# Progetto di Ricerca GNCS 2019: Tecniche adattive per metodi di ottimizzazione in Machine Learning

Responsabile: Stefania Bellavia
Dipartimento di Ingegneria Industriale
Università di Firenze


Convegno GNCS 2020
Montecatini Terme 11-13 Febbraio 2020

# Partecipanti (8 strutturati, 7 non strutturati)

- UNIFI: Stefania Bellavia, Benedetta Morini, Alessandra Papini, Gianmarco Gurioli, Cristina Sgattoni
- UNIBO: Elena Loli Piccolomini, Elena Morotti
- UNIMORE: Marco Prato, Luca Zanni, Carla Bertocchi, Giorgia Franchini, Mathilde Galinier
- UNINA: Gerardo Toraldo
- UNIFE: Gaetano Zanghirati, Serena Crisci

# The problem

## Finite-sums minimization problem

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

- $\phi_i : \mathbb{R}^n \to \mathbb{R}$, $\phi_i \in C^2(\mathbb{R}^n)$, $i = 1, \ldots, N$ and $f$ bounded below.
- We look for $\epsilon_g$- approximate first-order critical points:

$$\|\nabla f(\hat{x})\| \leq \epsilon_g$$

- When $N$ is large the evaluation of $f$ and its derivative information is computationally expensive.

# Optimization problems in ML

- Goal: determine a prediction function $h : \mathcal{A} \to \mathcal{B}$ such that, given $a \in \mathcal{A}$, the value $h(a)$ offers an accurate prediction about the true output $b$ associated to the input $a$.
- Choose a prediction function parametrized by a vector $x \in \mathbb{R}^n$,

$$h \in \mathcal{H} = \{h(\cdot; x) : x \in \mathbb{R}^n\}$$

minimizing a risk measure over $x$.

# Optimization problems in ML

- Goal: determine a prediction function $h : \mathcal{A} \to \mathcal{B}$ such that, given $a \in \mathcal{A}$, the value $h(a)$ offers an accurate prediction about the true output $b$ associated to the input $a$.
- Choose a prediction function parametrized by a vector $x \in \mathbb{R}^n$,

$$h \in \mathcal{H} = \{h(\cdot; x) : x \in \mathbb{R}^n\}$$

  minimizing a risk measure over $x$.
- Common settings: given a loss function $\ell : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ 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):

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

Supervised Learning- Empirical Risk

- Sample average approximation of $f(x) = \mathbb{E}[\ell(h(a; x), b)]$

# Nonlinear least-squares problems

Given $\{(a_i, b_i)\}_{i=1}^N$, $a_i \in \mathbb{R}^n$, $b_i \in [0,1]$

Minimize the empirical risk using the square-loss function and the sigmoid function as a prediction model:

$$f_N(x) = \frac{1}{N} \sum_{i=1}^{N} \left( b_i - \underbrace{\frac{1}{1 + e^{-a_i^T x}}}_{\text{sigmoid}} \right)^2 \qquad \text{non-convex}$$

- As $(-a_i^T x)$ goes from $-\infty$ to $\infty$, $\dfrac{1}{1 + e^{-a_i^T x}}$ goes from 0 to 1.
- $\dfrac{1}{1 + e^{-a_i^T x}}$ has a sigmoid shape (i.e., S-like shape).

# Binary classification problems

Given $\{(a_i, b_i)\}_{i=1}^N$, $a_i \in \mathbb{R}^n$, $b_i \in \{-1, +1\}$.

- Logistic loss:

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

Given $\hat{x}$ resulting from the classifier training, $\hat{a} \in \mathbb{R}^n$

$$\frac{1}{1 + e^{-\hat{a}^T \hat{x}}} \geq 0.5 \quad \hat{b} = 1$$

$$\frac{1}{1 + e^{-\hat{a}^T \hat{x}}} < 0.5, \quad \hat{b} = -1$$

- For correctly classified points $(-b_i a_i^T x)$ is negative, and $\log(1 + e^{-b_i a_i^T x})$ is near zero.
- For incorrectly classified points $(-b_i a_i^T x)$ is positive, and $\log(1 + e^{-b_i a_i^T x})$ can be large.

