A stochastic inexact restoration trust-region method with application to machine learning

Simone Rebegoldi[†]

(Joint work with S. Bellavia, N. Krejić and B. Morini)

[†]Department of Industrial Engineering, University of Florence

BOS/SOR Conference 2020

December 15 2020

Problem setting

We consider the following finite-sum minimization problem:

$$\min_{x \in \mathbb{R}^n} f_N(x) = \frac{1}{N} \sum_{i=1}^{N} \phi_i(x), \tag{1}$$

where $\phi_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, N$, are continuously differentiable.

- Several problems in machine learning can be cast in the form (1): binary or multinomial classification, data fitting, sample average approximation ...
- The loss function f_N is often nonconvex, e.g. in the case of neural networks
- Big data applications $\Rightarrow$ N very large $\Rightarrow$ $f_N(x)$ and $\nabla f_N(x)$ very expensive!

Stochastic gradient descent (SGD)

Given $x_0 \in \mathbb{R}^n$, compute

$$x_{k+1} = x_k - \alpha_k g_k, \quad k = 0, 1, \ldots$$

where $\alpha_k > 0$ is the learning rate and g_k is the stochastic gradient, defined by

$$g_k = \nabla f_{N_k}(x_k) = \frac{1}{N_k} \sum_{i \in I_{N_k}} \nabla \phi_i(x_k),$$

where $I_{N_k} \subset \{1, \ldots, N\}$ and $|I_{N_k}| = N_k$.

- If $N_k = 1 \Rightarrow$ standard SGD
- If $N_k > 1 \Rightarrow$ mini-batch SGD

Theorem (Bottou et al., SIAM Rev., 2018)

Suppose ∇f_N is $L-$Lipschitz continuous. Let Ω be an open set s.t. $\{x_k\} \subset \Omega$ and $f_N(x) \geq f_{low}$ for $x \in \Omega$. Assume there exist $\mu, M_1, M_2 > 0$ s.t.

$$\nabla f_N(x_k)^T \mathbb{E}(g_k) \geq \mu \|\nabla f_N(x_k)\|^2, \quad E(\|g_k\|^2) \leq M_1 + M_2 \|\nabla f_N(x_k)\|^2.$$

If $\alpha_k \equiv \alpha$, with $0 < \alpha \leq \mu/(LM_2)$, then

$$\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K} \|\nabla f_N(x_k)\|^2\right) \leq \frac{L\alpha M_1}{\mu} + \frac{2(f_N(x_1) - f_{low})}{\mu\alpha K}.$$

- ✓ The average norm of the gradients can be made arbitrarily small by picking a small α ...
- ✗ ... but the smaller α, the slower the convergence rate!
- ✗ Sublinear convergence for strongly convex f_N holds only if α_k is diminishing:

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Globalization strategies

SGD and its variants employ stochastic (possibly and occasionally full) gradient estimates and do not rely on any machinery from standard globally convergent optimization procedures, such as linesearch or trust-region.

On the other hand, a few recent papers rely on such strategies for selecting the steplength[1,2,3]. Part of them mimic traditional step acceptance rules using stochastic estimates of functions and gradients, which are required to be sufficiently accurate in probability.

The purpose of these methods is to partially overcome the dependence of the steplengths from the Lipschitz constant of the gradient.

[1] S. Bellavia, N. Krejić, B. Morini, Comput. Optim. Appl. 76, 701–736, 2020
[2] R. Chen, M. Menickelly, K. Scheinberg, Math. Progr. 169(2), 447–487, 2018
[3] C. Paquette, K. Scheinberg, arXiv:1807.07994, 2018

Example: STORM (STochastic Optimization with Random Models)[4,5]

0. Choose $x_0 \in \mathbb{R}^n$, $0 < \Delta_0 < \Delta_{\max}$, $\gamma > 1$, $\alpha, \beta, \eta_1 \in (0,1)$, $\eta_2 > 0$. Set $k = 0$.
1. Build a model $m_k(y)$ which is $\alpha-$probabilistically $\kappa-$fully linear, i.e.

   $$Pr\{|f_N(y) - m_k(y)| \leq \kappa\Delta_k^2, \quad \|\nabla f_N(y) - \nabla m_k(y)\| \leq \kappa\Delta_k, \ \forall\, y \in B(x_k, \Delta_k)\} \geq \alpha.$$

2. Compute $s_k = \operatorname{argmin}_{\|s\| \leq \Delta_k} m_k(s)$ approximately.
3. Compute f_k^0 and f_k^s which are $\beta-$probabilistically ϵ_F-accurate, i.e.

