A line search method with variable sample size

Nataša Krejić, Nataša Krklec

University of Novi Sad, Serbia

Scientific Computing 2011

Unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x). \tag{1}$$

$$f(x) = E(F(x, \xi))$$

- $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,
- ξ is a random vector $\xi : \Omega \to \mathbb{R}^m$
- $(\Omega, \mathcal{F}, P)$ is a probability space.

Sample average approximation

$$f(x) \approx \hat{f}_N(x) = \frac{1}{N} \sum_{i=1}^{N} F(x, \xi_i). \tag{2}$$

A random sample $\xi_1, \ldots, \xi_N$ is iid, Central Limit Theorem implies

$$\lim_{N \to \infty} \hat{f}_N(x) = f(x). \tag{3}$$

Sample path methods

$$\min f(x) \approx \hat{f}_N(x) \tag{4}$$

Variable sample size strategies

- Increasing sample sizes, $N \to \infty$ a.s convergence - Wardi, Homen de Mello, Polak, Royset etc
- Variable sample methods for TR approach - Deng & Ferris, 2008 - based on Bayesian approach, increasing sample size but N might be finite
- Variable (increasing and decreasing) sample sizes for TR approach, Bastin 2004, Bastin & Thoint 2009 - increasing and decreasing sample size, N_{max} defined

$$\min \hat{f}_{N_{max}}(x)$$

$$\min \hat{f}_{N_{max}}(x)$$

- search direction p_k
- sufficient decrease condition
- sample size, $x_k,\ \hat{f}_{N_k}, N_{k+1} \leq N_{max}$

- decrease measure
- lack of precision
- safeguard rule

lack of precision - width of confidence interval for $f(x)$

$$\varepsilon_\delta^{N_k}(x_k) \approx c$$

$$P(f(x_k) \in [\hat{f}_{N_k}(x_k) - c, \hat{f}_{N_k}(x_k) + c]) \approx \delta$$

$$c = \sigma(x_k)\alpha_\delta / \sqrt{N_k}$$

$$\hat{\sigma}^2_{N_k}(x_k) = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} (F(x_k, \xi_i) - \hat{f}_{N_k}(x_k))^2$$

$$\varepsilon_\delta^{N_k}(x_k) = \hat{\sigma}_{N_k}(x_k) \frac{\alpha_\delta}{\sqrt{N_k}} \tag{5}$$

decrease measure

$$dm_k = m_k^{N_k}(x_k) - m_k^{N_k}(x_{k+1}) = -\alpha_k p_k^T \nabla \hat{f}_{N_k}(x_k)$$

Sample size

$$N_k, N_k^{\min}, N_{\max}$$

Candidate sample size N_k^+ - based on dm_k and $\varepsilon_\delta^{N_k}(x_k)$

$$N_k \leq N_k^+ \text{ then } N_{k+1} = N_k^+$$

$$N_k > N_k^+ \text{ then } N_{k+1} = N_k^+ \text{ or } N_{k+1} = N_k$$

safeguard rule

$$\rho_k = \frac{\hat{f}_{N_k^+}(x_k) - \hat{f}_{N_k^+}(x_{k+1})}{\hat{f}_{N_k}(x_k) - \hat{f}_{N_k}(x_{k+1})}. \tag{6}$$

$$\rho_k < \eta_0 < 0 \text{ no decrease in the sample size}$$

Algorithm 1

S0 Input parameters:
$N_{max}, N_0^{min} \in \mathbb{N},\ x_0 \in \mathbb{R}^n,\ \delta, \eta, \beta, \gamma_3, \nu_1 \in (0,1),\ \eta_0 < 1.$

S1 Generate the sample realization: $\xi_1, \ldots, \xi_{N_{max}}$.
Put $k = 0$, $N_k = N_0^{min}$.

S2 Compute $\hat{f}_{N_k}(x_k)$ and $\varepsilon_\delta^{N_k}(x_k)$ using (2) and (5).