The logistic loss with $\ell_2$ regularization is strongly convex:

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

# Subsampling

$N$ is large

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

$$I_M \subseteq \{1, \ldots, N\}, \quad card(I_M) = M, \quad M \geq 1,$$

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

- A training set shows redundancy in the data $\Rightarrow$ using all the sample data in every optimization iteration is inefficient
- Work with small samples (at least initially)
- Methods in literature use subsampled $f$ and/or $\nabla f$ and/or $\nabla^2 f_M$

# Choosing the sample size

- Fixed fraction of all the sample data, at each iteration.
- Increase the sample size by a certain percentage in each iteration. If $f_k$ and derivatives have increasing, up to full, accuracy along the iterations then properties relies on the deterministic analysis.
- More elaborate schemes: function/gradient/Hessian estimates are supposed to be sufficiently accurate with prefixed probability. Then, $M$ is such that:

$$\begin{aligned} Pr(|f(x_k) - f_M(x_k)| \leq \tau_{k,0}) &\geq p_0 \\ Pr(\|\nabla f(x_k) - \nabla f_M(x_k)\| \leq \tau_{k,1}) &\geq p_1 \\ Pr(\|\nabla^2 f(x_k) - \nabla^2 f_M(x_k)\| \leq \tau_{k,2}) &\geq p_2 \end{aligned}$$

$\tau_{k,0},\ \tau_{k,1},\ \tau_{k,2} > 0,\ p_0,\ p_1,\ p_2 \in [0, 1)$.

- Convergence/complexity results in expectation/high probability.

# Stochastic Gradient methods

Choose $I_M$, $\alpha_k > 0$. Set

$$x_{k+1} = x_k - \alpha_k \nabla f_M(x_k) = x_k - \frac{\alpha_k}{M} \sum_{i \in I_M} \nabla \phi_i(x_k)$$

$M = 1$: simple/basic Stochastic Gradient;

$M > 1$: mini-batch Stochastic Gradient;

- Constant stepzise: $\alpha_k = \alpha$, $\forall k$ (knowledge of the Lipschitz constant needed).
- Diminishing stepsizes ( Robbins, Monro, 1951 ): $\sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty$
- If the approximated gradient has a large variance the convergence is slow $\Rightarrow$ variance reduction techniques where the full gradient is computed each $m$ iterations.

Convergence results in expectations on $f_N(x_k) - f_N(x^*)$ for strongly convex functions and results in expectations on $\nabla f_N(x_k)$ for nonconvex functions.

Bottou, Curtis and Nocedal, SIREV 2018

# Adaptive choice of the sample size

Choose $M$ such that: $Pr(|f(x_k) - f_M(x_k)| \leq \tau_{k,0}) \geq p_0$

- $M$ depends on the variance of the $\phi_i$ / $\nabla\phi_i$ / $\nabla^2\phi_i$,
- For random and uniform samples from $\{1, \ldots N\}$:

$$N \geq M \geq \mathcal{O}\left(\frac{V_f}{\tau_{k,0}^2} \ln \frac{2}{1-p_0}\right) \quad \text{or} \quad N \geq M \geq \mathcal{O}\left(\frac{\kappa_\phi^2}{\tau_{k,0}^2} \ln \frac{2}{1-p_0}\right)$$

$V_f$ variance of function realizations, $\max_{1 \leq i \leq N} |\phi_i(x_k)| \leq \kappa_\phi$

- Analogous results for $\nabla f_M(x_k)$, $\nabla^2 f_M(x_k)$.

Troop, Foundations and Trends in Machine Learning, 2015.

# GNCS 2020 project

- Limited-memory steplength rules in mini-batch stochastic gradient methods and box-constrained optimization
  *S. Crisci, V. Ruggiero, L. Zanni*, Applied Mathematics and Computation, 2019,
  *S. Crisci, F. Porta, V. Ruggiero, L. Zanni*, submitted,
  *G. Franchini, V. Ruggiero, L. Zanni*, submitted.
- Adaptive accuracy in functions and derivatives ( adaptive choice of the sample size)
  *S.B., G. Gurioli, B. Morini, Ph. Toint*, SIOPT 2019,
  *S.B., G. Gurioli, B. Morini*, IMA J. Numer. Anal., to appear,
  *S.B., G. Gurioli*, arXiv 2020,
  *S.B., B. Morini, N. Krejic*, arXiv 2019,
  *S.B., N. Krejic, N. Krklec Jerinkic*, IMA J. Numer. Anal., to appear.
- Semi-supervised algorithms producing models that take advantage of unlabeled samples
  *M. Viola, M. Sangiovanni, G. Toraldo, M. R. Guarracino*, Annals of Oper. Res., 2019.
- Deep learning techniques for adaptively estimating model/algorithm parameters in image reconstruction problems
  *C. Bertocchi, E. Chouzenoux, M. Corbineau, J.C. Pesquet, M. Prato*, Inverse Problems, 2019,
  *T.A. Bubba, M. Galinier, M. Lassas, M. Prato, L. Ratti, S. Siltanen*, in preparation.
- Accurate image super-resolution using convolutional networks
  *P. Cascarano, E. Loli Piccolomini, E. Polini*, in preparation.
- Matrix completion via SDP reformulation
  *S.B., J. Gondzio, M. Porcelli*, arXiv 2019.