   $$Pr\{|f_k^0 - f_N(x_k)| \leq \epsilon_F\Delta_k^2, \quad |f_k^s - f_N(x_k + s_k)| \leq \epsilon_F\Delta_k^2\} \geq \beta.$$

4. Compute $\rho_k = \frac{f_k^0 - f_k^s}{m_k(x_k) - m_k(x_k + s_k)}$.

   If $\rho_k \geq \eta_1$ and $\|g_k\| \geq \eta_2\Delta_k$ set $x_{k+1} = x_k + s_k$, $\Delta_{k+1} = \min\{\gamma\Delta_k, \Delta_{\max}\}$
   else $x_{k+1} = x_k$, $\Delta_{k+1} = \gamma^{-1}\Delta_k$, and go to Step 1.

✓ If $m_k(s) = f_k + g_k^T s$, STORM is an SGD method with adaptive steplength

✓ Probabilistic accuracy of m_k, f_k^0, f_k^s guarantees convergence in probability

✗ Function and gradient need to be estimated with increasingly high precision!

[4] S. Bandeira, K. Scheinberg, L.N. Vicente, SIAM J. Optim. 24(3), 1238–1264, 2014
[5] R. Chen, M. Menickelly, K. Scheinberg, Math. Progr. 169(2), 447–487, 2018

The proposed approach

We propose a stochastic first-order trust-region method with the following features.

- The trust-region model and acceptance rule employ both function and gradient estimates (similarly to STORM).
- The function sample size is computed dynamically according to a deterministic rule inspired by the Inexact Restoration (IR) Method[6] (unlike STORM).
- We require probabilistic accuracy for the gradient estimates only when the full function sample size is reached (unlike STORM).

GOAL: delay the use of the full function sample size and the adoption of probabilistically accurate random models as much as possible.

[6] J.M. Martinez, E.A. Pilotta, J. Optim. Theory Appl. 104, 135–163, 2000

The Inexact Restoration (IR) method is a constrained optimization tool suitable for problems where one does not want to enforce feasibility in all iterations.
The key idea is to improve feasibility and optimality in separate procedures. Each iteration ensures the sufficient decrease of a suitable merit function and, under certain assumptions, convergence to a feasible optimal point.

IDEA: apply the IR strategy to dynamically select the function sample size.
To this aim, let us rewrite the finite-sum minimization problem as

$$\min_{x \in \mathbb{R}^n} f_M(x) = \frac{1}{M} \sum_{i \in I_M} \phi_i(x)$$

$$\text{s.t. } M = N$$

where $I_M \subset \{1, \ldots, N\}$, $|I_M| = M$.

Ingredient #1: the Inexact Restoration method

We measure the level of infeasibility with respect to the constraint $M = N$ by using

$$h : \mathbb{N} \to \mathbb{R} \text{ decreasing function such that } h(1) > 0 \text{ and } h(N) = 0.$$

We also introduce the merit function

$$\Psi(x, M, \theta) = \theta f_M(x) + (1 - \theta)h(M), \quad \theta \in (0, 1).$$

Then, the IR method involves the following steps:

- compute $\tilde{N}_{k+1}$ such that $h(\tilde{N}_{k+1}) \leq r h(N_k)$, $r \in (0, 1)$ (restoration phase);
- compute the function sample size $N_{k+1} \leq \tilde{N}_{k+1}$;
- based on the inexact model $m_k(p)$ for the function $f_{N_{k+1}}$ around x_k, compute the trial point $x_k + p_k$;
- compute θ_{k+1}, consider the predicted and actual reduction defined as

$$\mathrm{Pred}_k(\theta_{k+1}) = \theta_{k+1}(f_{N_k}(x_k) - m_k(p_k)) + (1 - \theta_{k+1})(h(N_k) - h(\tilde{N}_{k+1}))$$
$$\mathrm{Ared}(x_k + p_k, \theta_{k+1}) = \Psi(x_k, N_k, \theta_{k+1}) - \Psi(x_{k+1}, N_{k+1}, \theta_{k+1})$$

and accept the trial point only if

$$\mathrm{Ared}(x_k + p_k, \theta_{k+1}) \geq \eta \, \mathrm{Pred}_k(\theta_{k+1}), \quad \eta \in (0, 1).$$

