# On the evaluation complexity of non-smooth composite minimization, with applications

**Coralia Cartis** (University of Edinburgh, UK)

joint with

**Nick Gould** (RAL, UK) & **Philippe Toint** (Namur, Belgium)

SIAM Conference on Imaging Science

Philadelphia, May 20–22, 2012

# Non-smooth composite function minimization

Consider the unconstrained problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad h(r(x)),$$

where $r : \mathbb{R}^n \to \mathbb{R}^p$ is smooth but nonconvex, and $h : \mathbb{R}^p \to \mathbb{R}$ is convex but possibly non-smooth. $[h = \|\cdot\|(+1)]$

[considered by Nesterov (2007, 2007) and CGT (2011)]

# Non-smooth composite function minimization

Consider the unconstrained problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad h(r(x)),$$

where $r : \mathbb{R}^n \to \mathbb{R}^p$ is smooth but nonconvex, and $h : \mathbb{R}^p \to \mathbb{R}$ is convex but possibly non-smooth. $[h = \|\cdot\|(+1)]$

[considered by Nesterov (2007, 2007) and CGT (2011)]

Evaluation complexity of generating an (approximate) first-order critical point?

$\Longleftrightarrow$ Global rates of convergence of algorithms

[does not include cost of solving the subproblem]

$\longrightarrow$ Suitable criticality measures for the non-smooth case

# Non-smooth composite function minimization

Consider the unconstrained problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad h(r(x)),$$

where $r : \mathbb{R}^n \to \mathbb{R}^p$ is smooth but nonconvex, and $h : \mathbb{R}^p \to \mathbb{R}$ is convex but possibly non-smooth. $[h = \|\cdot\|(+1)]$

[considered by Nesterov (2007, 2007) and CGT (2011)]

Evaluation complexity of generating an (approximate) first-order critical point?

$\Longleftrightarrow$ Global rates of convergence of algorithms

[does not include cost of solving the subproblem]

$\longrightarrow$ Suitable criticality measures for the non-smooth case

Methods with same worst-case evaluation complexity as in the smooth case: $\mathcal{O}(\epsilon^{-2})$ for first-order techniques. [sharp]

# Methods for non-smooth composite minimization

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad h(r(x))$$

First-order methods: compute a step $s_k$ by solving a (convex) subproblem as follows

- Quadratic regularization: for some weight $\sigma_k > 0$,

$$\underset{s \in \mathbf{R}^n}{\text{minimize}} \quad h(r(x_k) + A(x_k)s) + \frac{\sigma_k}{2}\|s\|^2.$$

  $\sigma_k$ is adaptively increased to ensure sufficient decrease.

- Trust-region: for some trust-region radius $\Delta_k > 0$,

$$\underset{s \in \mathbf{R}^n}{\text{minimize}} \quad \underbrace{h(r(x_k) + A(x_k)s)}_{l(x_k, s)} \quad \text{subject to} \quad \|s\| \leq \Delta_k.$$

  $\Delta_k > 0$ is adaptively decreased to ensure sufficient decrease.

# Quadratic Regularization Algorithm

Given $x_0$, and $\sigma_0 > 0$, for $k = 0, 1, \ldots$ until "termination",

- compute $s_k = \arg\min_{s \in \mathbb{R}^n} m_k(s) = l(x_k, s) + \frac{\sigma_k}{2}\|s\|^2$

  where $l(x_k, s) = h(r(x_k) + A(x_k)s)$

# Quadratic Regularization Algorithm

Given $x_0$, and $\sigma_0 > 0$, for $k = 0, 1, \ldots$ until "termination",

- compute $s_k = \arg\min_{s \in \mathbf{R}^n} m_k(s) = l(x_k, s) + \frac{\sigma_k}{2}\|s\|^2$

  where $l(x_k, s) = h(r(x_k) + A(x_k)s)$

- compute $\rho_k = \dfrac{h(r(x_k)) - h(r(x_k + s_k))}{h(r(x_k)) - m_k(s_k)}$

# Quadratic Regularization Algorithm

Given $x_0$, and $\sigma_0 > 0$, for $k = 0, 1, \ldots$ until "termination",

- compute $s_k = \arg\min_{s \in \mathbb{R}^n} m_k(s) = l(x_k, s) + \frac{\sigma_k}{2}\|s\|^2$

  where $l(x_k, s) = h(r(x_k) + A(x_k)s)$

- compute $\rho_k = \dfrac{h(r(x_k)) - h(r(x_k + s_k))}{h(r(x_k)) - m_k(s_k)}$

- set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > 0.1 \\ x_k & \text{otherwise} \end{cases}$