# Adaptive Regularisation Approaches for $\min_{x \in \mathbb{R}^n} f(x)$

At a generic iteration $k$, given $x_k$ and the regularization parameter $\sigma_k$, the step $s_k$ used to compute $x_{k+1} = x_k + s_k$ is an approximate minimizer of the regularized model of $f$ at $x_k$:

$$m(x_k, s, \sigma_k) = T_q(x_k, s) + \frac{\sigma_k}{q+1}\|s\|^{q+1} \qquad q = 1, 2.$$

$$\begin{aligned} T_1(x_k, s) &= f(x_k) + \nabla f(x_k)^T s && \text{first-order method} \\ T_2(x_k, s) &= f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s && \text{second-order method} \end{aligned}$$


Nesterov and Polyak Mathematical Programming Ser. A (2006).

Cartis, Gould, Toint Mathematical Programming Ser. A (2011).

Birgin, Gardenghi, Martínez, Santos, Toint Mathematical Programming Ser. A (2017).

# Optimal complexity

- $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2(\mathbb{R}^n)$ and bounded below.
- $\epsilon_g$ approximate first-order critical point:

$$\|\nabla f(\hat{x})\| \leq \epsilon_g$$

## Optimal Complexity

An $\epsilon_g$ approximate first-order critical point is found in at most

$$\begin{array}{ll} O\left(\epsilon_g^{-2}\right) & q = 1 \\ O\left(\epsilon_g^{-3/2}\right) & q = 2 \end{array}$$

iterations, (against the $O\left(\epsilon_g^{-2}\right)$ of Trust Region methods with $q = 2$).

# The Algorithm ($\epsilon_g$-approximate first-order critical point)

**Step 0: Initialization**. Given $x_0$, $\sigma_0 > 0$, the accuracy level $\epsilon_g$. Set $k = 0$.

**Step 1: Test for termination**. If $\|\nabla f(x_k)\| \leq \epsilon_g$, terminate with the approximate solution $\widehat{x} = x_k$. Otherwise, compute $\nabla^2 f(x_k)$ (if $q = 2$).

**Step 2: Step computation.** Compute $s_k$ approximate minimizer of $m(x_k, s, \sigma_k) = T_q(x_k, s) + \frac{\sigma_k}{q+1}\|s\|^{q+1}$ s.t.:

$$m(x_k, s_k, \sigma_k) < m(x_k, 0, \sigma_k), \qquad \|\nabla_s m(x_k, s_k, \sigma_k)\| \leq \theta \|s_k\|^2.$$

**Step 3: Acceptance of the trial step.** Compute $f(x_k + s_k)$ and set

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_q(x_k, 0) - T_q(x_k, s_k)}.$$

If $\rho_k \geq \eta_1$, define $x_{k+1} = x_k + s_k$; otherwise, define $x_{k+1} = x_k$.

**Step 4: Regularisation parameters update.** Set

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1 \sigma_k), \sigma_k], & \text{if } \rho_k \geq \eta_2 \quad (\textit{very successful iteration}) \\ [\sigma_k, \gamma_2 \sigma_k], & \text{if } \rho_k \in [\eta_1, \eta_2) \quad (\textit{successful iteration}) \\ [\gamma_2 \sigma_k, \gamma_3 \sigma_k], & \text{if } \rho_k < \eta_1 \quad (\textit{unsuccessful iteration}) \end{cases}$$

Set $k = k + 1$ and go to Step 1 if $\rho_k \geq \eta_1$, or to Step 2 otherwise.

# Our goal: reducing the overall computational cost

- Approximate function and derivatives preserving the optimal complexity;
- Adaptively chosen accuracy in function and derivatives;
- Computable accuracy requirements;

# Our goal: reducing the overall computational cost

- Approximate function and derivatives preserving the optimal complexity;
- Adaptively chosen accuracy in function and derivatives;
- Computable accuracy requirements;
- Accuracy requirements guaranteed only with high probability;
- Probabilities are not increasing, they need to be above a certain constant.

# Our goal: reducing the overall computational cost

- Approximate function and derivatives preserving the optimal complexity;
- Adaptively chosen accuracy in function and derivatives;
- Computable accuracy requirements;
- Accuracy requirements guaranteed only with high probability;
- Probabilities are not increasing, they need to be above a certain constant.
- Deterministic/high probability/stochastic complexity analysis

# Inexact Gradient Evaluation

## Level of Resemblance, case $q = 2$

Remind: $m(x_k, s, \sigma_k) = f(x_k) + \overline{\nabla f}(x_k)s + \frac{1}{2}s^\top \overline{\nabla^2 f}(x_k)s + \frac{\sigma_k}{3}\|s\|^3$

Approximation $\overline{\nabla f}(x_k) \in \mathbb{R}^n$ of $\nabla f(x_k) \in \mathbb{R}^n$ at the $k$-th iteration:

1. Possible choice ($\delta > 0$):

$$\|\nabla f(x_k) - \overline{\nabla f}(x_k)\| \leq \delta \|s_k\|^2$$


J. M. Kohler, A. Lucchi. "Sub-sampled cubic regularization for non-convex optimization". 34th ICML (2017).

C. Cartis, K. Scheinberg. "Global convergence rate analysis of unconstrained optimization methods based on probabilistic models".
Math. Progr., Ser. A (2018).

