) and (A2), $N^{1/2}(\hat{f}_N - f)$ converges in distribution to a random element Y of $C(\mathcal{X})$. For any finite set $\{x_1, ..., x_m\} \subset \mathcal{X}$, the random vector $(Y(x_1), ..., Y(x_m))$ has a multivariate normal distribution with zero mean and the same covariance matrix as the covariance matrix of $(F(x_1, \xi), ..., F(x_m, \xi))$. In particular, for fixed $x \in \mathcal{X}$, $Y(x) \sim N(0, \sigma^2(x))$ with $\sigma^2(x) := \mathsf{Var}[F(x, \xi)]$.

Theorem 5 *Suppose that the set $\mathcal{X} \subset \mathbb{R}^n$ is compact, and assumptions (A1) and (A2) hold. Then*

$$\hat{\vartheta}_N = \min_{x \in \mathcal{S}_0} \hat{f}_N(x) + o_p(N^{-1/2}),$$
$$N^{1/2}[\hat{\vartheta}_N - \vartheta_0] \xrightarrow{\mathcal{D}} \inf_{x \in \mathcal{S}_0} Y(x).$$

In particular, if the optimal set (of the true problem) $\mathcal{S}_0 = \{x_0\}$ is a singleton, then

$$N^{1/2}[\hat{\vartheta}_N - \vartheta_0] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(x_0)).$$

Second order asymptotics. Consider second order directional derivative of the optimal value function $V(Y) = \inf_{x \in \mathcal{X}} Y(x)$:

$$V''_\mu(d) := \lim_{\substack{t \downarrow 0 \\ d' \to d}} \frac{V(\mu + td') - V(\mu) - tV'_\mu(d')}{\frac{1}{2}t^2}.$$

If $\mathcal{S}_0 = \{x_0\}$, then under certain regularity conditions

$$V''_\mu(\delta) = \inf_{h \in C(x_0)} \Big\{ 2h^T \nabla \delta(x_0) + h^{\mathsf{T}} \nabla^2 f(x_0) h - \mathsf{s}(-\nabla f(x_0), T^2_{\mathcal{X}}(x_0, h)) \Big\},$$

where $\mathsf{s}(y, A) = \sup_{z \in A} y^{\mathsf{T}} z$ is the support function of set A,

$$C(x_0) = \Big\{ h \in T_{\mathcal{X}}(x_0) : h^T \nabla f(x_0) = 0 \Big\}$$

is the critical cone, $T_{\mathcal{X}}(x_0)$ is the tangent cone to $\mathcal{X}$ at x_0, and

$$T^2_{\mathcal{X}}(x, h) = \Big\{ z : \mathrm{dist}(x + th + \tfrac{1}{2}t^2 z, \mathcal{X}) = o(t^2),\ t \geq 0 \Big\}$$

is the second order tangent set.

Suppose that $N^{1/2}(\hat{f}_N - f)$ converges in distribution to a random element Y (in a functional space of Lipschitz continuous functions) and regularity conditions hold (in particular, the true problem has unique optimal solution x_0). Then

$$\hat{\vartheta}_N = \hat{f}_N(x_0) + \tfrac{1}{2}V''_f(\hat{f}_N - f) + o_p(N^{-1})$$

$$N\Big[\hat{\vartheta}_N - \hat{f}_N(x_0)\Big] \xrightarrow{\mathcal{D}} \tfrac{1}{2}V''_f(Y).$$

Moreover, if for all δ the optimization problem in the calculation of $V''_f(\delta)$ has unique optimal solution $\bar{h} = \bar{h}(\delta)$, then

$$N^{1/2}(\hat{x}_N - x_0) \xrightarrow{\mathcal{D}} \bar{h}(Y).$$

Bias of the optimal value $\widehat{\vartheta}_N$ of the SAA problem.