Ingredient #2: the trust-region approach

We consider a linear model $m_k(p)$ of $f_{N_{k+1}}$ around x_k of the form

$$m_k(p) = f_{N_{k+1}}(x_k) + g_k^T p,$$

and minimize it over the ball $B(0, \Delta_k)$, obtaining

$$p_k = \underset{\|p\| \leq \Delta_k}{\operatorname{argmin}}\ m_k(p).$$

- The trial point has the form

$$x_k + p_k = x_k - \frac{\Delta_k}{\|g_k\|} g_k,$$

  which is an SGD step with adaptive steplength!
- The stochastic gradient g_k is not necessarily computed using the same sample size as $f_{N_{k+1}}$. For instance, we allow for subsampling over $I_{N_{k+1}}$, i.e.

$$g_k = \frac{1}{N_{k+1,g}} \sum_{i \in I_{N_{k+1,g}}} \nabla \phi_i(x_k),$$

  where $I_{N_{k+1,g}} \subset I_{N_{k+1}}$ and $|I_{N_{k+1,g}}| = N_{k+1,g} \leq N_{k+1}$.

SIRTR - Stochastic Inexact Restoration Trust-Region algorithm

Choose $x_0 \in \mathbb{R}^n$, N_0 in $(0, N]$, $\theta_0, r, \eta \in (0, 1)$, $0 < \Delta_0 < \Delta_{\max}$, $\gamma > 1$, $\mu, \eta_2 > 0$.

STEP 0. Set $k = 0$.

STEP 1. If $N_k < N$, set $I_k = 0$ and find $\tilde{N}_{k+1}$ such that $N_k < \tilde{N}_{k+1} \leq N$ and

$$h(\tilde{N}_{k+1}) \leq r h(N_k).$$

Else set $I_k = 1$ and $\tilde{N}_{k+1} = N$.

STEP 2. Find N_{k+1} such that $N_{k+1} \leq \tilde{N}_{k+1}$ and

$$h(N_{k+1}) - h(\tilde{N}_{k+1}) \leq \mu \Delta_k^2.$$

STEP 3. Choose the stochastic gradient $g_k \in \mathbb{R}^n$ and set $p_k = -\Delta_k \frac{g_k}{\|g_k\|}$.

STEP 4. If $N_k = N$, $N_{k+1} < N$ and

$$f_N(x_k) - m_k(p_k) < \Delta_k \|g_k\|,$$

take $\Delta_k = \Delta_k / \gamma$ and go to STEP 2.

SIRTR - Stochastic Inexact Restoration Trust-Region algorithm

Choose $x_0 \in \mathbb{R}^n$, N_0 in $(0, N]$, $\theta_0, r, \eta \in (0,1)$, $0 < \Delta_0 < \Delta_{\max}$, $\gamma > 1$, $\mu, \eta_2 > 0$.

STEP 5. Compute the penalty parameter

$$\theta_{k+1} = \begin{cases} \theta_k, & \text{if } \mathrm{Pred}_k(\theta_k) \geq \eta(h(N_k) - h(\tilde{N}_{k+1})) \\ \dfrac{(1-\eta)(h(N_k)-h(\tilde{N}_{k+1}))}{m_k(p_k)-f_{N_k}(x_k)+h(N_k)-h(\tilde{N}_{k+1})}, & \text{otherwise.} \end{cases}$$

STEP 6. If $\mathrm{Ared}(x_k + p_k, \theta_{k+1}) \geq \eta\, \mathrm{Pred}_k(\theta_{k+1})$ and $(\|g_k\| - \eta_2 \Delta_k) I_k \geq 0$, set

$$x_{k+1} = x_k + p_k$$

$$\Delta_{k+1} = \begin{cases} \Delta_k, & \text{if } N_k < N, \\ \min\{\gamma \Delta_k, \Delta_{\max}\}, & \text{otherwise,} \end{cases}$$

set $k = k + 1$ and go to STEP 1.
Else set $x_{k+1} = x_k$, $\Delta_{k+1} = \Delta_k/\gamma$, $k = k + 1$, $\tilde{N}_{k+1} = \tilde{N}_k$, $I_{k+1} = I_k$ and go to STEP 2.

