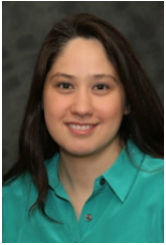
Scenario Generation and Sampling Methods

Güzin Bayraksan
Tito Homem-de-Mello

SVAN 2016 – IMPA
May 9th, 2016

The team:



Güzin Bayraksan
Integrated Systems Engineering,
The Ohio State University,
Columbus, Ohio

Tito Homem-de-Mello
School of Business,
Universidad Adolfo Ibañez,
Santiago, Chile

Linear programs with uncertainty

Many optimization problems can be formulated as linear programs:

$$\begin{aligned} &\min b^T x \\ &\text{s.t. } Ax \geq c. \end{aligned} \tag{LP}$$

Suppose there is some uncertainty in the coefficients A and c.

For example, the constraint $Ax \geq c$ could represent "total energy production must satisfy demand", but

- Demand is uncertain.
- Actual *produced* amount from each energy source is a (random) percentage of the *planned* amount.

What to do?

Dealing with uncertainty

Some possibilities:

- Impose that constraint $Ax \geq c$ must be satisfied regardless of the outcome of A and c.
- Impose that constraint $Ax \geq c$ must be satisfied with some probability, i.e., solve

$$\min \left\{ b^T x \;:\; P(Ax \geq c) \geq 1 - \alpha \right\} \qquad \text{for some small } \alpha > 0.$$

- Penalize the *expected constraint violation*, i.e., solve

$$\min\ b^T x + \mu \mathbb{E}[\max\{c - Ax, 0\}] \qquad \text{for some } \mu > 0.$$

Difficulty: How to solve any of these formulations?

The need for approximation

Even before we think of optimization methods to solve the above problems, we need to deal with an even more basic issue:

- How to *compute* quantities such as $P(Ax \geq c)$ or $\mathbb{E}[\max\{c - Ax, 0\}]$?
- Very hard to do! (except in special cases)
- We need to approximate these quantities with something we can compute.

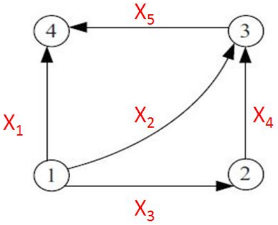
The estimation problem: an example

Suppose we have a vector of m random variables $X := (X_1, \ldots, X_m)$ and we want to calculate

$$g \;:=\; \mathbb{E}[G(X)] \;=\; \mathbb{E}[G(X_1, \ldots, X_m)],$$

where G is a function that maps m-dimensional vectors to the real numbers.

Example: find the expected completion time of a project.



Project has 3 components, given by activities

1

2 and 5

3, 4 and 5

.

$G(X) = \max\{X_1, X_2 + X_5, X_3 + X_4 + X_5\}$

The estimation problem

How to do that?

- Suppose that each variable X_k can take r possible values, denoted $x_k^1, \ldots, x_k^r$. If we want to compute the exact value, we have to compute

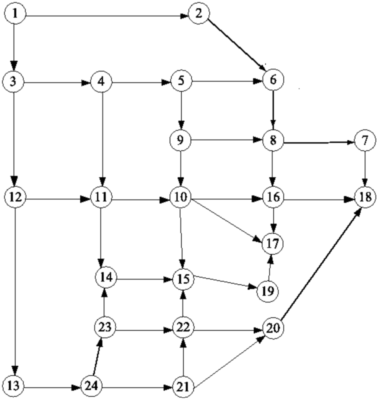
$$\mathbb{E}[G(X)] = \sum_{k_1=1}^{r} \sum_{k_2=1}^{r} \cdots \sum_{k_m=1}^{r} G(x_1^{k_1}, \ldots, x_m^{k_m})\, P(X_1 = x_1^{k_1}, \ldots, X_m = x_m^{k_m})$$

- In the above example, suppose each variable can take $r = 10$ values. If the travel times are independent, then we have a total of

$$10^5 \;=\; 100,000 \;\; \text{possible outcomes for } G(X)!$$

The estimation problem

Imagine now this project:



Path 1: 1-2-6-8-7-18
Path 2: 1-2-6-8-16-18
Path 3: 1-3-4-11-10-16-18
Path 4: 1-3-4-5-6-8-7-18
Path 5: 1-3-4-5-6-8-16-18
Path 6: 1-3-4-5-9-8-7-18
Path 7: 1-3-4-5-9-8-16-18
Path 8:1-3-4-5-9-10-16-18
Path 9:1-3-12-11-10-16-18
Path10:1-3-12-13-24-21-20-18
Path11:1-3-12-13-24-21-22-20-18
Path12:1-3-12-13-24-23-22-20-18

It is totally impractical to calculate the exact value!