For any *fixed* $x \in \mathcal{X}$, $\mathbb{E}\left[\widehat{f}_N(x)\right] = f(x)$, i.e., $\widehat{f}_N(x)$ is an unbiased estimator of $f(x) = \mathbb{E}[F(x,\xi)]$. However,

$$\mathbb{E}\left[\widehat{\vartheta}_N\right] = \mathbb{E}\left[\min_{x\in\mathcal{X}} \widehat{f}_N(x)\right] \le \mathbb{E}\left[\widehat{f}_N(x)\right], \;\; \forall x \in \mathcal{X},$$

and hence $\mathbb{E}\left[\widehat{\vartheta}_N\right] \le \vartheta_0 = \min_{x\in\mathcal{X}} f(x)$.

If $\mathcal{S}_0 = \{x_0\}$, then (under some regularity conditions)

$$\mathbb{E}\left[\widehat{\vartheta}_N\right] - \vartheta_0 = \tfrac{1}{2}N^{-1}\mathbb{E}[V_f''(Y)] + o(N^{-1}),$$

and hence the (negative) bias $\mathbb{E}\left[\widehat{\vartheta}_N\right] - \vartheta_0$ is of order $O(N^{-1})$. If $\mathcal{S}_0$ is "large", then

$$\mathbb{E}\left[\widehat{\vartheta}_N\right] - \vartheta_0 = N^{-1/2}\mathbb{E}\left[\inf_{x\in\mathcal{S}_0} Y(x)\right] + o(N^{-1/2}),$$

and the bias is of order $O(N^{-1/2})$.

Example (Average Value-at-Risk)
Sample estimate of $\theta = \mathsf{AVaR}_\alpha(Z)$ is obtained by replacing the expectation $\mathbb{E}[Z-t]_+$ with the corresponding sample average, that is

$$\hat{\theta}_N = \inf_{t\in\mathbb{R}} \left\{ t + \frac{1}{N(1-\alpha)} \sum_{i=1}^{N} [Z_i - t]_+ \right\}.$$

By the above theorem (assuming $\mathbb{E}|Z|^2 < \infty$) we have

$$\hat{\theta}_N = \inf_{t\in[\underline{t},\bar{t}]} \left\{ t + \frac{1}{N(1-\alpha)} \sum_{i=1}^{N} [Z_i - t]_+ \right\} + o_p(N^{-1/2}),$$

where $\underline{t}$ is the left side and $\bar{t}$ is the right side α-quantiles of the distribution of Z. In particular, if $\underline{t} = \bar{t}$, then $N^{1/2}(\hat{\theta}_N - \theta)$ converges in distribution to normal $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = (1-\alpha)^{-2}\mathsf{Var}([Z-\bar{t}\,]_+)$.

Second order expansion of the Average Value-at-Risk
Suppose that cdf $G(\cdot)$ of Z has unique α-quantile $\bar{t} = G^{-1}(\alpha)$. Suppose further that $G(\cdot)$ is continuous at $\bar{t}$ and has positive density $g(t) = dG(t)/dt$ at $t = \bar{t}$. Then

$$\begin{aligned}\widehat{\theta}_N - \widehat{f}_N(\bar{t}) &= N^{-1} \inf_{\tau \in \mathbb{R}} \left\{ \tau V + \tfrac{1}{2}\tau^2 f''(\bar{t}) \right\} + o_p(N^{-1}) \\ &= -\frac{(1-\alpha)V^2}{2Ng(\bar{t})} + o_p(N^{-1}),\end{aligned}$$

where $V \sim \mathcal{N}(0, \gamma^2)$ with

$$\gamma^2 := \mathrm{Var}\left((1-\alpha)^{-1} \frac{\partial [Z - \bar{t}]_+}{\partial t} \right) = \frac{G(\bar{t})(1 - G(\bar{t}))}{(1-\alpha)^2} = \frac{\alpha}{1-\alpha}.$$