S3 Test

If $\|\nabla \hat{f}_{N_k}(x_k)\| = 0$ and $N_k = N_{max}$ then STOP.
If $\|\nabla \hat{f}_{N_k}(x_k)\| = 0$, $N_k < N_{max}$ and $\varepsilon_\delta^{N_k}(x_k) > 0$ put $N_k = N_{max}$ and $N_k^{min} = N_{max}$ and go to step S2.
If $\|\nabla \hat{f}_{N_k}(x_k)\| = 0$, $N_k < N_{max}$ and $\varepsilon_\delta^{N_k}(x_k) = 0$ put $N_k = N_k + 1$ and $N_k^{min} = N_k^{min} + 1$ and go to step S2.

S4 Determine p_k such that $p_k^T \nabla \hat{f}_{N_k}(x_k) < 0$.

S5 Using the backtracking technique with the parameter β, find α_k such that

$$\hat{f}_{N_k}(x_k + \alpha_k p_k) \leq \hat{f}_{N_k}(x_k) + \eta \alpha_k p_k^T \nabla \hat{f}_{N_k}(x_k).$$

S6 Put $s_k = \alpha_k p_k, \quad x_{k+1} = x_k + s_k$ and compute dm_k

S7 Determine the candidate sample size N_k^+ using Algorithm 2.

S8 Determine the sample size N_{k+1} using Algorithm 3.

S9 Determine the lower bound of sample size N_{k+1}^{min}.

S10 Put $k = k + 1$ and go to step S2.

Algorithm 2

S0 Input parameters: $dm_k,\ N_k^{min},\ \varepsilon_\delta^{N_k}(x_k),\ \nu_1 \in (0,1)$

S1 Determine N_k^+ by the following

1) $dm_k = \varepsilon_\delta^{N_k}(x_k) \quad \rightarrow \quad N_k^+ = N_k$
2) $dm_k > \varepsilon_\delta^{N_k}(x_k) \quad \rightarrow$ starting with $N = N_k$, while $dm_k > \varepsilon_\delta^{N}(x_k)$ and $N > N_k^{min}$, decrease N by 1 and calculate $\varepsilon_\delta^{N}(x_k) \quad \rightarrow \quad N_k^+$
3) $dm_k < \varepsilon_\delta^{N_k}(x_k)$
   1) $dm_k \geq \nu_1 \varepsilon_\delta^{N_k}(x_k) \quad \rightarrow$ starting with $N = N_k$, while $dm_k < \varepsilon_\delta^{N}(x_k)$ and $N < N_{max}$, increase N by 1 and calculate $\varepsilon_\delta^{N}(x_k) \quad \rightarrow \quad N_k^+$
   2) $dm_k < \nu_1 \varepsilon_\delta^{N_k}(x_k) \quad \rightarrow \quad N_k^+ = N_{max}$

Algorithm 3

S0 Input parameters: $N_k^+, \; N_k, \; x_k, \; x_{k+1}$

S1 Determine N_{k+1}

- **1)** If $N_k^+ \geqslant N_k$ then $N_{k+1} = N_k^+$
- **2)** If $N_k^+ < N_k$ compute ρ_k using (6).
  - **1)** If $\rho_k \geqslant \eta_0$ put $N_{k+1} = N_k^+$
  - **2)** If $\rho_k < \eta_0$ put $N_{k+1} = N_k$

The lower bound N_k^{min}

- If $N_{k+1} \leq N_k$ then $N_{k+1}^{min} = N_k^{min}$,
- else $N_{k+1} > N_k$ and
  - if N_{k+1} is a sample size which we haven't had use or if we made big enough decrease in function $\hat{f}_{N_{k+1}}$ then $N_{k+1}^{min} = N_k^{min}$.
  - If we didn't make big enough decrease in function $\hat{f}_{N_{k+1}}$ then $N_{k+1}^{min} = N_{k+1}$.

$$\hat{f}_{N_{k+1}}(x_h) - \hat{f}_{N_{k+1}}(x_{k+1}) < \gamma_3 \nu_1 (k+1-h) \varepsilon_\delta^{N_{k+1}}(x_{k+1}), \qquad (7)$$

Assumptions

A1 $\xi_1, \ldots, \xi_{N_{max}}$ are independent, identically distributed random vectors.

