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

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

where

- $x \in \mathbb{R}^m$
- S is a compact subset of $\mathbb{R}^m$
- $\boldsymbol{\xi}$ 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)
- $G : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}$

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 $N \to \infty$, 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 $\psi : \mathrm{S} \to \mathbb{R}$, equipped with the sup-norm $\|\psi\| := \mathsf{sup}_{\mathrm{x}\in\mathrm{S}} |\psi|$

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. $\exists \overline{x} \in S$ such that $E_P[G(\overline{x}, \xi)^2] < \infty$.
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. ξ,

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

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

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 if 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 ϵ_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 $\{w_{kj},\ j = 1, 2, \ldots, k\}$ for each iteration.

We define

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

Retrospective approximation: principle

Step 0. Set $k = 1$ and choose $\overline{x}_0 = x_0 \in S$.

Step 1. Generate a sample-path problem with sample size N_k, with a "warm start", i.e. starting from $\overline{x}_{k-1}$ to solve the generated problem with the error-tolerance ϵ_k. Denote the obtained retrospective solution by x_k.

Step 2. Compute the solution $\overline{x}_k$.

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., $\mathrm{N_k} = \mathrm{cN_{k-1}} > 1$ for all k.
- The error-tolerances are chosen so that $\epsilon_{\mathrm{k}} = \mathrm{O}(1/\sqrt{\mathrm{N_k}})$.

Retrospective approximation: convergence rate

Under the previous assumptions,

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

as $k \to \infty$, 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 ith 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 that 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 α_δ be the quantile of a $\mathcal{N}(0,1)$ associated to some significance level δ, i.e.
$P_{\boldsymbol{\xi}}[-\alpha_\delta \leq Y \leq \alpha_\delta] = \delta$, 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)

$\epsilon_{\mathrm{N}}^{\delta}(\mathrm{x})$ is given by

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

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

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

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

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: Δ_0. Set η_1 and η_2 such that $0 < \eta_1 \leq \eta_2 < 1$ (for instance, $\eta_1 = 0.01$ and $\eta_2 = 0.75$), $N_{\mathsf{min}} = N^0_{\mathsf{min}}$ 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 $N_0 = N_{\mathsf{max}}$. 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 $N_k = N_{\mathsf{max}}$, 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. Comparaison 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 ρ_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 $N_{\min}^{k+1} > N_{\min}^k$.

Algorithme: 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 fixe N_k to N_{max} for all $k \geq 0$.

Variable sample size strategy

- Before the optimization, the user chooses a maximal sample size $\mathrm{N}_{\mathsf{max}}$ (for instance the number of observations).
- A minimum sample size $\mathrm{N}_{\mathsf{min}}^0$.
- Define $\mathrm{N}_0 = \mathsf{max}\{\mathrm{N}_{\mathsf{min}}^0, 0.1\mathrm{N}_{\mathsf{max}}\}$ if $\|\nabla \hat{\mathrm{g}}_{\mathrm{N}_0}(\mathrm{x}_0)\| \neq 0$ and $\epsilon_\delta^{\mathrm{N}_0}(\mathrm{x}_0) \neq 0$, $\mathrm{N}_0 = \mathrm{N}_{\mathsf{max}}$ otherwise.

Choice of N^+ in (Step 3)

Sample size required to obtain an accuracy equal to Δm_k:

$$N^s = \max\left\{N^k_{\min}, \left\lceil \frac{\alpha^2_\delta \hat{\sigma}^2_N(x)}{(\Delta m_k^{N_k})^2} \right\rceil\right\}.$$

Set

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

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

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

Variable sample size strategy (cont'd)

A possible value for χ_1 is 0.5.

- $\tau_1^k \geq 1$: $\Delta m_k \geq$ estimated accuracy.
- $\tau_1^k < 1$: $\Delta m_k <$ 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 Δm_k and the estimated error. We capitalize on τ_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 τ_1^j is similar to τ_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 τ_1^k is greater to some threshold ν_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 predicition 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 $\hat{g}_{N_k}(x_k + s_k)$. Set N^- to $\mathsf{max}\{N_k, N^+\}$, and

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

Additional tricks

- Minimum sample size update
  We want to be sure to use a sample size equal to N_{max} during the final iteration, in order to work with the desired accuracy. We increase N_{min} when the adaptive strategy does not deliver sufficient numerical gains, or simply by 1 at each iteration.
- Additional safeguards to avoid theoretical pathological cases (a critical point of the SAA with $N_k < N_{max}$ is found).