Consequently, under appropriate regularity conditions,

$$\begin{aligned} N\left[\widehat{\theta}_N - \widehat{f}_N(\bar{t})\right] &\xrightarrow{\mathcal{D}} -\left[\tfrac{\alpha}{2g(\bar{t})}\right] \chi_1^2, \\ \mathbb{E}[\widehat{\theta}_N] - \theta &= -\tfrac{\alpha}{2Ng(\bar{t})} + o(N^{-1}).\end{aligned}$$

Minimax stochastic programs

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

where $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ are closed sets, $F : \mathcal{X} \times \mathcal{Y} \times \Xi \to \mathbb{R}$ and $\xi = \xi(\omega)$ is a random vector whose probability distribution is supported on set $\Xi \subset \mathbb{R}^d$. The corresponding SAA problem is obtained by using the sample average as an approximation of the expectation $f(x,y)$, that is

$$\operatorname*{Min}_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \left\{ \hat{f}_N(x,y) := \frac{1}{N} \sum_{j=1}^{N} F(x,y,\xi^j) \right\}.$$

Suppose that: (i) the sets $\mathcal{X}$ and $\mathcal{Y}$ are nonempty, convex and compact, (ii) the function $F(x,y,\xi)$ is convex in $x \in \mathcal{X}$ and concave in $y \in \mathcal{Y}$, (iii) $F(x,y,\xi)$ is dominated by an integrable function, (iv) for some $(x,y) \in \mathcal{X} \times Y$, $\mathbb{E}[F(x,y,\xi)^2] < \infty$, (v) there is $C : \Xi \to \mathbb{R}_+$ such that for all $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$ and a.e. ξ,

$$|F(x,y,\xi) - F(x',y',\xi)| \leq C(\xi)(\|x - x'\| + \|y - y'\|).$$

Then

$$\hat{\vartheta}_N = \inf_{x \in \mathcal{S}_X} \sup_{y \in \mathcal{S}_Y} \hat{f}_N(x,y) + o_p(N^{1/2}),$$

where $\mathcal{S}_X$ and $\mathcal{S}_Y$ are the respective sets of optimal solutions. Moreover, if $\mathcal{S}_X = \{x_0\}$ and $\mathcal{S}_Y = \{y_0\}$ (i.e., (x_0, y_0) is the unique saddle point od the true problem), then $N^{1/2}(\hat{\vartheta}_N - \vartheta_0)$ converges in distribution to normal $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \text{Var}[F(x_0, y_0, \xi)]$.

Sample size estimates (by Large Deviations type bounds)

Consider an iid sequence $Y_1, \ldots, Y_N$ of replications of a real valued random variable Y, and let $Z_N := N^{-1} \sum_{i=1}^N Y_i$ be the corresponding sample average. Then for any real numbers a and $t > 0$ we have that $\text{Prob}(Z_N \geq a) = \text{Prob}(e^{tZ_N} \geq e^{ta})$, and hence, by Markov inequality

$$\text{Prob}(Z_N \geq a) \leq \quad e^{-ta}\mathbb{E}[e^{tZ_N}] = e^{-ta}[M(t/N)]^N,$$

where $M(t) := \mathbb{E}[e^{tY}]$ is the *moment generating function* of Y. Suppose that Y has finite mean $\mu := \mathbb{E}[Y]$ and let $a \geq \mu$. By taking the logarithm of both sides of the above inequality, changing variables $t' = t/N$ and minimizing over $t' > 0$, we obtain

$$\frac{1}{N} \log \left[\text{Prob}(Z_N \geq a) \right] \leq -I(a), \tag{5}$$

where $I(z) := \sup_{t \in \mathbb{R}} \{tz - \Lambda(t)\}$ is the conjugate of the logarithmic moment generating function $\Lambda(t) := \log M(t)$.

