Monte Carlo Sampling Approach to Stochastic Programming

A. Shapiro

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

The "true" (or expected value) optimization problem

$$\operatorname*{Min}_{x \in X} \ \{g(x) := \mathbb{E}_P[G(x, \xi(\omega))]\},$$

where $\xi(\omega)$ is a random vector having probability distribution P, $G(x, \xi)$ is a real valued function and $X \subset \mathbb{R}^n$. The random vector $\xi(\omega)$ represents the uncertain parameters (data) of the problem. In two-stage stochastic programming $G(x, \xi)$ is the optimal value of the second stage program.

The feasible set X can be finite, i.e., integer first stage problem. Both stages can be integer (mixed integer) problems.

- **How difficult is the above two-stage problem?**
- **What about multistage problems?**

Suppose that P has a finite support, i.e., $\xi(\omega)$ can take values $\xi_1, ..., \xi_K$ with respective probabilities $p_1, ..., p_K$. In that case $\mathbb{E}_P[G(x, \xi(\omega))] = \sum_{k=1}^K p_k G(x, \xi_k)$. The number K (number of scenarios), however, grows **exponentially** with dimension of the data $\xi(\omega)$.

Monte Carlo sampling approach

Let $\xi^1, ..., \xi^N$ be a generated (iid) random sample drawn from P. Then by the Law of Large Numbers, for a given $x \in X$, we have

$$N^{-1} \sum_{j=1}^{N} G(x, \xi^j) \to \mathbb{E}_P[G(x, \xi(\omega))] \quad w.p.1.$$

The sample average $\hat{g}_N(x) := N^{-1} \sum_{j=1}^{N} G(x, \xi^j)$ is an unbiased and consistent estimate of $g(x) = \mathbb{E}_P[G(x, \xi(\omega))]$.

Notoriously slow convergence of order $O_p(N^{-1/2})$. In order to improve the accuracy by one digit the sample size should be increased 100 times.

By the Central Limit Theorem

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

where $\sigma^2(x) := \mathbb{V}ar[G(x, \xi(\omega)]$.
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 approaches

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. 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:

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

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 $N^{1/2}\,[\hat{g}_N(x_1) - \hat{g}_N(x_2)]$ is

$$\mathbb{V}ar[G(x_1,\xi)] + \mathbb{V}ar[G(x_2,\xi)] - 2\,\mathbb{C}ov[G(x_1,\xi), G(x_2,\xi)].$$

It can be much smaller than $\mathbb{V}ar[G(x_1,\xi)]+\mathbb{V}ar[G(x_2,\xi)]$, 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 problem
$\hat{v}_N$ is the optimal value of the SAA problem
$\hat{S}_N^\varepsilon$ is the set of ε-optimal solutions of the SAA problem
$\hat{x}_N$ 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, $\hat{g}_N(x)$ converge (pointwise) to $g(x)$ w.p.1. Under mild additional conditions, this implies that $\hat{v}_N \to v^0$ and $\text{dist}(\hat{x}_N, S^0) \to 0$ 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.

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

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 $\hat{v}_N$ converges to v^0 at the same rate as $\hat{g}_N(x^0)$ converges to $g(x^0)$.

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

The required regularity conditions are that the expected value function $g(x)$ is smooth (twice differentiable) at x^0 and the Hessian matrix $\nabla^2 g(x^0)$ is positive definite. This typically happens if the probability distribution P is *continuous*. In such cases $\hat{x}_N$ 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. For any given $\epsilon > 0$, $\mathbb{P}(\|\hat{x}_N - x^0\| \geq \epsilon)$ approaches zero exponentially fast as $N \to \infty$ (Kaniovski, King and Wets, 1995).

Complexity issues

Suppose that the feasible set X is *finite*. Consider a mapping $u : X \setminus S^\varepsilon \to S^0$, and

$$H(x,\omega) := G(u(x),\xi(\omega)) - G(x,\xi(\omega)).$$

Suppose that for every $x \in X$ the moment generating function of $H(x,\omega)$ is finite valued in a neighborhood of zero. Let ε and δ be nonnegative numbers such that $\delta \leq \varepsilon$. Then there is $\gamma(\delta,\varepsilon) > 0$ such that

$$P\left(\widehat{S}_N^\delta \not\subset S^\varepsilon\right) \leq |X| e^{-N\gamma(\delta,\varepsilon)}.$$

The constant $\gamma(\delta,\varepsilon)$ can be estimated

$$\gamma(\delta,\varepsilon) \geq \frac{(\varepsilon^* - \delta)^2}{3\sigma^2} > \frac{(\varepsilon - \delta)^2}{3\sigma^2},$$

where

$$\varepsilon^* := \min_{x \in X \setminus S^\varepsilon} g(x) - v^0 \;\text{ and }\; \sigma^2 := \max_{x \in X \setminus S^\varepsilon} \mathbb{V}ar[H(x,\omega)].$$

Note that $\varepsilon^* > \varepsilon$. This gives the following estimate of the sample size N which guarantees that $\mathbb{P}\left(\widehat{S}_N^\delta \subset S^\varepsilon\right) \geq 1 - \alpha$, for a given $\alpha \in (0,1)$,

$$N \geq \frac{3\sigma^2}{(\varepsilon - \delta)^2} \log\left(\frac{|X|}{\alpha}\right).$$

Kleywegt, Shapiro, Homem-de-Mello (2000).

The required sample size grows as a logarithm of $|X|$.

