Adaptive Optimization Methods Using Random Models and Examples from Machine Learning

Stefania Bellavia
Dipartimento di Ingegneria Industriale
Università di Firenze

UNIVERSITÀ DEGLI STUDI FIRENZE

CALCOLO SCIENTIFICO E MODELLI MATEMATICI:
alla ricerca delle cose nascoste attraverso le cose manifeste

Rome, April 6th-8th 2022

Acknowledgements

- Gianmarco Gurioli, Benedetta Morini, Simone Rebegoldi
  University of Florence, Italy
- Nataša Krejić
  University of Novi Sad, Serbia
- Philippe Toint
  University of Namur, Belgium

Outline

- Introduction: random models and motivating applications.
- Trust-region procedures with random models: adaptive choice of sample size and learning rate.
- Complexity results in expectation.
- Finite sum: trust region & Inexact restoration.
- Conclusions.

Optimization with random models

Unconstrained Optimization Problems

$$\min_{x \in \mathbb{R}^n} f(x),$$

with $f : \mathbb{R}^n \to \mathbb{R}$ sufficiently smooth ($f \in C^2$ for second-order methods), bounded below, possibly nonconvex.

- $f(x)$, $\nabla f(x)$ and $\nabla^2 f(x)$ evaluations are subject to random noise and we can only compute random estimates

$$\overline{f}(x) = \overline{f}(x, \xi), \qquad \overline{\nabla f}(x) = \overline{\nabla f}(x, \xi) \qquad \overline{\nabla^2 f}(x) = \overline{\nabla^2 f}(x, \xi)$$

where ξ is a random variable.

Random model and optimality measures

First-order

- random model: $m_k(p) = f(x) + \overline{\nabla f}(x)^T p$
- ϵ- approximate first-order critical point:

$$\|\nabla f(\hat{x})\|_2 \le \epsilon.$$

Second-order

- random model: $m_k(p) = f(x) + \overline{\nabla f}(x)^T p + \frac{1}{2} p^T \overline{\nabla^2 f}(x) p$

ϵ approximate first and second-order critical point:

$$\begin{cases} \|\nabla f(\hat{x})\|_2 \le \epsilon \\ \lambda_{\min}(\nabla^2 f(\hat{x})) \ge -\epsilon. \end{cases}$$

.

Motivating applications

Finite-sum minimization problem:

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

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

- Several problems can be cast in the previous form: classification, data fitting, sample average approximation ...
- Supervised machine learning: given a family of prediction function $h(\cdot; x)$, $x \in \mathbb{R}^n$, a loss function ℓ and a set of examples $\{(a_i, b_i)\}_{i=1}^N$ (training set), $a_i \in \mathbb{R}^d$ (feature), $b_i \in \mathbb{R}$ (label),

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\ell(h(a_i; x), b_i)}_{\phi_i(x)}$$

- The function f is often nonconvex, e.g. in the case of neural networks
- Big data applications $\Rightarrow$ N very large $\Rightarrow$ f and derivatives are very expensive!

Subsampled functions, gradients and Hessians

N is large

- M: sample size
- I_M: a randomly selected nonempty subset of $\{1, \ldots, N\}$ of cardinality M

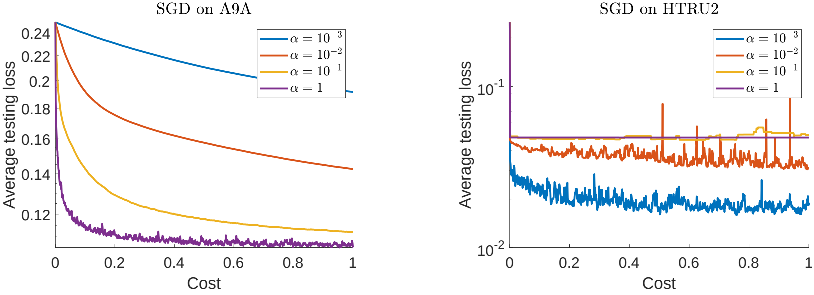
$$I_M \subseteq \{1, \ldots, N\}, \quad |I_M| = M, \quad M \geq 1,$$