Convergence

Theorem

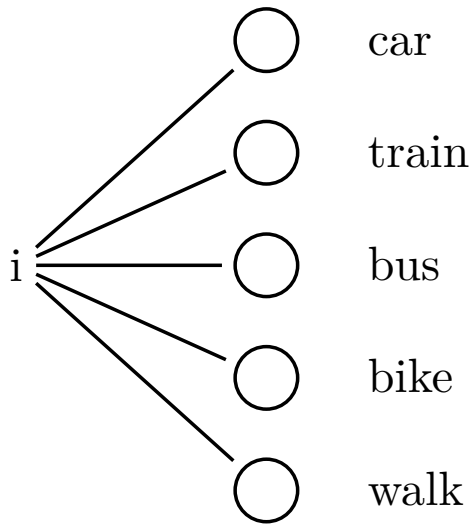
Under some regularity assumptions, if

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

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}.

Application to discrete choice theory

Discrete set of alternatives available for individual i: $\mathcal{A}(\mathrm{i})$.
Example: Which mode of transport have you chosen to come?



Each alternative has some utility.
Utility $\mathrm{U_{ij}}$ of $\mathrm{A_j} \in \mathcal{A}(\mathrm{i})$: $\mathrm{U_{ij}} = \mathrm{V}(\beta) + \boldsymbol{\epsilon}_{\mathrm{ij}}$.
β: parameters to be estimated.
rational individuals: a person chooses the alternative of maximum utility
$\rightarrow$ choice of $\mathrm{A_j}$ if $\mathrm{U_{ij}} \geq \mathrm{U_{in}}, \forall \mathrm{A_n} \in \mathcal{A}(\mathrm{i})$.

Important case: Gumbel distributed residuals ϵ_{ij} (mean 0, scale factor μ): $\rightarrow$ multinomial logit (MNL).

Probability that individual i choose $\mathrm{A_j}$: $\mathrm{P_{ij}} = \frac{\mathrm{e}^{\mu \mathrm{V_{ij}}(\beta)}}{\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{e}^{\mu \mathrm{V_{in}}(\beta)}}$

Mixed Logit models

Allow heterogeneity in parameters inside the population.

$$\boldsymbol{\beta} = \beta(\boldsymbol{\gamma}, \theta),$$

$\boldsymbol{\gamma}$: random vector, e.g. vector of independant $\mathrm{N}(0,1)$);
θ: vector of parameters, e.g. vector of means and std dev.
Probability choice of $\mathrm{A_j}$ by individual i:

$$\mathrm{P_{ij}}(\theta) = \mathrm{E}_{\boldsymbol{\xi}}\left[\mathrm{L_{ij}}(\gamma, \theta)\right] = \int \mathrm{L_{ij}}(\gamma, \theta)\mathrm{f}(\gamma)\mathrm{d}\gamma$$

This integral is approximated by

$$\mathrm{SP}^{\mathrm{R}}_{\mathrm{ij_i}} = \frac{1}{\mathrm{R}} \sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{L_{ij_i}}(\gamma_{\mathrm{r}}, \theta)$$

SAA problem:

$$\mathsf{max}_{\theta}\, \hat{\mathrm{g}}_{\mathrm{R}}(\theta) = \mathsf{max}_{\theta}\, \mathrm{SLL}(\theta) = \mathsf{max}_{\theta}\, \frac{1}{\mathrm{N}} \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathsf{ln}\, \mathrm{SP}^{\mathrm{R}}_{\mathrm{ij_i}}$$

$\rightarrow$ suggests to consider stochastic programming techniques.

Properties of Mixed Logit

Stochastic programming:

$$\min_z g(z) = \min_z E_{\boldsymbol{\xi}} [G(z, \boldsymbol{\xi})]$$

Mixed logit:

$$\max_\theta g(\theta) = \max_\theta LL(\theta) = \frac{1}{n} \sum_n \ln E_{\boldsymbol{\xi}} [L_{in}(\gamma, \theta)]$$

Consistency properties can be adapted. Assume I fixed and R grows toward ∞.

