

# Stochastic programming

## SAA: adaptive sampling

Fabian Bastin

`bastin@iro.umontreal.ca`

Université de Montréal – CIRRELT – IVADO – Fin-ML

# Motivation

Reminder: we consider the stochastic problem

$$\min_{x \in S} g(x) = E_P [G(x, \xi)],$$

where

- $x \in \mathbb{R}^m$
- $S$  is a compact subset of  $\mathbb{R}^m$
- $\xi$  is a real random vector defined on  $(\Xi, \mathcal{F}, P)$  and taking values in  $(\mathbb{R}^k, \mathcal{B}^k)$  ( $\mathcal{B}^k$  is the Borel measure)
- $G : \mathbb{R}^m \times \mathbb{R}^k \rightarrow \mathbb{R}$

Sample average approximation:

$$\min_{x \in S} \hat{g}_N(x) = \frac{1}{N} \sum_{i=1}^N G(x, \xi_i),$$

# Convergence

In addition to our previous consistency results for  $N \rightarrow \infty$ , the central limit theorem tells us that, if the draws are independent and identically distributed (i.i.d.) (and finite  $g(x)$ ),

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

where  $\sigma^2(x) = \text{var}(G(x, \xi))$ , and  $\Rightarrow$  denotes convergence in distribution.

This result is only valid for a given  $x$ . It is necessary to set stronger conditions in order to have a functional convergence.

Note: under our assumptions,  $\hat{g}_N(x)$  is continuous over  $S$ , and can thus be considered as a point in the Banach space  $C(S)$ .

## Banach space $C(S)$

$C(S)$  is the space of continuous functions  $\psi : S \rightarrow \mathbb{R}$ , equipped with the sup-norm  $\|\psi\| := \sup_{x \in S} |\psi(x)|$

$C(S)$  is a Banach space, i.e. a normed vectorial space, complete under the distance issued from its norm. A metric space  $M$  is said complete or complete space if every Cauchy suite in  $M$  has a limit in  $M$  (i.e. it converges in  $M$ ).

We will extend the (pointwise) central limit theorem to a functional central limit theorem, assuming as usual that the draws are i.i.d.

# Assumptions

1.  $\forall x \in S$ ,  $G(x, \cdot)$  is measurable (i.e. its expectation exists).
2.  $\exists x \in S$  such that  $E_P[G(\bar{x}, \xi)^2] < \infty$ .
3. (Lipschitz continuity condition)  $\exists K(\xi) \geq 0$  such that  $E[K(\xi)]$  is finite, and  $\forall x_1, x_2 \in S$  and a.e.  $\xi$ ,

$$|G(x_1, \xi) - G(x_2, \xi)| \leq K(\xi) \|x_1 - x_2\|,$$

We assume moreover that  $E[K^2(\xi)] < \infty$ .

# Functional central limit theorem

Under these conditions,

$$N^{1/2}[\hat{g}_n - g] \Rightarrow Y \in C(S).$$

If  $x_1, \dots, x_k \in S$ ,  $(Y(x_1), \dots, Y(x_k)) \sim N(0, \Sigma)$ , where  $\Sigma$  is the covariance matrix from  $(G(x_1, \xi), \dots, G(x_k, \xi))$ .

If  $N^{1/2}(\hat{g}_N - g) \Rightarrow Y \in C(S)$  (with  $\{\hat{g}_n\}$  and  $g$  in  $C(S)$ ), and

$$\hat{v}_N = \min_{x \in S} \hat{g}_N(x) \text{ et } v^* = \min_{x \in S} g(x),$$

then

$$N^{1/2}(\hat{v}_N - v^*) \Rightarrow \min_{x \in S^*} Y(x).$$

## Convergence of global solutions

If  $S^*$  is a singleton, under the previous assumptions, in the i.i.d. case,

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

Under some additional conditions, we also have the convergence of  $E[\hat{v}_N]$  to  $v^*$ .

But all these results become difficult to extend in the case of local optimization.