Assumptions

1. There exist $\Omega \subset \mathbb{R}^n$, f_{low}, f_{up} such that $\{x_k\} \subset \Omega$ and

$$f_{low} < f_M(x) < f_{up}, \quad \forall\, 1 \leq M \leq N,\ x \in \Omega.$$

2. The gradients $\nabla \phi_i$, $1 \leq i \leq N$, are Lipschitz continuous on Ω.
3. There exists $\Gamma > 0$ such that

$$\left\| g_k - \nabla f_{N_{k+1}}(x_k) \right\| \leq \Gamma, \quad \forall\, k \in \mathbb{N}.$$

Under these mild assumptions, we can prove some basic properties of SIRTR:

- if $N_k < N$, there exists $\Delta > 0$ such that iteration k is successful for $\Delta_k \leq \Delta$
- $\Delta_k \to 0$ as $k \to \infty$
- $N_k = N$ for all k sufficiently large.

The convergence analysis relies on the Lyapunov type function defined as

$$\Phi(x, N, \theta, \Delta) = v(\Psi(x, N, \theta) + \Sigma\theta_k) + (1 - v)\Delta^2,$$

where $v \in (0, 1)$ and Σ is such that $f_{N_k}(x) - h(N_k) + \Sigma \geq 0$ for $x \in \Omega$.

Our aim is to guarantee the sufficient decrease of Φ along successive iterations. Setting $\Phi_k = \Phi(x_k, N_k, \theta_k, \Delta_k)$, since $\{\theta_k\}$ is decreasing, we can show that

$$\Phi_{k+1} - \Phi_k \leq -v\,\mathrm{Ared}(x_{k+1}, \theta_{k+1}) + (1 - v)(\Delta_{k+1}^2 - \Delta_k^2).$$

Therefore, the possible decrease of $\{\Phi_k\}$ depends on both the actual reduction and the update rule of the trust-region radius.

Let us distinguish the iteration indexes k as below:

$$\begin{aligned}
\mathcal{I}_1 &= \{k \geq 0 \text{ s.t. } h(N_k) - h(\tilde{N}_{k+1}) > 0\},\\
\mathcal{I}_2 &= \{k \geq 0 \text{ s.t. } h(N_k) = h(\tilde{N}_{k+1}) = 0, N_{k+1} = N\},\\
\mathcal{I}_3 &= \{k \geq 0 \text{ s.t. } h(N_k) = h(\tilde{N}_{k+1}) = 0,\ N_{k+1} < N\}.
\end{aligned}$$

Lemma

Let Assumptions 1–3 hold, $k_\phi = \max\{|f_{low}|, |f_{up}|\}$, $\underline{h} = h(N-1)$, $\underline{\theta} = \inf_k \theta_k$.

1. If k is unsuccessful, we have

$$\Phi_{k+1} - \Phi_k \leq \left(v \left(\frac{2\kappa_\phi}{\Delta^2} + \mu \right) + (1-v) \frac{1-\gamma^2}{\gamma^2} \right) \Delta_k^2.$$

2. If k is successful and $k \in \mathcal{I}_1$ we have

$$\Phi_{k+1} - \Phi_k \leq -v \left(\frac{\eta^2 (1-r) \underline{h}}{\Delta_{\max}^2} \right) \Delta_k^2;$$

if k is successful and $k \in \mathcal{I}_2 \cup \mathcal{I}_3$, we have

$$\Phi_{k+1} - \Phi_k \leq \left(-v\eta\eta_2\underline{\theta} + (1-v)(\gamma^2 - 1) \right) \Delta_k^2.$$

A suitable choice of v and η_2 guarantees the sufficient decrease for all iterations.

Convergence properties

Theorem 1 (S. Bellavia, N. Krejić, B. Morini, S. Rebegoldi, 2020)

Let Assumptions 1–3 hold. If the following condition holds

$$\eta_2 > \frac{\gamma^2}{\eta\underline{\theta}}\left(\frac{2\kappa_\phi}{\Delta^2} + \mu\right), \tag{2}$$

then there exist $v \in (0,1)$ and $\sigma > 0$ such that

$$\Phi_{k+1} - \Phi_k \leq -\sigma\Delta_k^2, \quad \forall\, k \geq 0.$$

Consequently, we have

$$\sum_{k=0}^{\infty} \Delta_k^2 < \infty,$$

and $N_{k+1} = \tilde{N}_{k+1} = N$ for k sufficiently large.

- ✓ No probabilistic accuracy required
- ✗ The parameter η_2 may need to be large to satisfy (2) ...
- ✓ ... However, condition $\|g_k\| \geq \eta_2\Delta_k$ goes into action only when $N_k = N$.