Use:

$$\begin{aligned} \overline{f}(x) &= \frac{1}{M} \sum_{i \in I_M} \phi_i(x) \\ \overline{\nabla f}(x) &= \frac{1}{M} \sum_{i \in I_M} \nabla \phi_i(x) \\ \overline{\nabla^2 f}(x) &= \frac{1}{M} \sum_{i \in I_M} \nabla^2 \phi_i(x) \end{aligned}$$

- A training set shows redundancy in the data $\Rightarrow$ using all the sample data in every optimization iteration is inefficient
- Overall less expensive when N is large
- Computational evidence that they are more robust than fully deterministic approaches.

Stochastic gradient methods



$$x_{k+1} = x_k - \alpha_k \overline{\nabla f}(x_k), \qquad k = 0, 1, \ldots$$

- ✓ The expected value of the average norm of the gradients can be made small by picking a sufficiently small α
- ✗ … but the smaller α, the slower the convergence rate!
- ✗ The optimal α (and the mini-batch size) are problem-dependent!
- ✗ For large-scale, real-world systems, expensive parameter tuning efforts is required!

Bottou, Curtis and Nocedal, SIREV 2018, Curtis, Scheinberg, ArXiv, 2020.

Adaptive stochastic optimization methods

- 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.
- Strategies for selecting the steplength that mimic traditional step acceptance rules using stochastic estimates of functions and gradients:
  - Some criterion to accept/reject the step is tested
  - Stochastic estimates of functions and derivatives are computed.

$$\Downarrow$$

random models are employed.

Bandeira, Vicente Scheinberg, SIOPT, 2014
Chen, Menickelly, Scheinberg, Math. Prog., 2018
Bollapragada, Byrd, and Nocedal, SIOPT 2018
Blanchet, Cartis, Menickelly, Scheinberg, INFORMS J. on Opt. 2019
B., Gurioli, Morini, Toint, SIOPT 2019, & J. of Complexity 2021
Paquette, Scheinberg, SIOPT 2020
Xu, Roosta, Mahoney, Math. Prog. 2020
Berahas, Cao, Scheinberg, SIOPT 2021
B., Gurioli, Morini, Toint, ArXiv, 2021.
B., Krejić, Morini, Rebegoldi, ArXiv, 2021
di Serafino, Krejić, Krklec Jerinkić, Viola, ArXiv 2021
Bergou, Diouane, Kunc, Kungurstev, Royer, INFORMS J. Optim., 2022

0. Given $x_k \in \mathbb{R}^n$, $\eta \in (0,1)$, $\gamma > 1$, and the trust-region radius $\delta_k > 0$.
1. Compute a trial step
   Compute the model $m_k(p)$ and an (approximate) solution of the trust-region problem
   $$\min_p m_k(p) \quad \text{s.t.} \; \|p\| \leq \delta_k$$
2. Check decrease
   $$\rho_k(p_k) = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}$$
3. Successful iteration
   If $\rho_k \geq \eta$ then set $\delta_{k+1} = \gamma\delta_k$ and $x_{k+1} = x_k + p_k$.
4. Unsuccessful iteration
   If $\rho_k < \eta$ then $\delta_{k+1} = \gamma^{-1}\delta_k$ and $x_{k+1} = x_k$

Trust-Region method **with random models**

kth iteration

0. Given $x_k \in \mathbb{R}^n$, $\eta \in (0,1)$, $\gamma > 1$, and the trust-region radius $\delta_k > 0$.
1. Compute a trial step
   Compute a random model $\overline{m}_k(p)$ and an (approximate) solution of the trust-region problem
   $$\min_p \overline{m}_k(p) \quad \text{s.t.} \; \|p\| \le \delta_k$$
2. Check decrease
   $$\rho_k(p_k) = \frac{f(x_k) - f(x_k + p_k)}{\overline{m}_k(0) - \overline{m}_k(p_k)}$$
3. Successful iteration
   If $\rho_k \ge \eta$ then set $\delta_{k+1} = \gamma \delta_k$ and $x_{k+1} = x_k + p_k$.
4. Unsuccessful iteration
   If $\rho_k < \eta$ then set $\delta_{k+1} = \gamma^{-1}\delta_k$ and $x_{k+1} = x_k$