A2 For every $\xi, \quad F(\cdot, \xi) \in C^2(\mathbb{R}^n)$.

A3 There is a constant $M_1 > 0$ such that for every ξ and every $x \quad \|\nabla_x F(x, \xi)\| \leq M_1$.

A4 There are constants M_F, M_{FF} such that for every ξ and every $x \quad M_F \leq F(x, \xi) \leq M_{FF}$

$$\nabla_x E(F(x, \xi)) = E(\nabla_x F(x, \xi))$$

.

Convergence theory

Lemma

Suppose that assumptions A2 - A4 are true. Furthermore, suppose that there exist a positive constant κ *and number* $n_1 \in \mathbb{N}$ *such that* $\varepsilon_{\delta}^{N_k}(x_k) \geq \kappa$ *for every* $k \geq n_1$*. Then, either Algorithm 1 terminates after a finite number of iterations with* $N_k = N_{max}$ *or there exists* $q \in \mathbb{N}$ *such that for every* $k \geq q$ *the sample size is maximal, i.e.* $N_k = N_{max}$*.*

Theorem

Suppose that assumptions A2 - A5 are true. Furthermore, suppose that there exist a positive constant κ *and number* $n_1 \in \mathbb{N}$ *such that* $\varepsilon_{\delta}^{N_k}(x_k) \geq \kappa$ *for every* $k \geq n_1$ *and that the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by Algorithm 1 is bounded. Then, either Algorithm 1 terminates after a finite number of iterations at a stationary point of function* $\hat{f}_{N_{max}}$ *or every accumulation point of the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *is a stationary point of* $\hat{f}_{N_{max}}$.

Numerical results

- Noisy problems: Allufi - Pentini, Rosenbrook
- Discrete choice models - Mixed Logit problem

Directions

- $p_k = -\nabla \hat{f}_{N_k}(x_k)$
- $p_k = -H_k \nabla \hat{f}_{N_k}(x_k)$

$$f(x) = E(0.25(x_1\xi)^4 - 0.5(x_1\xi)^2 + 0.1\xi x_1 + 0.5x_2^2),$$

$$\xi : \mathcal{N}(1, \sigma^2). \tag{8}$$

σ^2	global minimizer - x^*	local minimizer	maximizer	$f(x^*)$
0.01	$(-1.02217, 0)$	$(0.922107, 0)$	$(0.100062, 0)$	-0.340482
0.1	$(-0.863645, 0)$	$(0.771579, 0)$	$(0.092065, 0)$	-0.269891
1	$(-0.470382, 0)$	$(0.419732, 0)$	$(0.05065, 0)$	-0.145908

Table 1: Stationary points for Allufi - Pentini's problem

$\sigma^2 = 0.01,\ N_{max} = 100$				
Algorithm	$\|\nabla \hat{f}_{N_{max}}\|$	$\|\nabla f\|$	*fev*	*fevNmax*
NG	0.008076	0.014906	1402	1868
NG - ρ	0.008002	0.013423	1286	
BFGS	0.003575	0.011724	840	928
BFGS - ρ	0.003556	0.012158	793	

$\sigma^2 = 0.1\ ,\ N_{max} = 200$				
Algorithm	$\|\nabla \hat{f}_{N_{max}}\|$	$\|\nabla f\|$	*fev*	*fevNmax*
NG	0.007545	0.027929	3971	4700
NG - ρ	0.006952	0.028941	3537	
BFGS	0.003414	0.027991	2155	2968
BFGS - ρ	0.003879	0.027785	2152	