Suppose that $|\mathcal{X}| < \infty$, i.e., the set $\mathcal{X}$ is finite. Let $\mathcal{S}_\varepsilon$ be the set of ε-optimal solutions of the true problem and $\hat{\mathcal{S}}_{\delta N}$ be the set of δ-optimal solutions of the corresponding SAA problem. Suppose that: (i) for every $x \in \mathcal{X}$ the expected value $f(x) = \mathbb{E}[F(x, \xi)]$ is finite, (ii) there are constants $\sigma > 0$ and $a \in (0, +\infty]$ such that

$$M_x(t) \leq \exp\{\sigma^2 t^2/2\}, \ \ \forall t \in [-a, a], \ \forall x \in \mathcal{X} \setminus \mathcal{S}^\varepsilon,$$

where $M_x(t)$ is the moment generating function of the random variable $F(u(x), \xi) - F(x, \xi) - \mathbb{E}[F(u(x), \xi) - F(x, \xi)]$ and $u(x)$ is a point of the optimal set $\mathcal{S}_0$. Choose $\varepsilon > 0$, $\delta \geq 0$ and $\alpha \in (0, 1)$ such that $0 < \varepsilon - \delta \leq a\sigma^2$. Then for sample size

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

we are guaranteed, with probability at least $1 - \alpha$, that any δ-optimal solution of the SAA problem is an ε-optimal solution of the true problem, i.e., $\text{Prob}(\hat{\mathcal{S}}_{\delta N} \subset \mathcal{S}_\varepsilon) \geq 1 - \alpha$.

Let $\mathcal{X} = \{x_1, x_2\}$ with $f(x_2) - f(x_1) > \varepsilon > 0$ and suppose that random variable $F(x_2, \xi) - F(x_1, \xi)$ has normal distribution with mean $\mu = f(x_2) - f(x_1)$ and variance σ^2. By solving the corresponding SAA problem we make the correct decision (that x_1 is the minimizer) if $\hat{f}_N(x_2) - \hat{f}_N(x_1) > 0$. Probability of this event is $\Phi(\mu\sqrt{N}/\sigma)$. Therefore we need the sample size $N > z_\alpha^2 \sigma^2/\varepsilon^2$ in order for our decision to be correct with probability at least $1 - \alpha$.

In order to solve the corresponding optimization problem we need to test $H_0 : \mu \le 0$ versus $H_a : \mu > 0$. Assuming that σ^2 is known, by Neyman-Pearson Lemma, the uniformly most powerful test is: "reject H_0 if $\hat{f}_N(x_2) - \hat{f}_N(x_1)$ is bigger than a specified critical value".

Now let $\mathcal{X} \subset \mathbb{R}^n$ be a set of finite diameter $D := \sup_{x',x\in\mathcal{X}} \|x'-x\|$. Suppose that: (i) for every $x \in \mathcal{X}$ the expected value $f(x) = \mathbb{E}[F(x,\xi)]$ is finite, (ii) there is a constant $\sigma > 0$ such that

$$M_{x',x}(t) \leq \exp\{\sigma^2 t^2/2\}, \;\; \forall t \in \mathbb{R}, \; \forall x', x \in \mathcal{X},$$

where $M_{x',x}(t)$ is the moment generating function of the random variable $F(x',\xi) - F(x,\xi) - \mathbb{E}[F(x',\xi) - F(x,\xi)]$, (iii) there exists $\kappa : \Xi \to \mathbb{R}_+$ such that its moment generating function is finite valued in a neighborhood of zero (with $L = \mathbb{E}[\kappa(\xi)]$) and

$$\left| F(x',\xi) - F(x,\xi) \right| \leq \kappa(\xi) \|x' - x\|, \;\; \forall \xi \in \Xi, \; \forall x', x \in \mathcal{X}.$$

Choose $\varepsilon > 0$, $\delta \in [0, \varepsilon)$ and $\alpha \in (0, 1)$. Then for sample size

$$N \geq \frac{8\sigma^2}{(\varepsilon - \delta)^2} \left[n \log\left(\frac{O(1)DL}{(\varepsilon - \delta)^2} \right) + \log\left(\frac{2}{\alpha} \right) \right] \bigvee \left[\beta^{-1} \log\left(\frac{2}{\alpha} \right) \right],$$

we are guaranteed that $\text{Prob}\left(\hat{\mathcal{S}}_{\delta N} \subset \mathcal{S}_{\varepsilon} \right) \geq 1 - \alpha$.