Stochastic Trust-Region

kth iteration

0. Given $x_k \in \mathbb{R}^n$, $\eta \in (0,1)$, $\gamma > 1$, and the trust-region radius $\delta_k > 0$.
1. Compute a trial step
   Compute a random model $\overline{m}_k(p)$ and an (approximate) solution of the trust-region problem
   $$\min_p \overline{m}_k(p) \quad \text{s.t.} \ \|p\| \leq \delta_k$$
2. Guess decrease
   Compute $\overline{f}(x_k)$ and $\overline{f}(x_k + p_k)$ estimate of $f(x_k)$ and $f(x_k + p_k)$ and
   $$\rho_k(p_k) = \frac{\overline{f}(x_k) - \overline{f}(x_k + p_k)}{\overline{m}_k(0) - \overline{m}_k(p_k)}$$
3. Successful iteration
   If $\rho_k \geq \eta$ then set $\delta_{k+1} = \gamma\delta_k$ and $x_{k+1} = x_k + p_k$.
4. Unsuccessful iteration
   If $\rho_k < \eta$ then set $\delta_{k+1} = \gamma^{-1}\delta_k$ and $x_{k+1} = x_k$

Blanchet, Cartis, Menickelly, Scheinberg, INFORMS J. on Opt. (2019)

B., Gurioli, Morini, Toint, arXiv:2112.06176 (2021)

Stochastic Trust-Region -First order method

kth iteration

0. Given $x_k \in \mathbb{R}^n$, $\eta \in (0,1)$, $\gamma > 1$, and the trust-region radius $\delta_k > 0$,
1. Compute a trial step
   Compute a a random estimate $\overline{\nabla f}(x_k)$ of $\nabla f(x_k)$ and set

$$p_k = -\frac{\delta_k}{\|\overline{\nabla f}(x_k)\|}\overline{\nabla f}(x_k)$$

2. Guess decrease
   Compute $\overline{f}(x_k)$ and $\overline{f}(x_k + p_k)$ estimate of $f(x_k)$ and $f(x_k + p_k)$ and

$$\rho_k(p_k) = \frac{\overline{f}(x_k) - \overline{f}(x_k + p_k)}{\|\overline{\nabla f}(x_k)\|\delta_k}$$

3. Successful/unsuccesful iteration
   If $\rho_k \geq \eta$ then set $\delta_{k+1} = \gamma\delta_k$ and $x_{k+1} = x_k + p_k$.
   If $\rho_k < \eta$ then set $\Delta_{k+1} = \gamma^{-1}\delta_k$ and $x_{k+1} = x_k$

Stochastic Trust-Region -First order method

kth iteration

0. Given $x_k \in \mathbb{R}^n$, $\eta \in (0,1)$, $\gamma > 1$, and the trust-region radius $\delta_k > 0$,
1. Compute a trial step
   Compute a a random estimate $\overline{\nabla f}(x_k)$ of $\nabla f(x_k)$ and set

$$p_k = -\frac{\delta_k}{\|\overline{\nabla f}(x_k)\|}\overline{\nabla f}(x_k)$$

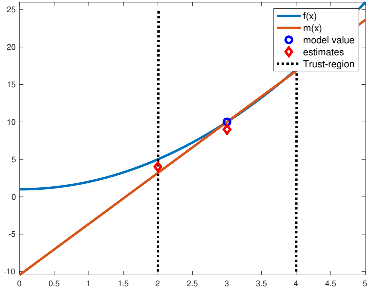
2. Guess decrease
   Compute $\overline{f}(x_k)$ and $\overline{f}(x_k + p_k)$ estimate of $f(x_k)$ and $f(x_k + p_k)$ and