- The problem is even worse if the distributions are *continuous*.

The need for scenarios

The example shows that we need a method that can help us approximate distributions with a finite (and not too large) set of scenarios.

Issues:

- How to select such a set of scenarios?
- What guarantees can be given about the quality of the approximation?

As we shall see, there are two classes of approaches:

- *Sampling* methods
- *Deterministic* methods

Each class requires its own tools to answer the two questions above.

The estimation problem via sampling

Idea: Let $X^j := (X_1^j, \ldots, X_m^j)$ denote one sample from the random vector X.

- Draw N *independent and identically distributed* (iid) samples $X^1, \ldots, X^N$.
- Compute

$$\hat{g}_N \; := \; \frac{1}{N} \sum_{j=1}^{N} G(X^j).$$

Recall the *Strong Law of Large Numbers*: as N goes to infinity,

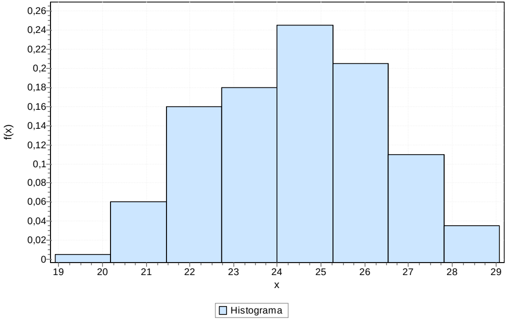
$$\lim_{N \to \infty} \hat{g}_N \; = \; \mathbb{E}[G(X)] \;\text{ with probability one (w.p.1)}$$

so we can use $\hat{g}_N$ as an approximation of $g = \mathbb{E}[G(X)]$.

Assessing the quality of the approximation

ISSUE: $\hat{g}_N$ is a random variable, since it depends on the sample.

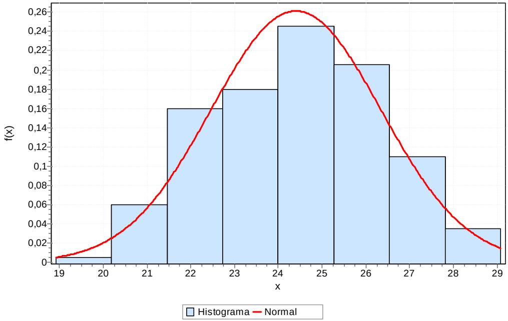
- That is, in one experiment $\hat{g}_N$ may be close to g while in another it may differ from g by a large amount!
- Example: 200 runs of the completion time problem with $N = 50$.



Assessing the quality of the approximation

ISSUE: $\hat{g}_N$ is a random variable, since it depends on the sample.

- That is, in one experiment $\hat{g}_N$ may be close to g while in another it may differ from g by a large amount!
- Example: 200 runs of the completion time problem with $N = 50$.



The Central Limit Theorem

Note that

$$\mathbb{E}[\hat{g}_N] \;=\; \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} G(X^j)\right] \;=\; \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[G(X^j)\right] \;=\; g.$$

Also,

$$\mathrm{Var}(\hat{g}_N) \;=\; \mathrm{Var}\left(\frac{1}{N}\sum_{j=1}^{N} G(X^j)\right) \;=\; \frac{1}{N^2}\sum_{j=1}^{N}\mathrm{Var}(G(X^j)) \;=\; \frac{1}{N}\mathrm{Var}(G(X)).$$

The Central Limit Theorem asserts that, for N sufficiently large,

$$\frac{\sqrt{N}\,(\hat{g}_N - g)}{\sigma} \;\approx\; \mathrm{Normal}(0,1),$$

where $\sigma^2 = \mathrm{Var}(G(X))$.

Computing the margin of error of the estimate

The CLT implies that

$$P\left(\hat{g}_N - 1.96\frac{\sigma}{\sqrt{N}} \leq g \leq \hat{g}_N + 1.96\frac{\sigma}{\sqrt{N}}\right) \;=\; 0.95.$$

That is, out of 100 experiments, on average in 95 of those the interval given by

$$\left[\hat{g}_N - 1.96\frac{\sigma}{\sqrt{N}},\; \hat{g}_N + 1.96\frac{\sigma}{\sqrt{N}}\right]$$

will contain the true value g.

The above interval is called a 95% confidence interval for g.

- Note that σ^2 is usually unknown. Again, when N is large enough we can approximate σ^2 with

$$S_N^2 \;:=\; \frac{\sum_{j=1}^N \left(G(X^j) - \hat{g}_N\right)^2}{N-1}.$$

The estimation problem via deterministic approximation

One idea is to approximate the distribution of each X_i with a discrete distribution with small number of points (say, 3 points).