# Inexact Gradient Evaluation

## Level of Resemblance, case $q = 2$

Remind: $m(x_k, s, \sigma_k) = f(x_k) + \overline{\nabla f}(x_k)s + \frac{1}{2}s^\top \overline{\nabla^2 f}(x_k)s + \frac{\sigma_k}{3}\|s\|^3$

Approximation $\overline{\nabla f}(x_k) \in \mathbb{R}^n$ of $\nabla f(x_k) \in \mathbb{R}^n$ at the $k$-th iteration:

1. Possible choice ($\delta > 0$):

$$\|\nabla f(x_k) - \overline{\nabla f}(x_k)\| \leq \delta \|s_k\|^2$$

J. M. Kohler, A. Lucchi. "Sub-sampled cubic regularization for non-convex optimization". 34th ICML (2017).

C. Cartis, K. Scheinberg. "Global convergence rate analysis of unconstrained optimization methods based on probabilistic models". Math. Progr., Ser. A (2018).

2. Alternative choice:

$$\|\nabla f(x_k) - \overline{\nabla f}(x_k)\| \leq \kappa \left(\frac{1-\beta}{\sigma_k}\right)^2 \|\overline{\nabla f}(x_k)\|^2,$$

for $\kappa \geq 0$ and with $\|\overline{\nabla f}(x_k)\| \leq \kappa_g$, $\exists \kappa_g > 0$.

S. B., G. Gurioli. "Complexity Analysis of a Stochastic Cubic Regularisation Method under Inexact Gradient Evaluation and Dynamic Hessian Accuracy". arXiv.org/abs/2001.10827 (2020).

# Inexact Hessian Information

## Level of Resemblance

1. $\|\nabla^2 f_k - \overline{\nabla^2 f}_k\| \leq \chi \|s_k\|, \quad \chi > 0.$

J. M. Kohler, A. Lucchi. "Sub-sampled cubic regularization for non-convex optimization". 34th ICML (2017).

# Inexact Hessian Information

## Level of Resemblance

1. $\|\nabla^2 f_k - \overline{\nabla^2 f}_k\| \leq \chi \|s_k\|, \quad \chi > 0.$

   J. M. Kohler, A. Lucchi. "Sub-sampled cubic regularization for non-convex optimization". 34th ICML (2017).

2. $\|\nabla^2 f_k - \overline{\nabla^2 f}_k\| < \epsilon_g$

   P. Xu, F. Roosta-Khorasani, M.W. Mahoney. "Newton-Type Methods for Non-Convex Optimization Under Inexact Hessian Information". arXiv:1708.07164 (2018).

# Inexact Hessian Information

## Level of Resemblance

1. $\|\nabla^2 f_k - \overline{\nabla^2 f}_k\| \leq \chi \|s_k\|, \quad \chi > 0.$

   J. M. Kohler, A. Lucchi. "Sub-sampled cubic regularization for non-convex optimization". 34th ICML (2017).

2. $\|\nabla^2 f_k - \overline{\nabla^2 f}_k\| < \epsilon_g$

   P. Xu, F. Roosta-Khorasani, M.W. Mahoney. "Newton-Type Methods for Non-Convex Optimization Under Inexact Hessian Information". arXiv:1708.07164 (2018).

3. $\|\nabla^2 f_k - \overline{\nabla^2 f}_k\| \leq c_k, \quad c_k > 0 \quad \Rightarrow \quad \|\nabla^2 f_k - \overline{\nabla^2 f}_k\| \leq \bar{c}\|s_k\|, \quad \exists \bar{c} > 0,$

   S. B., G. Gurioli, B. Morini. "Adaptive Cubic Regularization Methods with Dynamic Inexact Hessian Information and Applications to Finite-Sum Minimization". IMA J. Numer. Anal. , to appear.

# Dynamic Inexact Hessian Information

## Adaptive Choice of the Accuracy $C_k$

Let $\overline{\nabla^2 f}(x_k) \in \mathbb{R}^{n\times n}$, approximation of $\nabla^2 f(x_k)$, $x_k \in \mathbb{R}^n$, satisfy

$$\|\nabla^2 f(x_k) - \overline{\nabla^2 f}(x_k)\| \leq c_k,$$

**Adaptive choice of the accuracy $c_k$ on the inexact Hessian**

$$c_k \leq \begin{cases} c, \quad c > 0, & \text{if} \quad \|s_k\| \geq 1, \\ \alpha(1-\beta)\|\overline{\nabla f}(x_k)\|, & \text{if} \quad \|s_k\| < 1, \end{cases}$$

for all $k \geq 0$, with $\alpha > 0$, $s_k \in \mathbb{R}^n$ and $\beta \in [0, 1)$.