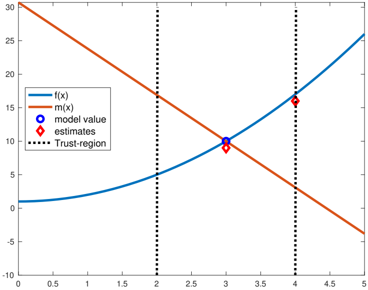
$$\rho_k(p_k) = \frac{\overline{f}(x_k) - \overline{f}(x_k + p_k)}{\|\overline{\nabla f}(x_k)\|\delta_k}$$

3. Successful/unsuccesful iteration
   If $\rho_k \geq \eta$ then set $\delta_{k+1} = \gamma\delta_k$ and $x_{k+1} = x_k + p_k$.
   If $\rho_k < \eta$ then set $\Delta_{k+1} = \gamma^{-1}\delta_k$ and $x_{k+1} = x_k$

Stochastic gradient method with adaptive choice of the steplenght (learning rate)!

Possible iteration outcomes

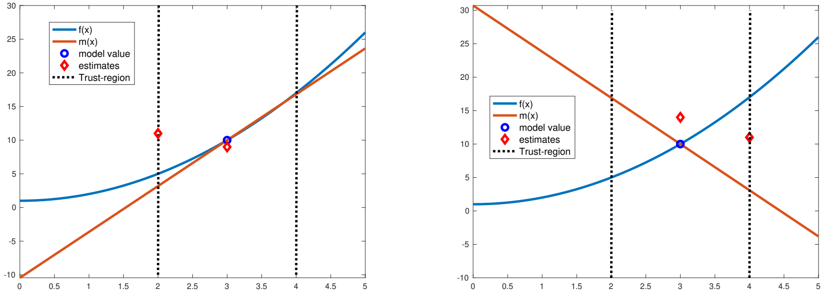




Good model, good estimates (left), Bad model, good estimates (right)

True successful   True unsuccessful

Possible iteration outcomes



Good model, bad estimates (left), Bad model, bad estimates (right)

False unsuccessful   False successful

- What does it mean "good" model/estimations?
- How often can we have false successful/unsuccessful iterations?

Adaptive accuracy requirements

An $\boxed{\textit{informal}}$ statement of our assumptions:

Consider the events

$$\mathcal{M}_k = \{\|G_k - \nabla f(X_k)\| \le \nu \|G_k\|)\}$$

$$\mathcal{F}_k = \{max\{|F_k^0 - f(X_k)|, |F_k^p - f(X_k + P_k)|\} \le \nu \|G_k\| \Delta_k \quad \}$$

We assume that

$$\textit{Probability}\left[\mathcal{M}_k \cap \mathcal{F}_k | \text{conditioned by the past}\right] = p_* > \tfrac{1}{2}$$

the expected value of $f(X_k) - f(X_k + Pk)$ at false successful iterations, conditioned by the past, is positive.

+ f bounded below and Lipschitz continuity of $\nabla f(x)$

============

X_k, Δ_k, P_k are the random variables corresponding to the realizations x_k, δ_k, p_k.
G_k is the random variable associate with the realization $\overline{\nabla f}(x_k)$.
F_k^0, F_k^p are the random variables associated with the realizations $\overline{f}(x_k), \overline{f}(x_k + p_k)$.

Iteration complexity

Let

$$N_\epsilon = \inf \{k \geq 0 \ | \ \|\nabla f(X_k)\| \leq \epsilon\}.$$

If the stochastic Trust-region algorithm is applied to the problem

$$\min f(x)$$

then, under the stated assumptions,

$$\mathrm{E}[\mathrm{N}_\epsilon] = \mathrm{O}\left(\epsilon^{-2}\right)$$

$O\left(\epsilon^{-2}\right)$ iteration bound is sharp for TR methods using exact function and gradient evaluations.

Probability p_* is constant along the algorithm and we only require $p_* > 1/2$

B., Gurioli, Morini, Toint arXiv:2112.06176 (2021)

Ensuring the Accuracy Requirements

The Finite-Sum Minimisation Setting - Uniform Random Subsampling

Consider the *finite-sum* minimisation setting: $\min_{x\in\mathbb{R}^n} f(x),\ \ f = \frac{1}{N}\sum_{i=1}^{N}\phi_i(x)$.