Convergence properties

Theorem 2 (S. Bandeira et al., SIAM J. Optim., 2014)

Let Assumptions 1–3 and condition (2) hold.
Denote with $\bar{k}$ the first index such that $N_{k+1} = \tilde{N}_{k+1} = N$ for all $k \geq \bar{k}$.
If model $m_k(p)$ is $\alpha-$probabilistically $\kappa-$fully linear with $\alpha \geq 1/2$ for $k \geq \bar{k}$, then

$$Pr\left\{\lim_{k\to\infty} \|\nabla f_N(x_k)\| = 0\right\} = 1.$$

- ✓ Under the assumptions of Theorem 2 and for k sufficiently large, SIRTR reduces to STORM with exact function evaluations and random gradients
- ✓ Convergence in probability
- ✗ In order to impose the probabilistic accuracy, one should take $N_{k+1,g} \geq \mathcal{O}(1/\Delta_k^2)$ $\Rightarrow$ very large $N_{k+1,g}$ when Δ_k is small!

Preliminary numerical results

We tested our method on a nonconvex problem arising in binary classification. Let $\{(a_i, b_i)\}_{i=1}^N$ denote the pairs forming the training set, being $a_i \in \mathbb{R}^n$ the vector containing the entries of the i-th example and $b_i \in \{0, 1\}$ its label. Then the classification problem is solved by minimizing

$$f_N(x) = \frac{1}{N} \sum_{i=1}^{N} \left(b_i - \frac{1}{1 + e^{-a_i^T x}} \right)^2 .$$

	Training set		Testing set
Data set	N	n	N_T
A9A	22793	123	9768
IJCNN1	49990	22	91701
MNIST	60000	784	10000
HTRU2	10000	8	7898

Table: Datasets used. For each data set, N is the number of training examples, n the dimension of each instance, and N_T the number of elements in the testing set.

Preliminary numerical results

- Infeasibility measure

$$h(M) = \frac{N - M}{N}, \quad 1 \leq M \leq N.$$

- Function sample size

$$\tilde{N}_{k+1} = \min\{N, \lceil \tilde{c} \cdot N_k \rceil\}, \quad \tilde{c} > 1$$

$$N_{k+1} = \begin{cases} \lceil \tilde{N}_{k+1} - \mu N \Delta_k^2 \rceil, & \text{if } \lceil \tilde{N}_{k+1} - \mu N \Delta_k^2 \rceil \in [N_0, 0.95N] \\ \tilde{N}_{k+1}, & \text{if } \lceil \tilde{N}_{k+1} - \mu N \Delta_k^2 \rceil < N_0 \\ N, & \text{if } \lceil \tilde{N}_{k+1} - \mu N \Delta_k^2 \rceil > 0.95N. \end{cases}$$

- Stochastic gradient

$$g_k = \nabla f_{N_{k+1,g}}(x_k), \quad \text{where } N_{k+1,g} = \lceil c \cdot N_{k+1} \rceil, \quad c \in (0, 1].$$

- Stopping criterion

$$\|\nabla f_{N_{k,g}}(x_k)\| \leq \epsilon, \qquad |f_{N_k}(x_k) - f_{N_{k-1}}(x_{k-1})| \leq \epsilon |f_{N_{k-1}}(x_{k-1})| + \epsilon.$$

Preliminary numerical results

$\tilde{N}_{k+1}$	$\min\{N, \lceil 1.05N_k \rceil\}$			$\min\{N, \lceil 1.1N_k \rceil\}$			$\min\{N, \lceil 1.2N_k \rceil\}$		
	`cost`	`err`	`sub`	`cost`	`err`	`sub`	`cost`	`err`	`sub`
A9A	18	0.169	46	29	0.165	32	55	0.164	10
IJCNN1	16	0.092	48	26	0.089	31	27	0.087	22
MNIST	20	0.152	46	40	0.144	37	81	0.141	18
HTRU2	31	0.022	40	38	0.024	25	58	0.024	13

Table: Average results obtained running SIRTR 50 times with $N_0 = \lceil 0.1N \rceil$.
`cost` is the overall number of full function/gradient evaluations.
`err` is the classification error obtained with the final iterate.
`sub` is the number of times the problem is solved without reaching the full sample size.

Preliminary numerical results

N_0	$\lceil 0.001N \rceil$			$\lceil 0.01N \rceil$			$\lceil 0.1N \rceil$		
	cost	err	sub	cost	err	sub	cost	err	sub
A9A	18	0.185	48	14	0.179	49	18	0.169	46
IJCNN1	13	0.096	49	14	0.081	49	16	0.092	48
MNIST	3	0.267	50	11	0.170	49	20	0.152	46
HTRU2	4	0.045	49	25	0.025	42	31	0.022	40

Table: Average results obtained running SIRTR 50 times with $\tilde{N}_{k+1} = \min\{N, \lceil 1.05 N_k \rceil\}$.
`cost` is the overall number of full function/gradient evaluations.
`err` is the classification error obtained with the final iterate.
`sub` is the number of times the problem is solved without reaching the full sample size.