However, we see that the results should be better with larger  $N$ , but at a higher computational cost, as

$$\hat{g}_N(x) = \frac{1}{N} \sum_{i=1}^N G(x, \xi_i).$$

# External adaptive method

What does interest us?

$$\min_{x \in S} \hat{g}_N(x) = \frac{1}{N} \sum_{i=1}^N G(x, \xi_i).$$

We can start with a small sample and extend it over the iterations: the adaptive sampling procedure can be **external** to the algorithm, or **internal**.

An external approach is known as **retrospective approximation** consists to repeatedly apply the optimization algorithm with samples of increasing sizes.



# Retrospective approximation (RA)

Introduced by Healy and Schruben in 1991.

Let  $k$  be the iteration index. Components:

1. A procedure for solving a generated sample-path problem to specified tolerance vector  $\epsilon_k$ , delivering a solution  $x_k$ .
2. A sequence  $\{N_k\}$  of sample sizes tending to infinity.
3. A sequence  $\{\epsilon_k\}$  of error-tolerances tending to zero.
4. A sequence of weights  $\{w_{kj}, j = 1, 2, \dots, k\}$  for each iteration.

We define

$$\bar{x}_k := \sum_{j=1}^k w_{kj} x_j.$$

# Retrospective approximation: principle

- Step 0. Set  $k = 1$  and choose  $\bar{x}_0 = x_0 \in S$ .
- Step 1. Generate a sample-path problem with sample size  $N_k$ , with a “warm start”, i.e. starting from  $\bar{x}_{k-1}$  to solve the generated problem with the error-tolerance  $\epsilon_k$ . Denote the obtained retrospective solution by  $x_k$ .
- Step 2. Compute the solution  $\bar{x}_k$ .
- Step 3. Set  $k \leftarrow k + 1$  and go to Step 1.

## Retrospective approximation: assumptions

- The true function  $g$  has a unique minimizer  $x^* \in S$ .
- $G(\cdot, \xi)$  is Lipschitz with Lipschitz constant  $L(\xi)$  on  $S$  a.s., and  $E[L(\xi)] < \infty$ .
- $G(\cdot, \xi)$  is continuously differentiable at any  $x \in \mathcal{B}(x^*, \epsilon)$ ,  $\epsilon > 0$ , a.s.
- $\exists x \in S$  such that  $E[\|\nabla_x G(x, \xi)\|^2] < \infty$ .
- The sample function  $\hat{g}_N(x)$  has a unique minimum  $x_N^*$  a.s.
- When  $\hat{g}_N(x)$  attains a unique minimum  $x_N^*$ ,  $\hat{g}_N(x)$  is twice differentiable at  $x_N^*$ . Furthermore, the  $\nabla^2 \hat{g}_N(x_N^*)$  is positive definite with smallest eigenvalue uniformly bounded away from 0 a.s.
- The solution  $x_k$  obtained from the  $k^{th}$  iteration of RA satisfies  $\|\nabla \hat{g}_{N_k}(x_k)\| \leq \epsilon_k$ .

## Retrospective approximation: assumptions

- The numerical procedure used to solve the sample-path problems in RA exhibits  $p^{th}$ -order sublinear convergence or  $p^{th}$ -order linear convergence with respect to the observed derivatives.
- The sample sizes are increased linearly, i.e.,  $N_k/N_{k-1} = c > 1$  for all  $k$ .
- The error-tolerances are chosen so that  $\epsilon_k = O(1/\sqrt{N_k})$ .

## Retrospective approximation: convergence rate

Under the previous assumptions,

$$C_k \|x_k - x^*\|^2 = O_p(1),$$

as  $k \rightarrow \infty$ , where  $C_k$  is the total amount of computational work done until the  $k^{th}$  iteration and is given by  $C_k = \sum_{i=1}^k Q_i N_i$ . Here  $Q_i$  is the number of points visited by the numerical procedure during the  $i$ th iteration.

We recover the convergence rate of stochastic approximation method.

## Internal-external adaptive method

The major issue with this procedure is how to quantify the word “approximative” in Step 1. If no care is taken, the resulting algorithm can in fact be more time-consuming than the direct minimization of  $\hat{g}_{N_{\max}}$ .

We can also replace the stopping test on  $N_{\max}$  by a test of the criticality conditions of optimality.

The **internal approach** is a non-monotone strategy that depends on the underlying optimization methods. Here, we consider the unconstrained case.