If θ_R^*, $R = 1, \ldots$, are first (second)-order critical for the corresponding SAA problem, under some regularity conditions, for almost every sequence of random draws, there exists some limit point θ^* of $(\theta_R^*)_{R=1}^{\infty}$ that is first (second)-order critical.

Error estimation

With an i.i.d. sample for each individual, we have, from the delta method (see for instance Shapiro and Rubinstein):

$$\mathrm{LL}(\theta) - \mathrm{SLL}^{\mathrm{R}}(\theta) \Rightarrow \mathrm{N}\left(0, \frac{1}{\mathrm{I}}\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{I}} \frac{\sigma^2_{\mathrm{ij_i}}(\theta)}{\mathrm{R}\left(\mathrm{P}_{\mathrm{ij_i}}(\theta)\right)^2}}\right)$$

Asymptotic value of the confidence interval radius:

$$\epsilon_\delta = \alpha_\delta \frac{1}{\mathrm{I}}\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{I}} \frac{\sigma^2_{\mathrm{ij_i}}(\theta)}{\mathrm{R}\left(\mathrm{P}_{\mathrm{ij_i}}(\theta)\right)^2}}$$

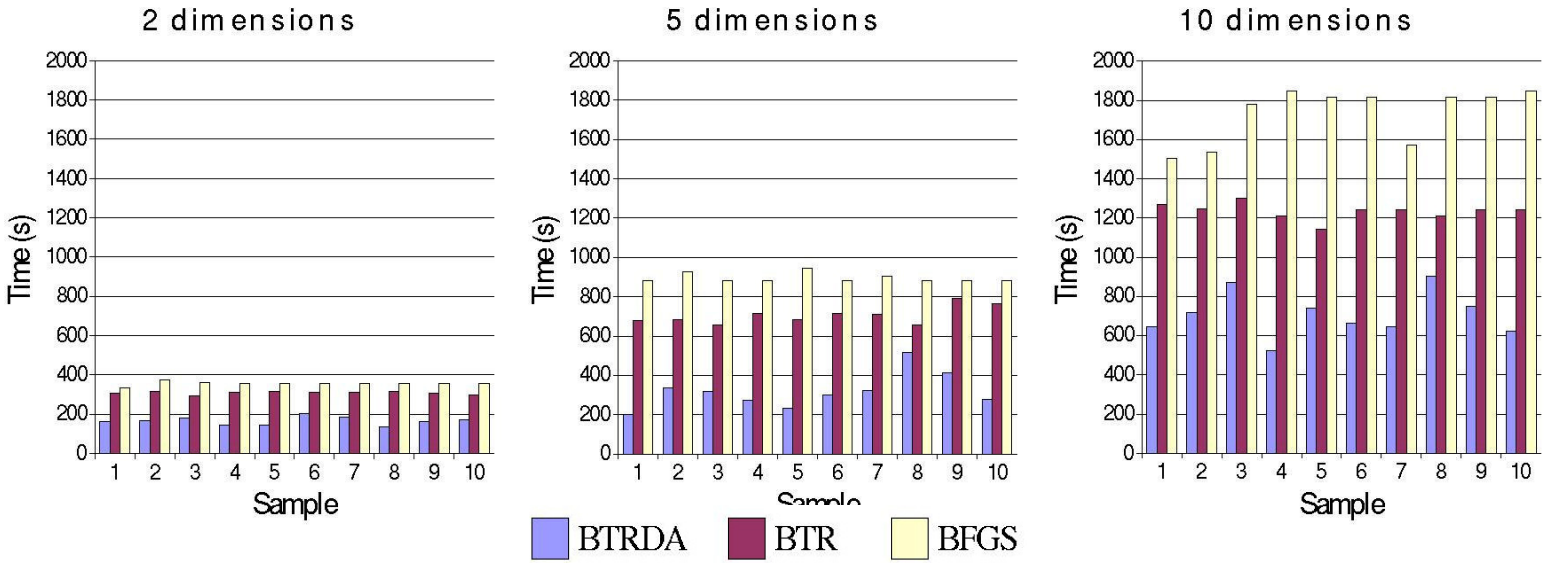
δ: signification level; $\alpha_{0.9} \approx 1.65$.
Bias of simulation (Taylor expansion):

$$\mathrm{B} := \mathrm{E}[\mathrm{SLL}^{\mathrm{R}}(\theta)] - \mathrm{LL}(\theta) = -\frac{\mathrm{I}\epsilon^2_\delta}{2\alpha^2_\delta}$$

In practice, use of SAA estimators $\sigma^{\mathrm{R}}_{\mathrm{ij_i}}(\theta)$ and $\mathrm{P}^{\mathrm{R}}_{\mathrm{ij_i}}(\theta)$.