$\sigma^2 = 1\ ,\ N_{max} = 600$				
Algorithm	$\|\nabla \hat{f}_{N_{max}}\|$	$\|\nabla f\|$	*fev*	*fevNmax*
NG	0.006072	0.050208	13731	15444
NG - ρ	0.005149	0.058036	10949	
BFGS	0.003712	0.054871	7829	14760
BFGS - ρ	0.002881	0.055523	8372	

$$f(x) = E(100(x_2 - (x_1\xi)^2)^2 + (x_1\xi - 1)^2), \tag{9}$$

σ^2	global minimizer - x^*	$f(x^*)$
0.001	$(0.711273, 0.506415)$	0.186298
0.01	$(0.416199, 0.174953)$	0.463179
0.1	$(0.209267, 0.048172)$	0.634960

Table 4: Rosenbrock problem - the global minimizers

$\sigma^2 = 0.001$, $N_{max} = 3500$				
Algorithm	$\|\nabla \hat{f}_{N_{max}}\|$	$\|\nabla f\|$	*fev*	*fevNmax*
BFGS	0.003413	0.137890	56857	246260
BFGS - ρ	0.003068	0.137810	49734	
$\sigma^2 = 0.01$, $N_{max} = 3500$				
Algorithm	$\|\nabla \hat{f}_{N_{max}}\|$	$\|\nabla f\|$	*fev*	*fevNmax*
BFGS	0.002892	0.114680	56189	213220
BFGS - ρ	0.003542	0.114160	52875	
$\sigma^2 = 0.1$, $N_{max} = 3500$				
Algorithm	$\|\nabla \hat{f}_{N_{max}}\|$	$\|\nabla f\|$	*fev*	*fevNmax*
BFGS	0.003767	0.093363	67442	159460
BFGS - ρ	0.003561	0.093290	59276	

Table 5: Rosenbrock problem

Safeguard rule

- Decreasing sample size proposed in 11% - 32% iterations
- Rejection of the decrease by the safeguard rule: 25% - 66%

Mixed Logit Models

r_a agents, r_m alternatives, r_k characteristics
Utility of agent i for alternative j

$$U_{i,j} = V_{i,j} + \varepsilon_{i,j},$$

$$V_{i,j} = V_{i,j}(\beta^i) = m_j^T \beta^i.$$

$$\beta^i = (\beta_1^i, ..., \beta_{r_k}^i)^T = (\mu_1 + \xi_1^i \sigma_1, ..., \mu_{r_k} + \xi_{r_k}^i \sigma_{r_k})^T,$$

Task: Estimate μ_k, σ_k

$$L_{i,j}(x,\bar{\xi}^i) = \frac{e^{V_{i,j}(x,\bar{\xi}^i)}}{\sum_{s=1}^{r_m} e^{V_{i,s}(x,\bar{\xi}^i)}}.$$

$$\min f(x) := -\frac{1}{r_a}\sum_{i=1}^{r_a} \ln E(L_{i,j(i)}(x,\xi^i)),$$

$$\hat{f}_N(x) = -\frac{1}{r_a}\sum_{i=1}^{r_a} \ln(\frac{1}{N}\sum_{s=1}^{N} L_{i,j(i)}(x,\xi_s^i)).$$

$$\varepsilon_\delta^N(x) = \frac{\alpha_\delta}{r_a}\sqrt{\sum_{i=1}^{r_a} \frac{\hat{\sigma}^2_{N,i,j(i)}(x)}{NP^2_{i,j(i)}(x)}}. \tag{10}$$

$$\hat{\sigma}^2_{N,i,j(i)}(x) = \frac{1}{N-1}\sum_{s=1}^{N}(L_{i,j(i)}(x,\xi_s^i) - \frac{1}{N}\sum_{k=1}^{N}(L_{i,j(i)}(x,\xi_k^i))^2.$$

Algorithm	$\|\nabla \hat{f}_{N_{max}}\|$	$\tilde{g}$	*fev*	*fevNmax*
NG	0.008888	0.008101	4.4668E+07	9.5300E+07
NG - ρ	0.009237	0.008530	3.8611E+07	
BFGS	0.004128	0.003498	6.2430E+06	1.7750E+07
BFGS - ρ	0.004616	0.004256	5.7895E+06	

Table 7 : Mixed Logit Problem