More precisely, we generate a sample before the optimization process, with  $N_{\max}$  i.i.d. random draws. At iteration  $k$ , we will use a subset of this initial sample, using  $N_k$  of the  $N_{\max}$  random draws, typically the first ones.

## Accuracy estimation

This implies that  $\hat{g}_N$  is a smooth function, well defined for each choice of  $N$ .

In order to determine a sample size, we can measure the approximation accuracy. Let  $\alpha_\delta$  be the quantile of a  $\mathcal{N}(0, 1)$  associated to some significance level  $\delta$ , i.e.

$P_\xi[-\alpha_\delta \leq Y \leq \alpha_\delta] = \delta$ , where  $Y \sim \mathcal{N}(0, 1)$ .

We will use the central limit theorem

$$g(x) - \hat{g}_N(x) \Rightarrow \mathcal{N}\left(0, \frac{\sigma^2(x)}{N}\right),$$

where  $\sigma^2(x)$  is the variance of  $g$ , taken at the point  $x$ , in order to build a confidence interval for  $g(x)$  around  $\hat{g}_N(x)$ , as

$$[\hat{g}_N(x) - \epsilon_N^\delta(x), \hat{g}_N(x) + \epsilon_N^\delta(x)],$$

## Accuracy estimation (cont'd)

$\epsilon_N^\delta(x)$  is given by

$$\epsilon_N^\delta(x) = \alpha_\delta \frac{\sigma(x)}{\sqrt{N}}.$$

Typically, we will choose  $\alpha_{0.9} \approx 1.64$  or  $\alpha_{0.95} \approx 1.96$ .

In practice, we do not know  $\sigma^2(x)$ , but we can use its estimator

$$\hat{\sigma}_N^2(x) = \frac{1}{N-1} \sum_{i=1}^N (G(x, \xi_i) - \hat{g}_N(x))^2.$$

We will exploit this error estimation in the context of trust-region methods.



## Basic principles

The basic idea is that if the model approximates the objective function well enough w.r.t. the accuracy of the objective function (which depends on the sample size), we presume that we could work with a less accurate approximation, and therefore reduce the sample size.

On the other hand, if the adequation of the model with respect to the accuracy of the objective function is poor, we can increase the sample size in an attempt to correct this deficiency.

We assume the assumptions developed for the consistency analysis hold.

A formal algorithm description follows.

# Algorithm BTRDA (Bastin, Cirillo, Toint, 2006)

(Basic) trust-region algorithm with dynamic accuracy.

**Step 0. Initialization.** initial point:  $x_0$ , initial trust-region radius:  $\Delta_0$ . Set  $\eta_1$  and  $\eta_2$  such that  $0 < \eta_1 \leq \eta_2 < 1$  (for instance,  $\eta_1 = 0.01$  and  $\eta_2 = 0.75$ ),  $N_{\min} = N_{\min}^0$  and  $N_0$  satisfying  $\|\nabla \hat{g}_{N_0}(x_0)\| \neq 0$  if  $\epsilon_{\delta}^{N_0}(x_{k+1}) \neq 0$ , except if  $N_0 = N_{\max}$ . Compute  $\hat{g}_{N_0}(x_0)$  and set  $k = 0$ ,  $t = 0$ .

**Step 1. Stopping test.** Stop if  $\|\nabla \hat{g}_{N_k}(x_k)\| = 0$  and either  $N_k = N_{\max}$ , either  $\epsilon_{\delta}^{N_k}(x_k) = 0$ . Otherwise, go to Step 2.

**Step 2. Model definition** Define a model  $m_k^{N_k}$  of  $\hat{g}_{N_k}(x)$  in  $\mathcal{B}_k$ . Compute a new adequate sample size  $N^+$ , and set  $N^- = N_k$ .