- The resulting approximation

$$\overline{f}(x_k) = \frac{1}{|\mathcal{D}_k^f|}\sum_{i\in\mathcal{D}_k^f}\phi_i(x_k), \quad \overline{\nabla f}(x_k) = \frac{1}{|\mathcal{D}_k^g|}\sum_{i\in\mathcal{D}_k^g}\nabla\phi_i(x_k),$$

  with $\mathcal{D}_k^f, \mathcal{D}_k^g \subseteq \{1,2,\ldots,N\}$ (*randomly* and *uniformly* taken).

Ensuring the Accuracy Requirements

The Finite-Sum Minimisation Setting - Uniform Random Subsampling

Consider the *finite-sum* minimisation setting: $\min_{x\in\mathbb{R}^n} f(x),\ f = \frac{1}{N}\sum_{i=1}^{N}\phi_i(x).$

- The resulting approximation

$$\overline{f}(x_k) = \frac{1}{|\mathcal{D}_k^f|}\sum_{i\in\mathcal{D}_k^f}\phi_i(x_k), \quad \overline{\nabla f}(x_k) = \frac{1}{|\mathcal{D}_k^g|}\sum_{i\in\mathcal{D}_k^g}\nabla\phi_i(x_k),$$

  with $\mathcal{D}_k^f, \mathcal{D}_k^g \subseteq \{1,2,\dots,N\}$ (*randomly* and *uniformly* taken).

- *Probability* $\left[\mathcal{M}_k \cap \mathcal{F}_k | \text{conditioned by the past}\right] \geq p_* = \alpha_*\beta_*$ if

Adaptive choice of the sample size

$$|\mathcal{D}_k^f| = O\left(\frac{1}{\nu\|\overline{\nabla f}(x_k)\|^2\delta_k^2}\log(\frac{1}{1-\beta_*})\right) \quad |\mathcal{D}_k^g| = O\left(\frac{1}{\zeta_k^2}\log(\frac{1}{1-\alpha_*})\right)$$

where $\zeta_k < \nu\|\overline{\nabla f}(x_k)\|$ (requires an inner loop).

========================

$\mathcal{M}_k = \{\|G_k - \nabla f(X_k)\| \leq \nu\|G_k\|)\}$, $\mathcal{F}_k = \{max\{|F_k^0 - f(X_k)|, |F_k^p - f(X_k + P_k)|\} \leq \nu\|G_k\|\Delta_k\}$

An example: classification problems

- Logistic loss: given $\{(a_i, b_i)\}_{i=1}^N$

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\log(1 + e^{-b_i a_i^T x})}_{\phi_i(x)} + \frac{1}{2N}\|x\|^2,$$

The classifier is such that

$$\frac{1}{1+e^{-a_i^T x}} \geq 0.5 \qquad b_i = 1$$
$$\frac{1}{1+e^{-a_i^T x}} < 0.5, \qquad b_i = 0$$

- Nonlinear least squares problems: given $\{(a_i, b_i)\}_{i=1}^N$

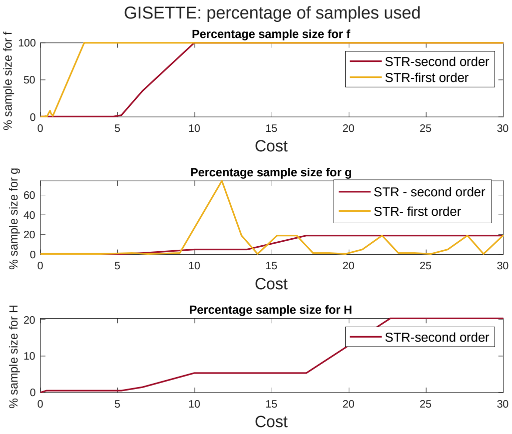
$$f(x) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\left(b_i - \frac{1}{1+e^{-a_i^T x}}\right)^2}_{\phi_i(x)}$$

- The main cost in the computation of ϕ_i is the scalar product $a_i^T x$.
- Props: Number of Propagations (1 full function and gradient evaluation is counted as 2 Prop). A maximum number of Props is considered as a termination criterion.
- Computing $\overline{f}(x)$ and $\overline{\nabla f}(x)$ costs $\frac{|\mathcal{D}_k^f|+|\mathcal{D}_k^g|}{N}$ props.