Preliminary numerical results

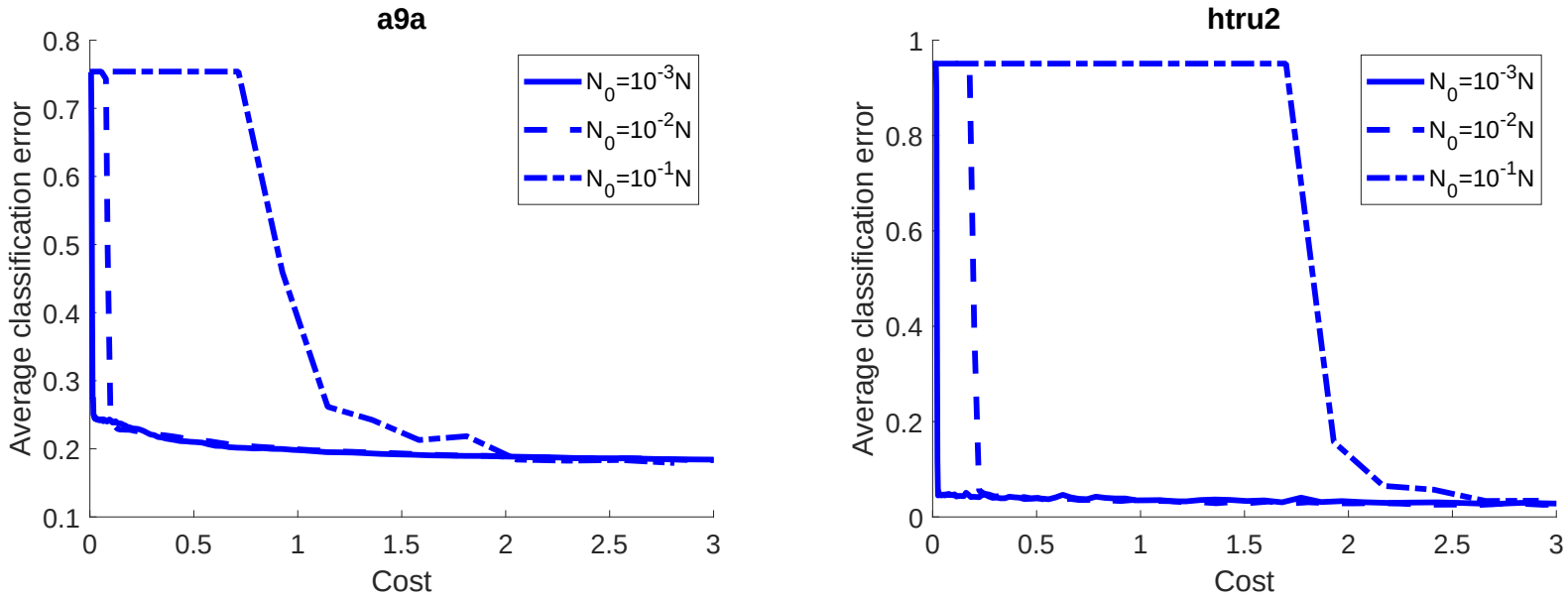

Figure: Decrease of the average classification error VS average computational cost.

Conclusions and future work

We have proposed a first-order trust-region method with function and gradient estimates built via subsampling techniques. The choice of the function sample size is deterministic and ruled by the inexact restoration approach.

The proposed method eventually reaches full precision in evaluating the objective function. However, numerical tests show that the method is stable with respect to the parameters, which makes easy to delay the use of the full sample size.

Future work will concern the numerical assessment of the algorithm in comparison to other stochastic trust-region algorithms (STORM or TRish) and its extensive application to neural networks.

References

- S. Bellavia, N. Krejić, B. Morini, Inexact restoration with subsampled trust-region methods for finite-sum minimization, *Comput. Optim. Appl. 76*, 701–736, 2020
- S. Bellavia, N. Krejić, B. Morini, S. Rebegoldi, A stochastic inexact restoration trust-region algorithm with subsampling, *in preparation*, 2020