Algorithms comparison

5000 individuals, 5 alternatives, coefficients $\sim$ N(0.5, 1)



	2 dimensions	5 dimensions	10 dimensions
BTRDA	167s	319s	709s
BTR	310s	706s	1235s
BFGS	357s	894s	1736s

Results obtained with a Pentium IV, 2 Ghz. BFGS method has been implemented using the More-Thuente linesearch.

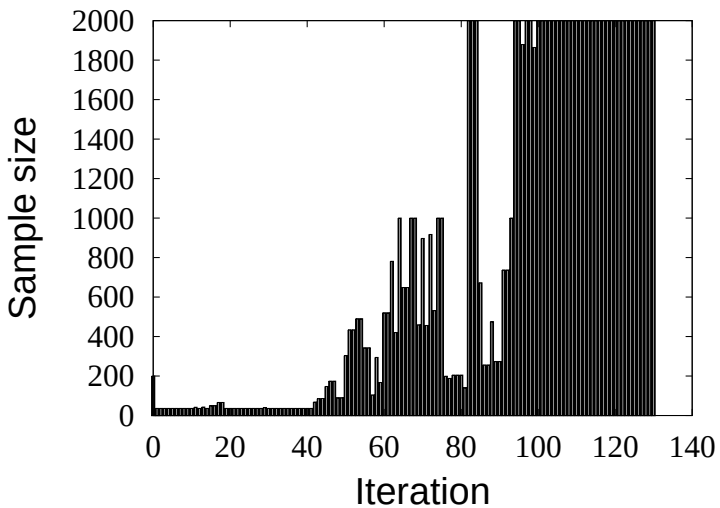
Example

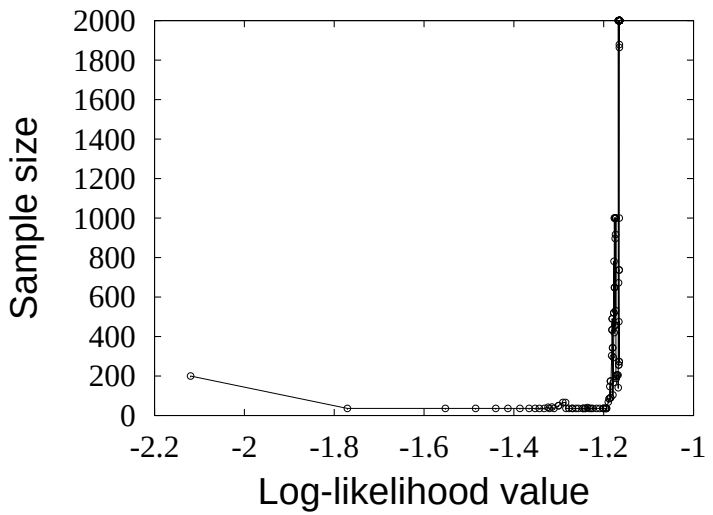
Mode choice model: Mobidrive data (Axhausen and al.)

→ 5799 observations;

→ 5 alternatives;

→ 14 parameters (3 random, normally distributed).