# Dynamic Inexact Hessian Information

## Adaptive Choice of the Accuracy $C_k$

Let $\overline{\nabla^2 f}(x_k) \in \mathbb{R}^{n\times n}$, approximation of $\nabla^2 f(x_k)$, $x_k \in \mathbb{R}^n$, satisfy

$$\|\nabla^2 f(x_k) - \overline{\nabla^2 f}(x_k)\| \leq c_k,$$

### Adaptive choice of the accuracy $c_k$ on the inexact Hessian

$$c_k \leq \begin{cases} c, \quad c > 0, & \text{if} \quad \|s_k\| \geq 1, \\ \alpha(1-\beta)\|\overline{\nabla f}(x_k)\|, & \text{if} \quad \|s_k\| < 1, \end{cases}$$

for all $k \geq 0$, with $\alpha > 0$, $s_k \in \mathbb{R}^n$ and $\beta \in [0, 1)$.

### Main advantages of the adaptive design

- Cheap step restoration (arbitrary choice of $c > 0$), if $\|s_k\| \geq 1$;
- $c_k \approx \epsilon$ only when $\|\overline{\nabla f}(x_k)\| \approx \epsilon$ and $\|s_k\| < 1$.

# Regularization Algorithm with Inexact Derivatives ($q = 2$)

$k$-th iteration Hess. Accuracy: $c_k = c$, if $\|s_k\| \geq 1$; $c_k \leq \alpha(1-\beta)\|\overline{\nabla f}(x_k)\|$, if $\|s_k\| < 1$

**Step 1.1: Gradient approximation.** Compute $\overline{\nabla f}(x_k)$ such that (s.t.)

$$\|\nabla f(x_k) - \overline{\nabla f}(x_k)\| \leq \kappa \left(\frac{1-\beta}{\sigma_k}\right)^2 \|\overline{\nabla f}(x_k)\|^2.$$

**Step 1.2: Hessian approximation.** Compute $\overline{\nabla^2 f}(x_k)$ s.t.

$$\|\nabla^2 f(x_k) - \overline{\nabla^2 f}(x_k)\| \leq c_k.$$

**Step 2: Step computation.** Choose $\beta_k \leq \beta$. Compute the step $s_k$ satisfying $m(x_k, s_k, \sigma_k) < m(x_k, 0, \sigma_k)$ and $\|\nabla_s m(x_k, s_k, \sigma_k)\| \leq \beta_k \|\overline{\nabla f}(x_k)\|$.

**Step 3: Check on $\|s_k\|$.**

If $\|s_k\| < 1$ and flag $= 1$ and $c > \alpha(1-\beta)\|\overline{\nabla f}(x_k)\|$,

set $x_{k+1} = x_k$, $\sigma_{k+1} = \sigma_k$, (*unsuccessful iteration*)

set $c_{k+1} = \alpha(1-\beta)\|\overline{\nabla f}(x_k)\|$, flag $= 0$, $k = k+1$ and go to Step 1.2.

**Step 4: Acceptance of the trial step and parameters update.**

Compute $f(x_k + s_k)$ and the relative decrease $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_2(x_k, 0) - T_2(x_k, s_k)}$.

If $\rho_k \geq \eta$, define $x_{k+1} = x_k + s_k$,

set $\sigma_{k+1} = \max\left(\sigma_{\min}, \frac{1}{\gamma}\sigma_k\right)$ (*successful iteration*)

If $\|s_k\| \geq 1$, set $c_{k+1} = c$, flag $= 1$; else: $c_{k+1} = \alpha(1-\beta)\|\overline{\nabla f}_{k+1}\|$, flag $= 0$.

Set $k = k+1$ and "Test for termination".

else, define $x_{k+1} = x_k$, $\sigma_{k+1} = \gamma\sigma_k$, (*unsuccessful iteration*)

$c_{k+1} = c_k$, $\overline{\nabla^2 f}(x_{k+1}) = \overline{\nabla^2 f}(x_k)$, set $k = k+1$ and go to Step 2.

## Optimal Complexity

Suppose that the objective function $f$ is bounded below by $f_{\text{low}}$ .
The Regularized Algorithm with Dynamic derivative Information requires at most

$$O\left(\left\lfloor \frac{f(x_0) - f_{\text{low}}}{\epsilon_g^{3/2}} \right\rfloor\right)$$

iterations to produce an iterate $x_k$ satisfying $\|\nabla f(x_k)\| \leq \epsilon_g$.