Now let X be a bounded subset of $\mathbb{R}^n$. Then for a given $\nu > 0$, consider a finite subset X_ν of X such that for any $x \in X$ there is $x' \in X_\nu$ satisfying $\|x - x'\| \le \nu$. If D is the diameter of the set X, then such set X_ν can be constructed with $|X_\nu| \le \left(\frac{D}{\nu}\right)^n$. Reducing the feasible set X to its subset X_ν, we obtain the following estimate of the required sample size to solve the reduced problem

$$N \ge \frac{3\sigma^2}{(\varepsilon - \delta)^2} \left[n \log \left(\frac{D}{\nu} \right) - \log \alpha \right].$$

Suppose that $g(x)$ is Lipschitz continuous modulus L. By taking $\nu := (\varepsilon - \delta)/(2L)$ we obtain the following estimate of the required sample size to solve the the true problem

$$N \ge \frac{12\sigma^2}{(\varepsilon - \delta)^2} \left[n \log \left(\frac{2DL}{(\varepsilon - \delta)^2} \right) - \log \alpha \right].$$

This suggests a **linear** growth of the required sample size with the dimensionality n of the first stage problem.

Convergence of subdifferentials

Suppose that $G(\cdot, \xi(\omega))$ is convex for a.e. $\omega \in \Omega$ and $g(\cdot)$ is finite. Then

$$g'(x, d) = \mathbb{E}_P\left[G'_\omega(x, d)\right],$$

$$\lim_{N\to\infty} \sup_{\|d\|\le 1} \left|g'(x, d) - \widehat{g}'_N(x, d)\right| = 0, \;\; w.p.1,$$

$$\lim_{N\to\infty} \mathcal{H}\left(\partial g(x), \partial \widehat{g}_N(x)\right) = 0, \;\; w.p.1,$$

where $\mathcal{H}(\cdot, \cdot)$ denotes the Hausdorff distance between sets and $G'_\omega(x, d)$ is the directional derivative of $G(\cdot, \xi(\omega))$.

Suppose, further, that:
(i) the distribution P has a finite support, i.e., finite number of scenarios,
(ii) for every $\omega \in \Omega$ the function $G(\cdot, \xi(\omega))$ is piecewise linear and convex.

Then the expected value function $g(x)$ is convex piecewise linear, and
(a) the subdifferentials $\partial g(x)$, $\partial \widehat{g}_N(x)$ are polyhedrons,
(b) there is a correspondence between extreme points of $\partial \widehat{g}_N(x)$ and a subset of extreme points of $\partial g(x)$,
(c) w.p.1 for N large enough there is one-to-one correspondence between extreme points of $\partial \widehat{g}_N(x)$ and extreme points of $\partial g(x)$, and distances between these extreme points tend to zero as $N \to \infty$.

Suppose that the true problem is **convex piecewise linear**, i.e.,
(i) the distribution P has a finite support,
(ii) for every $\omega \in \Omega$ the function $G(\cdot, \xi(\omega))$ 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 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 $\hat{S}_N$ of optimal solutions of the SAA problem is nonempty and forms a face of the (polyhedral) set S^0.

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

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

(Shapiro & Homem-de-Mello, 2000)

Well and ill conditioned problems

Suppose that the problem is *convex piecewise linear*, and let x^0 be unique optimal solution of the true problem. Then

$$g'(x^0, d) > 0, \quad \forall\, d \in T_X(x^0) \setminus \{0\}.$$

Furthermore, there exists a *finite* set $\{d_1, ..., d_\ell\} \subset T_X(x_0)$ of nonzero directions, independent of the sample, such that if $\widehat{g}'_N(x^0, d_j) > 0$ for $j = 1, ..., \ell$, then $\widehat{x}_N = x^0$.

We call

$$\kappa := \max_{j \in \{1,...,\ell\}} \frac{\mathbb{V}ar[G'_\omega(x^0, d_j)]}{[g'(x^0, d_j)]^2}$$

the condition number of the true problem. Recall that $\mathbb{E}\left[G'_\omega(x^0, d)\right] = g'(x^0, d)$.

For convex piecewise linear problems with unique optimal solution, the exponential rate holds and the corresponding constant γ is approximately equal to $(2\kappa)^{-1}$. This means that the sample size N required to achieve a given probability of the event "$\widehat{x}_N = x^0$" is roughly proportional to the condition number κ. More accurately, for large N and κ,

$$P(\widehat{x}_N \neq x^0) \approx \frac{C e^{-N/(2\kappa)}}{\sqrt{4\pi N/(2\kappa)}},$$

where C is a positive constant independent of the sample.

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 $g(\hat{x}) - v^0$.

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

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

(i) Estimate $g(\hat{x})$ by the sample average $\hat{g}_{N'}(\hat{x})$, 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

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

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 $(G(x_1, \xi(\omega)), ..., G(x_k, \xi(\omega)))$.

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

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

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 $G(\cdot, \xi(\omega))$ is differentiable at $\hat{x}$ w.p.1 and

$$\nabla g(\hat{x}) = \mathbb{E}_P \left[\nabla_x G(\hat{x}, \xi(\omega))\right].$$

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

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

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

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

by using the sample estimator

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

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

Multistage stochastic programming

Nested formulation

$$\operatorname*{Min}_{\substack{A_{11}x_1=b_1\\ x_1\geq 0}} c_1^T x_1+\mathbb{E}\left[\operatorname*{Min}_{\substack{A_{21}x_1+A_{22}x_2=b_2\\ x_2\geq 0}} c_2^T x_2+\cdots+\mathbb{E}[\operatorname*{Min}_{\substack{A_{T,T-1}x_{T-1}+A_{TT}x_T=b_T\\ x_T\geq 0}} c_T^T x_T]\right].$$

Scenario tree

Scenario is a path. What is a right way of sampling? Conditional sampling versus scenario sampling.