↓ Starting value

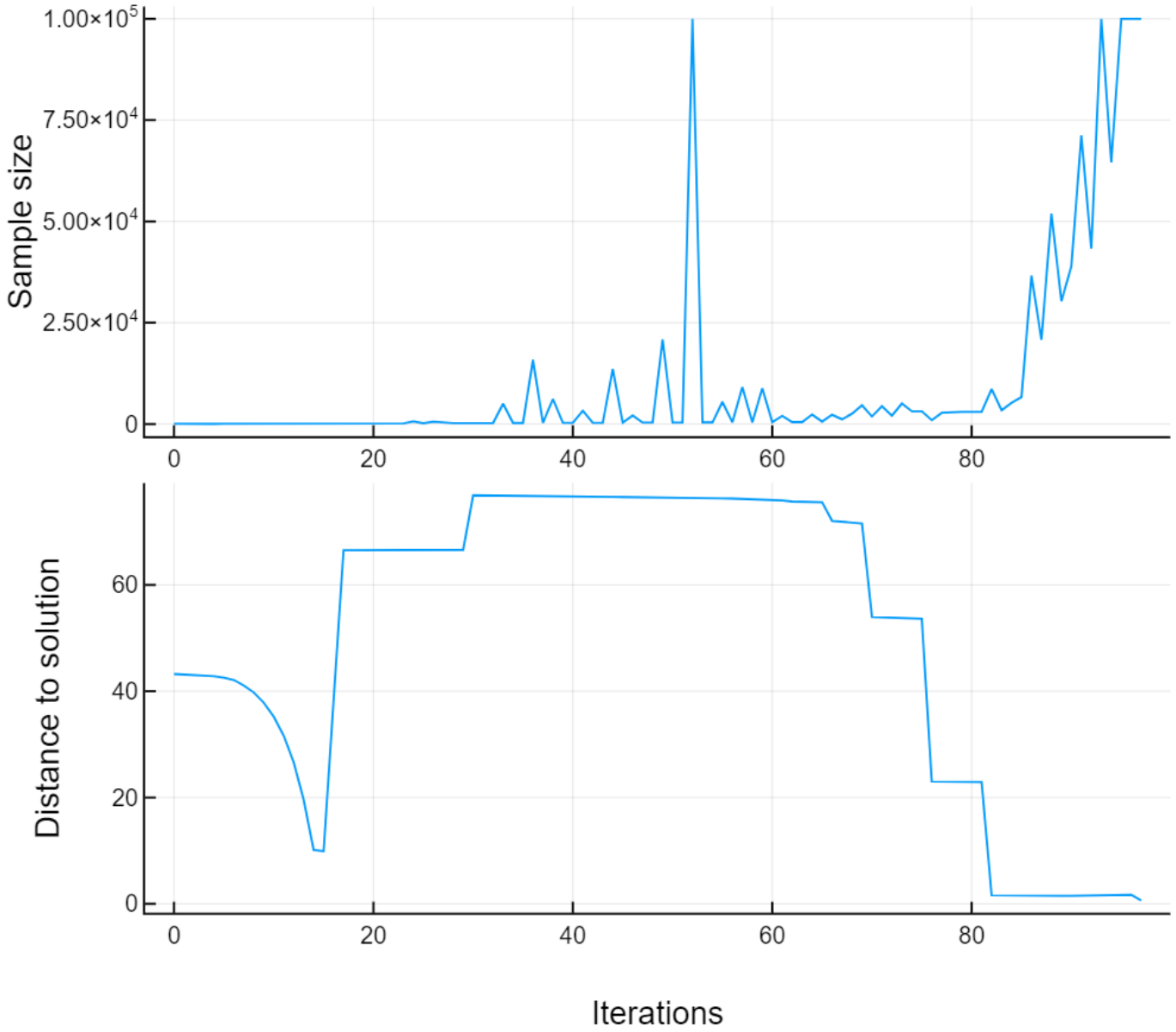
↓ Maximum value

Possible improvmet: how to draw samples?

Sampling methods are notorious for generalization issues.
Common strategies:

- Independent Random Numbers (IRN) (e.g. stochastic gradient descent)
- Incremental Common Random Numbers (CRN): take the first $\mathrm{N_K}$ numbers from a predefined sequance of random draws: $\mathcal{N}_\mathrm{k} = \{\xi_1, \ldots, \xi_{\mathrm{N_k}}\}$

Incremental Common Random Numbers



IRN/CRN combination (I/CRN)

1. $N_{k+1} > N_k$
   Draw new samples $\xi_{N_k+1,\ldots,N_{k+1}}$ and set

$$\mathcal{N}_{k+1} = \mathcal{N}_k \bigcup \left\{ \xi_{N_k+1,\ldots,\xi_{N_{k+1}}} \right\}.$$

2. $N_{k+1} < N_k$

$$\mathcal{N}_{k+1} = \mathcal{N}_k \backslash \{\text{unformly randomly selected elements in } \mathcal{N}_k\}$$

This heuristic reduces the risk to be trapped in a local minimizer or a saddle point.