STR - first and second order: Adaptive sample size choice

N: 4800 $n = 5000$, Testing 1200

Average Accuracy STR- first order 87.85%, STR- second order 94.67%



Stocastic trust region & inexact restoration

kth iteration

0. Given $x_k \in \mathbb{R}^n$, $\eta \in (0,1)$, $\gamma > 1$, and the trust-region radius $\Delta_k > 0$,
1. Compute a trial step
   Choose *randomly* and *uniformly* $\mathcal{D}_k^g, \subseteq \{1,2,\dots,N\}$, compute $\overline{\nabla f}(x_k) = \frac{1}{|\mathcal{D}_k^g|}\sum_{i\in\mathcal{D}_k^g}\nabla\phi_i(x_k)$ and set

$$p_k = -\frac{\delta_k}{\|\overline{\nabla f}(x_k)\|}\overline{\nabla f}(x_k)$$

2. Guess decrease
   Compute $\overline{f}(x_k + p_k)$ and $\overline{f}(x_k)$ by subsampling in $\mathcal{D}_k^g$
   and $\rho_k(p_k)$ given by the inexact-restoration step acceptance rule.
3. Successful/unsuccessful iteration
   If $\rho_k \geq \eta$ and $\|\overline{\nabla f}(x_k)\| \geq \eta_2\delta_k$ then set $\delta_{k+1} = \gamma\delta_k$ and $x_{k+1} = x_k + p_k$.
   Otherwise set $\delta_{k+1} = \gamma^{-1}\delta_k$ and $x_{k+1} = x_k$

Stocastic trust region & inexact restoration

kth iteration

0. Given $x_k \in \mathbb{R}^n$, $\eta \in (0,1)$, $\gamma > 1$, and the trust-region radius $\Delta_k > 0$,
1. Compute a trial step
   Choose *randomly* and *uniformly* $\mathcal{D}_k^g, \subseteq \{1, 2, \ldots, N\}$, compute $\overline{\nabla f}(x_k) = \frac{1}{|\mathcal{D}_k^g|} \sum_{i \in \mathcal{D}_k^g} \nabla \phi_i(x_k)$ and set
   $$p_k = -\frac{\delta_k}{\|\overline{\nabla f}(x_k)\|} \overline{\nabla f}(x_k)$$
2. Guess decrease
   Compute $\overline{f}(x_k + p_k)$ and $\overline{f}(x_k)$ by subsampling in $\mathcal{D}_k^g$ and $\rho_k(p_k)$ given by the inexact-restoration step acceptance rule.
3. Successful/unsuccessful iteration
   If $\rho_k \geq \eta$ and $\|\overline{\nabla f}(x_k)\| \geq \eta_2 \delta_k$ then set $\delta_{k+1} = \gamma \delta_k$ and $x_{k+1} = x_k + p_k$.
   Otherwise set $\delta_{k+1} = \gamma^{-1} \delta_k$ and $x_{k+1} = x_k$

The function approximation is computed averaging in the same subsample used for the gradient approximation!

B., Krejić, Morini, Rebegoldi *A stochastic first-order trust-region method with inexact restoration for finite-sum minimization*, Arxiv2107.03129, 2021

Inexact-restoration step acceptance

Given $x_k, \mathcal{D}_k^g, \mathcal{D}_{k-1}^g, \theta_k, p_k$.