# Dynamic Inexact Derivatives Information

## Optimal Complexity and convergence Results

### Optimal Complexity

Suppose that the objective function $f$ is bounded below by $f_{\text{low}}$ .
The Regularized Algorithm with Dynamic derivative Information requires at most

$$O\left(\left\lfloor \frac{f(x_0) - f_{\text{low}}}{\epsilon_g^{3/2}} \right\rfloor\right)$$

iterations to produce an iterate $x_k$ satisfying $\|\nabla f(x_k)\| \leq \epsilon_g$.

### First-order convergence results

- $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$ follows from the complexity results
- $\lim_{k\to\infty} \|s_k\| = 0$
- $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$.

Unsuccessful iterations in the sense of Step 3 do not occur eventually.

# Finite Sum Minimisation

## Optimal Complexity in Probability

- Let $\tau_{1,k} = \kappa \left(\frac{1-\beta}{\sigma_k}\right)^2 \|\overline{\nabla f}(x_k)\|^2$, $\tau_{2,k} = c_k$.
- Assume

$$Pr\left(\|\overline{\nabla f}(x_k) - \nabla f(x_k)\| \leq \tau_{1,k}\right) \geq p_1,$$
$$Pr\left(\|\overline{\nabla^2 f}(x_k) - \nabla^2 f(x_k)\| \leq \tau_{2,k}\right) \geq p_2,$$

  then, the probability of having sufficiently accurate gradients and Hessians along $K$ iterations is $(p_1 p_2)^K$ (independent events).
- Then, the algorithm requires at most

$$O\left(\epsilon_g^{-3/2}\right)$$

  iterations to satisfy $\|\nabla f(x_k)\| \leq \epsilon_g$ *with probability* at least $\bar{p}$ whenever $(1 - p_1 p_2) = O((1 - \bar{p})\epsilon_g^{3/2})$.

# Finite Sum Minimisation

## Optimal Complexity in Probability

- Let $\tau_{1,k} = \kappa \left(\frac{1-\beta}{\sigma_k}\right)^2 \|\overline{\nabla f}(x_k)\|^2$, $\tau_{2,k} = c_k$.
- Assume

$$Pr\left(\|\overline{\nabla f}(x_k) - \nabla f(x_k)\| \leq \tau_{1,k}\right) \geq p_1,$$
$$Pr\left(\|\overline{\nabla^2 f}(x_k) - \nabla^2 f(x_k)\| \leq \tau_{2,k}\right) \geq p_2,$$

  then, the probability of having sufficiently accurate gradients and Hessians along $K$ iterations is $(p_1 p_2)^K$ (independent events).
- Then, the algorithm requires at most

$$O(\epsilon_g^{-3/2})$$

  iterations to satisfy $\|\nabla f(x_k)\| \leq \epsilon_g$ *with probability* at least $\bar{p}$ whenever $(1 - p_1 p_2) = O((1-\bar{p})\epsilon_g^{3/2})$.

- $|\mathcal{D}_{k,j}| \geq \min\left\{N, \left\lceil \frac{4\kappa_{\varphi,j}(x_k)}{\tau_{j,k}}\left(\frac{2\kappa_{\varphi,j}(x_k)}{\tau_{j,k}} + \frac{1}{3}\right)\log\left(\frac{jn-j+2}{1-p_j}\right)\right\rceil\right\}, \qquad j = 1, 2.$

S. B., G. Gurioli, B. Morini. "Adaptive Cubic Regularization Methods with Dynamic Inexact Hessian Information and Applications to Finite-Sum Minimization". IMA J. Numer. Anal. , to appear.

# Stochastic Regularization Algorithm

Remark: The previous analysis does not say anything about the expected number of iterations needed to reach a first-order critical point.

## The Random Process

The previous algorithm generates a random process

$$\{X_k, S_k, M_k, \Sigma_k, C_k\}.$$

## Notations

- Capital letters: random variables.
- Small letters: their realisations.
- $\mathbb{E}[X]$ denotes the expected value of the random variable $X$.
- $\mathbb{1}_A$ refers to the indicator of the random event $A$ occurring (i.e. $\mathbb{1}_A(a) = 1$ if $a \in A$, otherwise $\mathbb{1}_A(a) = 0$).

# Stochastic Regularization Algorithm

Remark: The previous analysis does not say anything about the expected number of iterations needed to reach a first-order critical point.

## The Random Process

The previous algorithm generates a random process

$$\{X_k, S_k, M_k, \Sigma_k, C_k\}.$$

## Notations

- Capital letters: random variables.
- Small letters: their realisations.
- $\mathbb{E}[X]$ denotes the expected value of the random variable $X$.
- $\mathbb{1}_A$ refers to the indicator of the random event $A$ occurring (i.e. $\mathbb{1}_A(a) = 1$ if $a \in A$, otherwise $\mathbb{1}_A(a) = 0$).