- But even then we have to sum up 3^m terms!
- Also, it is difficult to assess the quality of the approximation...
- How about quadrature rules to approximate integrals (e.g., Simpson's rule)?
  - They work well for *low-dimensional* problems.

From estimation to optimization

Consider a generic stochastic optimization problem of the form

$$\min_{x \in X} \{ g(x) := \mathbb{E}[G(x,\xi)] \}, \qquad \text{(SP)}$$

where:

- G is a real-valued function representing the quantity of interest (cost, revenues, etc.).
- The inputs for G are the decision vector x and a random vector ξ that represents the uncertainty in the problem.
- X is the set of feasible points.

The need for approximation

As before, if G is not a simple function, or if ξ is not low-dimensional, then we need to approximate the problem, since we cannot evaluate $g(x)$ exactly.

- As before, we can use either sampling or deterministic approximations.
- **Issue:** What is the effect of the approximation on the optimal value and/or optimal solutions of the problem?

The newsvendor problem, revisited

Newsvendor purchases papers in the morning at price c and sells them during the day at price r

Unsold papers are returned at the end of the day for salvage value s.

If we want to maximize the *expected* revenue, then we have to solve

$$\min_{x \geq 0} \{g(x) := \mathbb{E}[G(x,\xi)]\},$$

where

$$\begin{aligned} G(x,\xi) &:= -cx + r\min\{x,\xi\} + s(x - \min\{x,\xi\}) \\ &= (s-c)x + (r-s)\min\{x,\xi\} \end{aligned}$$

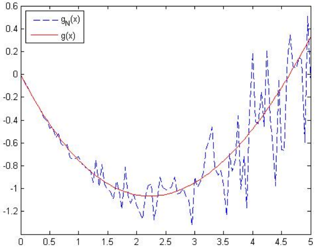
Approximation with sampling

As we saw before, we can approximate the value of $g(x)$ (for each given x) with a sample average.

That is, for each $x \in X$ we can draw a sample $\{\xi_x^1, \dots, \xi_x^N\}$ from the distribution of ξ, and approximate $g(x)$ with

$$\tilde{g}_N(x) \ := \ \frac{1}{N} \sum_{j=1}^{N} G(x, \xi_x^j).$$

But: It is useless to generate a new approximation for each x!

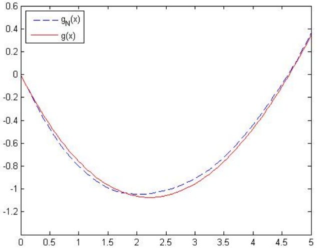


The Sample Average Approximation approach

The idea of the Sample Average Approximation (SAA) approach is to use the same sample for all x.

That is, we draw a sample $\{\xi^1, \ldots, \xi^N\}$ from the distribution of ξ, and approximate $g(x)$ with

$$\hat{g}_N(x) \;:=\; \frac{1}{N}\sum_{j=1}^{N} G(x, \xi^j).$$



The Sample Average Approximation approach

We can see that the approximation is very close to the real function.

This suggests replacing the original problem with

$$\min_{x \in X} \hat{g}_N(x),$$

which can be solved using a deterministic optimization algorithm!

Questions:

- Does that always work, i.e. for any function $G(x, \xi)$?
- What is a "good" sample size to use?
- What can be said about the quality of the solution returned by the algorithm?

Asymptotic properties

Let us study first what happens as the sample size N goes to infinity.

It is important to understand what that means. Consider the following *hypothetical* experiment:

- We draw a sample of infinite size, call it $\{\xi^1, \xi^2, \ldots\}$. We call that a sample path.
- Then, for each N, we construct the approximation

$$\hat{g}_N(\cdot) = \frac{1}{N} \sum_{j=1}^{N} G(\cdot, \xi^j)$$

  using the first N terms of that sample path, and we solve

$$\min_{x \in X} \hat{g}_N(x). \qquad (\mathrm{SP}_N)$$

Asymptotic properties

Let

$$\begin{aligned} \hat{x}_N &:= \text{an optimal solution of } (\mathrm{SP}_N) \\ S_N &:= \text{the set of optimal solutions of } (\mathrm{SP}_N) \\ \nu_N &:= \text{the optimal value of } (\mathrm{SP}_N) \end{aligned}$$

and

$$\begin{aligned} x^* &:= \text{an optimal solution of } (\mathrm{SP}) \\ S^* &:= \text{the set of optimal solutions of } (\mathrm{SP}) \\ \nu^* &:= \text{the optimal value of } (\mathrm{SP}) \end{aligned}$$

As the sample size N goes to infinity, does

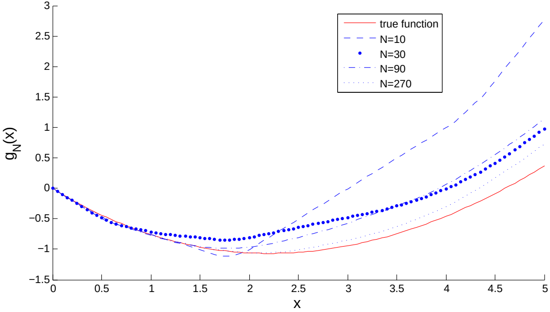
- $\hat{x}_N$ converge to some x^*?
- S_N converge to the set S^*?
- ν_N converge to ν^*?