- Let $\overline{f}_{k-1}(x_k) = \frac{1}{|\mathcal{D}_{k-1}^g|} \sum_{i \in \mathcal{D}_{k-1}^g} \phi_i(x_k)$ be the estimate computed at the previous iteration and

$$\rho_k = \frac{\mathrm{Ared}_k(\theta_{k+1})}{\mathrm{Pred}_k(\theta_{k+1})}$$

- $\mathrm{Pred}_k(\theta_{k+1}) = \theta_{k+1}(\overline{f}_{k-1}(x_k) - \underbrace{(\overline{f}(x_k) + \overline{\nabla f}(x_k)^T p_k)}_{m_k(p_k)}) + (1 - \theta_{k+1})\frac{|\mathcal{D}_k^g| - |\mathcal{D}_{k-1}^g|}{N}$
- $\mathrm{Ared}_k(\theta_{k+1}) = \theta_{k+1}(\overline{f}_{k-1}(x_k) - \overline{f}(x_k + p_k)) + (1 - \theta_{k+1})\frac{|\mathcal{D}_k^g| - |\mathcal{D}_{k-1}^g|}{N}$
- $\theta_{k+1} \in (0, 1)$ s.t.

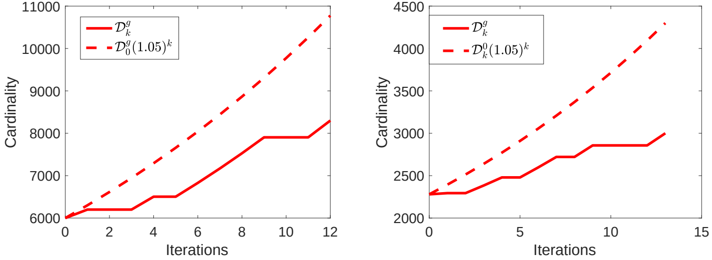
$$\mathrm{Pred}_k(\theta_{k+1}) \geq \eta \frac{|\mathcal{D}_k^g| - |\mathcal{D}_{k-1}^g|}{N}.$$

We balance the increase/decrease in the approximated objective function with the increase/decrease in the sample size.

History: sample size versus iterations

MNIST problem $N = 60000$
Average accuracy: 86,90%

A9A problem $N = 22793$.
Average accuracy: 98,32%



'-': SIRTR '- -': $\mathcal{D}_{k+1}^g = 1.05\mathcal{D}_k^g$

TRISH: trust-region without adaptive choice of the learning rate

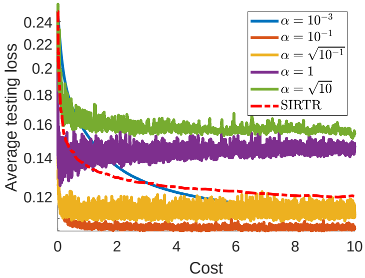
SIRTR versus Trust-Region-ish algorithm (TRish)[1].
TRish is a stochastic gradient method based on a trust-region methodology. Normalized steps are used in a dynamic manner whenever the norm of the stochastic gradient is within a prefixed interval. The $k-$th iteration of TRish is given by

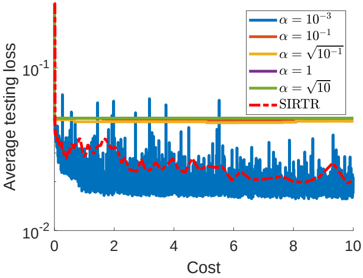
$$x_{k+1} = x_k - \begin{cases} \gamma_{1,k}\alpha_k \overline{\nabla f}_k, & \text{if } \|\overline{\nabla f}_k\| \in \left[0, \frac{1}{\gamma_{1,k}}\right) \\ \alpha_k \frac{\overline{\nabla f}_k}{\|\overline{\nabla f}_k\|}, & \text{if } \|\overline{\nabla f}_k\| \in \left[\frac{1}{\gamma_{1,k}}, \frac{1}{\gamma_{2,k}}\right] \\ \gamma_{2,k}\alpha_k \overline{\nabla f}_k, & \text{if } \|\| \in \left(\frac{1}{\gamma_{2,k}}, \infty\right) \end{cases}$$

where $\alpha_k > 0$ is the steplength parameter, $0 < \gamma_{2,k} < \gamma_{1,k}$ are positive constants.

[1] F.E. Curtis, K. Scheinberg, R. Shi, INFORMS Journal on Optimization 1(3), 200–220, 2019.