## $k$-th Iteration: Formalising the Conditioning on the Past

$\mathcal{F}_{k-1}^M$ denotes the $\hat{\sigma}$-algebra induced by the random variables $M_0$,..., $M_{k-1}$, with $\mathcal{F}_{-1}^M = \hat{\sigma}(x_0)$.

# Stochastic Algorithm with Inexact Derivatives

## The Stopping Time

For a given tolerance $\epsilon_g$, let $N_{\epsilon_g}$ denotes a random variable corresponding to the number of steps until $\|\nabla f(X_k)\| \leq \epsilon_g$ occurs for the first time:

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

- $N_{\epsilon_g}$ can be seen as a stopping time for the stochastic process generated by the algorithm.

### Main goal

Bounding the expected number of iterations $\mathbb{E}[N_{\epsilon_g}]$ which is needed, in the worst-case, to reach a first-order optimal point.

# Expected number of iterations

The analysis relies on requirements that estimates are sufficiently accurate with sufficiently high probability.

- The function does not increase even if the derivatives are inaccurate.
- Probabilities are not increasing, they need to be above a certain constant.
- If gradient and/or Hessian estimate are inaccurate, they can be arbitrarily inaccurate.

# Expected number of iterations

The analysis relies on requirements that estimates are sufficiently accurate with sufficiently high probability.

- The function does not increase even if the derivatives are inaccurate.
- Probabilities are not increasing, they need to be above a certain constant.
- If gradient and/or Hessian estimate are inaccurate, they can be arbitrarily inaccurate.

## Theorem (Upper Bound on $\mathbb{E}[N_{\epsilon_g}]$)

Let $p = p_1 p_2$ ($p$ is the probability that the model is accurate).
If $p > 2/3$, then

$$\mathbb{E}[N_{\epsilon_g}] \leq \frac{3p}{(3p-2)(2p-1)} \left[ (f_0 - f_{low}) \left( 2\kappa_s \epsilon^{-3/2} + \kappa_u \right) + log_\gamma \left( \frac{\overline{\sigma}}{\sigma_0} \right) \right]$$

# Adaptive accuracy in function evaluations

Function evaluations are needed, after step computation, to compute the ratio:

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_q(x_k, 0) - T_q(x_k, s_k)} \Rightarrow \bar{\rho}_k = \frac{\bar{f}(x_k) - \bar{f}(x_k + s_k)}{\bar{T}_q(x_k, 0) - \bar{T}_q(x_k, s_k)}$$

- Function accuracy requirement ($\chi > 0$):

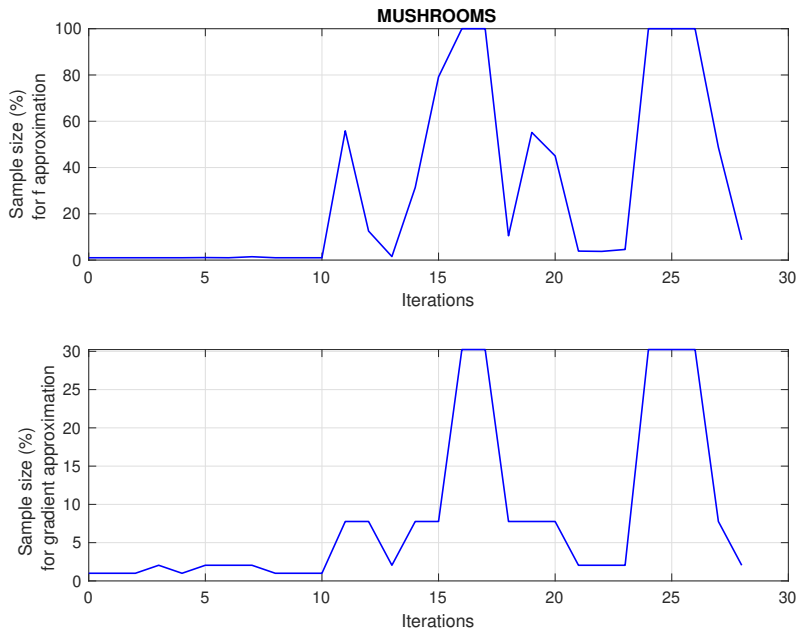
$$\begin{aligned} |\bar{f}(x_k + s_k) - f(x_k + s_k)| &\leq \chi \|s_k\|^3 \\ |\bar{f}(x_k) - f(x_k)| &\leq \chi \|s_k\|^3 \end{aligned}$$

S. B., B. Morini, G. Gurioli, Ph. L. Toint. "Adaptive Regularization Algorithms with Inexact Evaluations for Nonconvex Optimization". SIOPT (2019).