## Algorithm BTRDA (cont'd)

**Step 3. Step computation** Compute a step  $s_k$  that sufficiently reduces  $m_k^{N_k}$  and s.t.  $x_k + s_k \in \mathcal{B}_k$ . Set

$$\Delta m_k^{N_k} = m_k^{N_k}(x_k) - m_k^{N_k}(x_k + s_k).$$

**Step 4. Comparaision of decreases** Compute  $\hat{g}_{N^+}(x_k + s_k)$  and

$$\rho_k = \frac{\hat{g}_{N_k}(x_k) - \hat{g}_{N^+}(x_k + s_k)}{\Delta m_k^{N_k}}.$$

**Step 5. Sample size update** If  $\rho_k < \eta_1$  and  $N_k \neq N^+$ , modify  $N^-$  or the candidate sample size  $N^+$  in order to take account of variance differences. Update  $\rho_k$ .

## Algorithm BTRDA (cont'd)

Step 6. Candidate iterate acceptance If  $\rho_k < \eta_1$ , define

$x_{k+1} = x_k$ ,  $N_{k+1} = N^-$ . Otherwise, define  
 $x_{k+1} = x_k + s_k$  and set  $N_{k+1} = N^+$ ; increment  $t$ .

If  $N_{k+1} \neq N^{\max}$ ,  $\|\nabla \hat{g}_{N_{k+1}}(x_{k+1})\| = 0$ , and  
 $\epsilon_{\delta}^{N_{k+1}}(x_{k+1}) \neq 0$ , increase  $N_{k+1}$  to some size less  
or equal to  $N_{\max}$  such that  $\|\nabla \hat{g}_{N_{k+1}}(x_{k+1})\| \neq 0$  if  
 $N_{k+1} \neq N_{\max}$ , and compute  $\hat{g}_{N_{k+1}}(x_{k+1})$ .

If  $N_k = N_{k+1}$  or if a sufficient decrease has been  
observed since the last evaluation of  $\hat{g}_{N_{k+1}}$ , set  
 $N_{\min}^{k+1} = N_{\min}^k$ . Otherwise, set  $N_{\min}^{k+1} > N_{\min}^k$ .

## Algorithm: BTRDA (cont'd)

### Step 7. Trust region radius update

$$\Delta_{k+1} \in \begin{cases} [\Delta_k, \infty) & \text{if } \rho_k \geq \eta_2, \\ [\gamma_2 \Delta_k, \Delta_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k] & \text{if } \rho_k < \eta_1, \end{cases}$$

In this algorithm the variable  $t$  is used to count the number of successful iterations.

Note: the algorithms BTR and BTRDA coincide if we fix  $N_k$  to  $N_{\max}$  for all  $k \geq 0$ .

## Variable sample size strategy

- Before the optimization, the user chooses a maximal sample size  $N_{\max}$  (for instance the number of observations).
- A minimum sample size  $N_{\min}^0$ .
- Define  $N_0 = \max\{N_{\min}^0, 0.1 N_{\max}\}$  if  $\|\nabla \hat{g}_{N_0}(x_0)\| \neq 0$  and  $\epsilon_{\delta}^{N_0}(x_0) \neq 0$ ,  $N_0 = N_{\max}$  otherwise.

## Choice of $N^+$ in (Step 3)

Sample size required to obtain an accuracy equal to the model decrease:

$$N^S = \max \left\{ N_{\min}^k, \left\lceil \frac{\alpha_{\delta}^2 \hat{\sigma}_N^2(x)}{(\Delta m_k^{N_k})^2} \right\rceil \right\}.$$

Set

$$\tau_1^k = \frac{\Delta m_k^{N_k}}{\epsilon_{\delta}^{N_k}(x_k)}, \quad \tau_2^k = \frac{N_k}{\min\{N_{\max}, N^S\}}.$$

Set  $N^+ = \max\{N', N_{\min}^k\}$ , where

$$N' = \begin{cases} \min\{\lceil \chi_1 N_{\max} \rceil, \lceil N^S \rceil\} & \text{if } \tau_1^k \geq 1, \\ \min\{\lceil \chi_1 N_{\max} \rceil, \lceil \tau_1^k N^S \rceil\} & \text{if } \tau_1^k < 1 \text{ and } \tau_1^k \geq \tau_2^k, \\ \lceil \chi_1 N_{\max} \rceil & \text{if } \nu_1 \leq \tau_1^k < 1 \text{ and } \tau_1^k < \tau_2^k, \\ N_{\max} & \text{if } \tau_1^k < \nu_1 \text{ and } \tau_1^k < \tau_2^k. \end{cases}$$

## Variable sample size strategy (cont'd)

A possible value for  $\chi_1$  is 0.5.