In particular, if $\kappa(\xi) \equiv L$, then the estimate takes the form

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

Suppose further that for some $c > 0$, $\gamma \geq 1$ and $\bar{\varepsilon} > \varepsilon$ the following growth condition holds

$$f(x) \geq \vartheta^* + c[\mathrm{dist}(x, \mathcal{S}_0)]^\gamma, \ \ \forall x \in \mathcal{S}_{\bar{\varepsilon}},$$

and that the problem is convex. Then, for $\delta \in [0, \varepsilon/2]$, we have the following estimate of the required sample size:

$$N \geq \left(\frac{O(1)LD}{c^{1/\gamma}\varepsilon^{(\gamma-1)\gamma}}\right)^2 \left[n \log \left(\frac{O(1)\bar{D}L}{\varepsilon}\right) + \log \left(\frac{1}{\alpha}\right)\right],$$

where $\bar{D}$ is the diameter of $\mathcal{S}_{\bar{\varepsilon}}$. In particular, if $\mathcal{S}_0 = \{x_0\}$ is a singleton and $\gamma = 1$, we have the estimate (independent of ε):

$$N \geq O(1)c^{-2}L^2 \left[n \log(O(1)c^{-1}L) + \log(\alpha^{-1})\right].$$

Sample complexity of multistage stochastic programming

Conditional sampling. Let ξ_2^i, $i = 1, ..., N_1$, be an iid random sample of ξ_2. Conditional on $\xi_2 = \xi_2^i$, a random sample ξ_3^{ij}, $j = 1, ..., N_2$, is generated and etc. The obtained scenario tree is considered as a sample approximation of the true problem. Note that the total number of scenarios $N = \prod_{t=1}^{T-1} N_t$ and each scenario in the generated tree is considered with the same probability $1/N$. Note also that in the case of stagewise independence of the corresponding random process, we have two possible strategies. We can generate a different (independent) sample ξ_3^{ij}, $j = 1, ..., N_2$, for every generated node ξ_2^i, or we can use the same sample ξ_3^j, $j = 1, ..., N_2$, for every ξ_2^i. In the second case we preserve the stagewise independence condition for the generated scenario tree.

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

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

we have that any first-stage $\varepsilon/2$-optimal solution of the SAA problem is an ε-optimal first-stage solution of the true problem with probability at least $1-\alpha$. (Here $D_1, D_2, L_1, L_2, \sigma_1, \sigma_2$ are certain analogues of similar constants in the sample size estimate of two stage problem.)

In particular, suppose that $N_1 = N_2$ and take $L := \max\{L_1, L_2\}$, $D := \max\{D_1, D_2\}$, $\sigma^2 := \max\{\sigma_1^2, \sigma_2^2\}$ and $n := \max\{n_1, n_2\}$. Then the required sample size $N_1 = N_2$:

$$N_1 \geq \frac{O(1)\sigma^2}{\varepsilon^2}\left[n \log\left(\frac{O(1)DL}{\varepsilon}\right) + \log\left(\frac{1}{\alpha}\right)\right],$$

with total number of scenarios $N = N_1^2$.

That is, the total number of scenarios needed to solve a T-stage stochastic program with a reasonable accuracy by the SAA method grows exponentially with increase of the number of stages T. Another way of putting this is that the number of scenarios needed to solve T-stage problem would grow as $O\left(\varepsilon^{-2(T-1)}\right)$ with decrease of the error level $\varepsilon > 0$.