# Binary classification problem, Sample size

## The Mushrooms Dataset

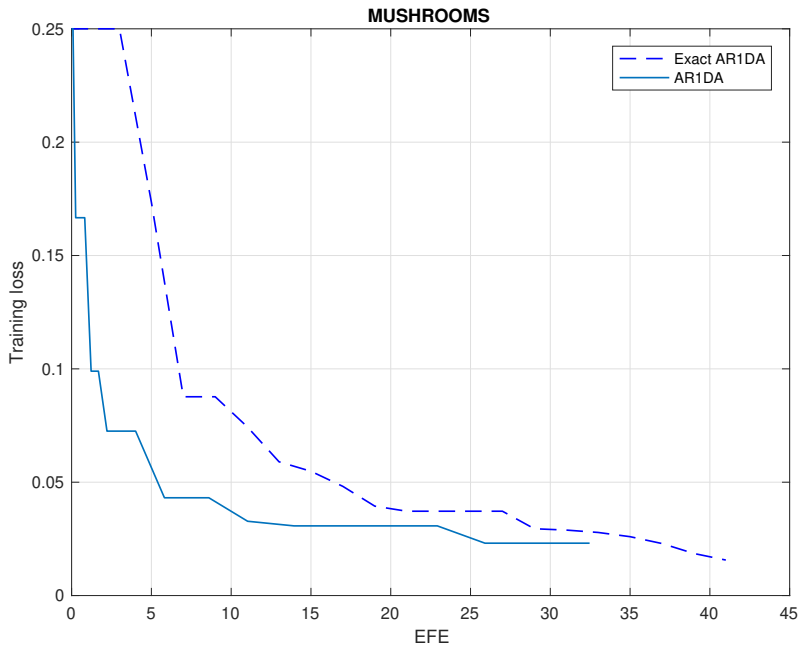
| Training Size $N$ | Test Size | $n$ | $\epsilon_g$ | Class. Rate | Class. Rate (exact method) |
|---|---|---|---|---|---|
| 6503 | 1621 | 112 | $10^{-3}$ | 98.89% | 98.33% |

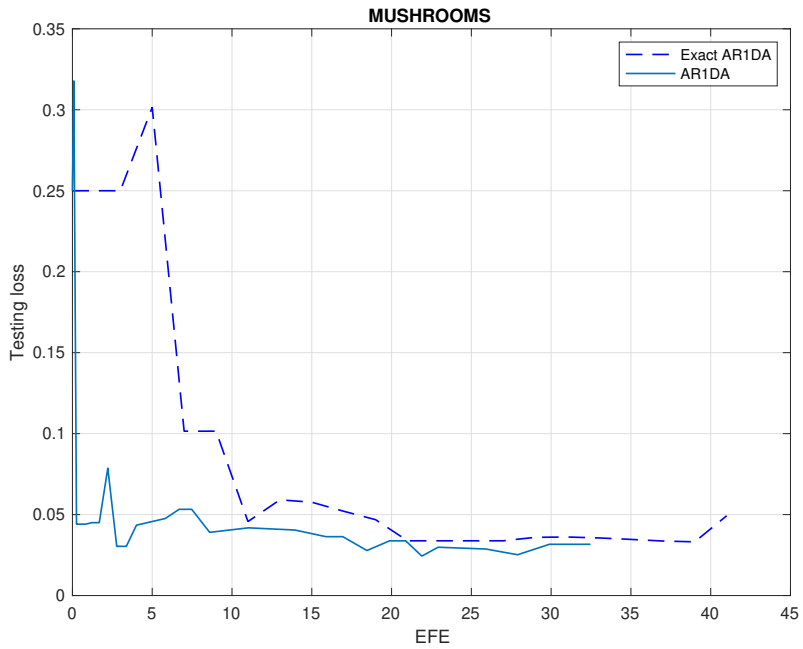



AR1DA: first-order Adaptive Regularization method, Dynamic Accuracy in function and gradient

$$|\mathcal{D}_{k,j}| \geq \min\left\{N, \left\lceil \frac{4\kappa}{\tau_{j,k}}\left(\frac{2\kappa}{\tau_{j,k}}+\frac{1}{3}\right)\log\left(\frac{j(n-1)+2}{\overline{p}_j}\right)\right\rceil\right\}, \quad \kappa = 10^{-3}, p_j = 0.85,\ j = 1, 2.$$

# Training and Testing error




Effective Function Evaluations (EFE); Sum of function and gradient evaluations;

# Budget

- Visiting Professor grant: Prof. J. Gondzio (20-29 March 2019);
- Giornata di Lavoro: January 31st 2010;
- Conferences:
  - IMA Conference of Mathematics of Operational Research (Optimization for Machine Learning and Big Data);
  - ICIAM 2019 (Nonlinear Optimization for Inverse Problems and Applications);
  - UMI 2019 (Sessione Speciale “Ottimizzazione e Problemi inversi”);
  - FGS 2019 Conference on Optimization;

Grazie per l’attenzione