- $\tau_1^k \geq 1$ : the model decrease is greater or equal to the estimated accuracy.
- $\tau_1^k < 1$ : the improvement is smaller than the accuracy. However, a sufficient improvement during several consecutive iterations can lead to a significant decrease.
  - If  $\tau_1^k \geq \tau_2^k$ , the ratio between the current sample size and the potential next one is smaller than the ratio between the model decrease and the estimated error. We capitalize on  $\tau_1^k$  by computing a sample size smaller than  $N^s$ , such that an improvement of the order  $\epsilon_\delta^{N_k}(z_k)$  would be reached in approximatively  $\lceil \tau_1^k \rceil$  iterations if  $\tau_1^j$  is similar to  $\tau_1^k$  for  $j$  close to  $k$ .
  - If  $\tau_1^k < \tau_2^k$ , it can nevertheless be cheaper to continue to work with a smaller sample size, defined again as  $\lceil \chi_1 N_{\max} \rceil$ , as long as  $\tau_1^k$  is greater to some threshold  $\nu_1$  (for instance 0.2).



## Accuracy differences

If  $N^+$  is not equal to  $N_k$ , the computation of

$$\hat{g}_{N_k}(x_k) - \hat{g}_{N^+}(x_k + s_k)$$

is affected by the change in approximation variance. This can lead to  $\rho_k < \eta_1$ , even if the model  $m_k^{N_k}$  gives a good prediction for the sample size  $N^k$ .

If  $N^+ > N_k$ , compute  $\hat{g}_{N^+}(x_k)$ ,  $\Delta m_k^{N^+}$  and  $\epsilon_\delta^{N^+}(x_k)$ , otherwise if  $N^+ < N_k$  compute  $\hat{g}_{N^+}(x_k + s_k)$ . Set  $N^-$  to  $\max\{N_k, N^+\}$ , and

$$\rho_k = \frac{\hat{g}_{N^-}(x_k + s_k) - \hat{g}_{N^-}(x_k)}{\Delta m_k^{N^-}}.$$

## Minimum sample size update

While we expect to take advantage of smaller sample sizes when we are far from the solution, we should be sure to use a sample size equal to  $N_{\max}$  during the final iteration, in order to work with the desired accuracy.

To this end, we increase the minimum sample size when the adaptive strategy does not deliver sufficient numerical gains. A simpler strategy consists to increase  $N_{\min}$  by 1 at each iteration.

## Additional safeguards

Finally, note that if  $N_k \neq N_{\max}$ , we cannot exclude the pathological case in which  $z_k$  is a critical first-order point for  $\hat{g}_{N_k}$ . If  $\epsilon_\delta^{N_k}(x_k) \neq 0$ , the algorithm does not stop, but since the model is quadratic, no decrease is achieved if  $H_k$  is positive definite. In order to avoid this situation, we force  $N_{k+1}$  to increase when it occurs.

In practice, the gradient norm usually changes slowly in the neighborhood of such a critical point, and a small gradient typically leads to a small decrease of the model, implying the increase of the sample size, and  $N_{\max}$  is reached before this safeguard is deployed.

# Convergence

## Theorem

*Under some regularity assumptions, if*

$$\exists \kappa > 0 \text{ such that } \epsilon_{\delta}^{N_k}(x_k) \geq \kappa,$$

*for all  $k$  large enough, then, almost surely, the algorithm converges in a finite number of iterations with a final sample size equal to  $N_{\max}$ , or the number of iterations is infinite and there exists some  $j$  such that for all the iterations  $i$ ,  $i \geq j$ ,  $N_i$  is equal to  $N_{\max}$ .*

## Example

Mixed logit model: stochastic maximum likelihood

$$\max_{\theta} LL(\theta) = \max_{\theta} \frac{1}{N} \sum_{n=1}^N \ln E[P_{ij}(\mathbf{x}, \theta, \xi)].$$

Mode choice model: **Mobidrive** data (Axhausen et al., 2002)  
 $N = 5799$  observations,  $R_{\max} = 2000$  draws per individual, 14 parameters (integration dimension: 3 normal variables).