Avoiding learning rate tuning



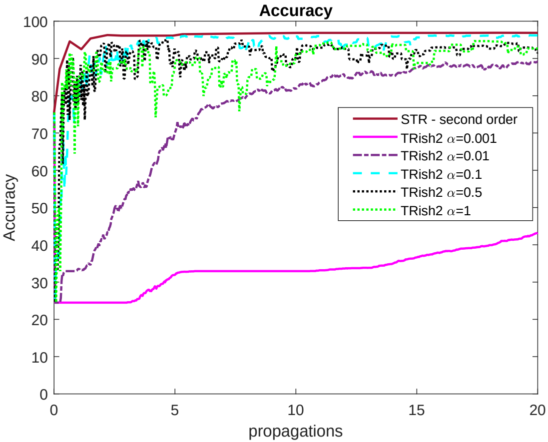


SIRTR versus Trust-Regionish algorithm for several choices of the steplength α.
Decrease of the (average) testing loss $\overline{f}(x_k)$ w.r.t. the (average) computational time.
From left to right: a9a and htru2 datasets.

Avoiding Learning rate tuning (2)

Mushrooms dataset, Training $N = 5000$, $n = 112$, Testing 1600, batch-size=50

TRrish2 $\gamma_1 = 4/G$, $\gamma_2 = 1/(2G)$ G : average norm of stochastic gradient estimates provided by SDG, $\alpha = 0.1$.



Conclusion, remarks ..

- The stocastic trust-region approach has been extended to polynomial models of arbitrary degree:
  seek for first- and second-order critical points,
  and also for critical points of arbitrary order
- Adaptive accuracy, finite sum context:
  adaptive choice of the steplength and of the subsample sizes
- Second order methods:
  Inexact steps + matrix-free implementation produce
  a significative reduction of each iteration cost

Conclusion, remarks ..

- The stocastic trust-region approach has been extended to polynomial models of arbitrary degree:
  seek for first- and second-order critical points,
  and also for critical points of arbitrary order
- Adaptive accuracy, finite sum context:
  adaptive choice of the steplength and of the subsample sizes
- Second order methods:
  Inexact steps + matrix-free implementation produce
  a significative reduction of each iteration cost

Thank you!

Some references

- S.B., N.Krejić, B.Morini, S.Rebegoldi, *A stochastic first-order trust-region method with inexact restoration for finite-sum minimization*, Arxiv: 2107.03129, 2021
- S. B., G. Gurioli, B. Morini and Ph. L. Toint, *"Trust-region algorithms: probabilistic complexity and intrinsic noise with applications to subsampling techniques"*, arXiv:2112.06176, 2021.
- S. B., G. Gurioli, B. Morini and Ph. L. Toint, *"Trust-region algorithms: probabilistic complexity and intrinsic noise with applications to subsampling techniques"*, arXiv:2104.02519, 2021.
- S. B., G. Gurioli, B. Morini and Ph. L. Toint, *"Adaptive regularization for nonconvex optimization using inexact function values and randomly perturbed derivatives"*, Journal of Complexity, 2022.
- S. B., G. Gurioli, *"Complexity Analysis of a Stochastic Cubic Regularisation Method under Inexact Gradient Evaluations and Dynamic Hessian Accuracy"*, Optimization, 2021.
- S. B., G. Gurioli, B. Morini, *"Adaptive cubic regularization methods with dynamic inexact Hessian information and applications to finite-sum minimization"*, IMA Journal Numerical Analysis, 2021.
- S. B., N.Krejić, B. Morini *"Inexact restoration with subsampled trust-region methods for finite-sum minimization"*, COAP, 2020.
- S. B., N.Krejić, N. Krklec Jerinkic, *"Subsampled Inexact Newton methods for minimizing large sums of convex functions"*, IMA J. Numer. Anal. 2020.

Grants: Indam-GNCS, UNIFI Internationalization plan, "Second order methods for optimisation problems in machine learning", project for the exchange of researchers within the frame of the executive programme of Scientific and Technological cooperation between the Italian Republic and the Republic of Serbia for the years 2019-2022.