Asymptotic properties, continuous distributions

We illustrate the asymptotic properties with the newsvendor problem.

We will study separately the cases when demand ξ has a continuous and a discrete distribution.

Suppose first demand has an Exponential(10) distribution.



Asymptotic properties, continuous distributions

It seems the functions $\hat{g}_N$ are converging to g. The table lists the values of $\hat{x}_N$ and ν_N ($N = \infty$ corresponds to the true function):

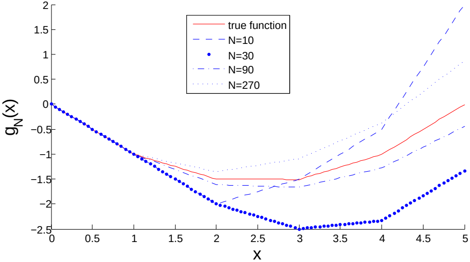
N	10	30	90	270	∞
$\hat{x}_N$	1.46	1.44	1.54	2.02	2.23
ν_N	-1.11	-0.84	-0.98	-1.06	-1.07

So, we see that $\hat{x}_N \to x^*$ and $\nu_N \to \nu^*$!

Asymptotic properties, discrete distributions

Now let us look at the case when ξ has a discrete distribution.

Suppose demand has discrete uniform distribution on $\{1, 2, \ldots, 10\}$.



Asymptotic properties, discrete distributions

Again, it seems the functions $\hat{g}_N$ are converging to g. The table lists the values of $\hat{x}_N$ and ν_N ($N = \infty$ corresponds to the true function):

N	10	30	90	270	∞
$\hat{x}_N$	2	3	3	2	[2,3]
ν_N	-2.00	-2.50	-1.67	-1.35	-1.50

We see that $\nu_N \to \nu^*$. However, $\hat{x}_N$ does not seem to be converging at all.

- On the other hand, $\hat{x}_N$ is oscillating between two optimal solutions of the true problem!

How general is this conclusion?

Convergence result

We can see from both figures that $\hat{g}_N(\cdot)$ converges *uniformly* to $g(\cdot)$.

- Uniform convergence occurs for example when the functions are *convex*.

The following result is general:

Theorem

When uniform convergence holds, we have the following results:

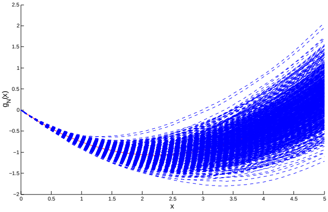
1. $\nu_N \to \nu^*$ *with probability one (w.p.1),*
2. *Suppose that there exists a compact set* C *such that (i)* $\emptyset \neq S^* \subseteq C$ *and* $\emptyset \neq S_N \subseteq C$ *w.p.1 for* N *large enough, and (ii) the objective function is finite and continuous on* C*. Then,* $dist(S_N, S^*) \to 0$ *w.p.1.*

Convergence result (cont.)

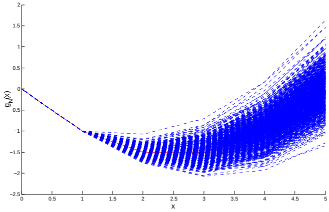
What does "convergence with probability one" means?

- Recall that the functions $\hat{g}_N$ in the above example were constructed from a *single* sample path.
- The theorem tells us that, *regardless of the sample path* we pick, we have convergence as $N \to \infty$!

So, let us repeat the above experiment (only for $N = 270$) multiple times, each time with a different sample path:



(a) Exponential demand



(b) Discrete uniform demand

Convergence result (cont.)

We see that for some sample paths we have a very good approximation for this N, (in this case, $N = 270$) but for others we don't.

Why? Don't we have convergence for all sample paths?

- The problem is the theorem only guarantees convergence as $N \to \infty$.
- So, for some path we quickly get a good approximation, whereas for others we may need a larger N to achieve the same quality.

So, if we pick *one sample of size N* and solve $\min \hat{g}_N(x)$ as indicated by the SAA approach, how do we know if we are on a "good" or on a "bad" sample path?

- The answer is...we don't!
- So, we need to have some probabilistic guarantees.