# Quadratic Regularization Algorithm

Given $x_0$, and $\sigma_0 > 0$, for $k = 0, 1, \ldots$ until "termination",

- compute $s_k = \arg\min_{s \in \mathbf{R}^n} m_k(s) = l(x_k, s) + \frac{\sigma_k}{2}\|s\|^2$

  where $l(x_k, s) = h(r(x_k) + A(x_k)s)$

- compute $\rho_k = \dfrac{h(r(x_k)) - h(r(x_k + s_k))}{h(r(x_k)) - m_k(s_k)}$

- set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > 0.1 \\ x_k & \text{otherwise} \end{cases}$

- given $\gamma_2 \geq \gamma_1 > 1$, set

$$\sigma_{k+1} \in \begin{cases} (0, \sigma_k] = \frac{1}{2}\sigma_k & \text{if } \rho_k > 0.9 \quad \textbf{very successful} \\ [\sigma_k, \gamma_1\sigma_k] = \sigma_k & \text{if } 0.1 \leq \rho_k \leq 0.9 \quad \textbf{successful} \\ [\gamma_1\sigma_k, \gamma_2\sigma_k] = 2\sigma_k & \text{otherwise} \quad \textbf{unsuccessful} \end{cases}$$

# Quadratic Regularization Algorithm: termination

Criticality measure for $h(r(x))$:

$$\Psi(x) = l(x,0) - \min_{\|s\| \leq 1} l(x,s),$$

where $l(x,0) = h(r(x))$ and $l(x,s) = h(r(x) + A(x)s)$.

[Y. Yuan (1985)]

- $\Psi(x)$ continuous for all $x$; $x_*$ minimizer $\Longrightarrow \Psi(x_*) = 0$.

Termination condition for algorithm:

$$\Psi(x_k) \leq \epsilon,$$

for a(ny) user-provided tolerance $\epsilon > 0$.

# Evaluation complexity of non-smooth composite min

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad h(r(x))$$

Main result: Assume $h$ and Jacobian $A$ are globally Lipschitz continuous and $h(r(\cdot))$ is bounded below. Then the quadratic regularization algorithm takes at most

$$\left\lceil \frac{\kappa_{\mathbf{qr}}}{\epsilon^2} \right\rceil$$

residual-evaluations to achieve

$$\Psi(x_k) = l(x_k, 0) - \min_{\|s\| \leq 1} l(x_k, s) \leq \epsilon,$$

where $\kappa_{\mathbf{qr}}$ depends on $h(r(x_0)) - h(r)_{\mathbf{low}}$, Lipschitz constants $L_h$ and $L_A$, and other parameters. □

- similar result for trust-region algorithm.

# Evaluation complexity of non-smooth composite min..

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad h(r(x))$$

Key ingredients:

- Sufficient function decrease on successful iterations:

$$\begin{aligned} h(r(x_k)) - h(r(x_{k+1})) &\geq \eta_1 [h(r(x_k)) - m_k(s_k)] \\ &\geq \frac{\eta_1}{2} \min \left\{ 1, \frac{\Psi(x_k)}{\sigma_k} \right\} \Psi(x_k). \end{aligned}$$

- Regularization weight $\sigma_k \leq \sigma_{\max}$ for all $k$ as step is successful: $h(r(x_k + s_k)) \leq l(x_k, s_k) + \frac{L_h L_A}{2} \|s_k\|^2$.

While $\Psi(x_k) > \epsilon$,

$$h(r(x_0)) - h(r)_{\text{low}} = \sum_{i=0}^{j} [h(r(x_k)) - h(r(x_{k+1}))] \geq \frac{\eta_1 \epsilon^2}{2\sigma_{\max}} k_\epsilon \quad \ldots$$

# Evaluation complexity of constrained optimization

Consider now the EC-NLO problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c(x) = 0,$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$ are smooth & nonconvex; $m \leq n$

# Evaluation complexity of constrained optimization

Consider now the EC-NLO problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c(x) = 0,$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$ are smooth & nonconvex; $m \leq n$

Evaluation complexity of generating an (approximate) first-order critical (i.e., KKT) point?

$$g(x_*) + J(x_*)^T y_* = 0 \ \& \ c(x_*) = 0$$

[does not include cost of solving the subproblem]

$\longrightarrow$ As far as we can go:

[computing second-order critical points of (NLO) is at least NP-hard; cf Vavasis at al.]

# Evaluation complexity of constrained optimization

Consider now the EC-NLO problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c(x) = 0,$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$ are smooth & nonconvex; $m \leq n$

Evaluation complexity of generating an (approximate) first-order critical (i.e., KKT) point?

$$g(x_*) + J(x_*)^T y_* = 0 \ \& \ c(x_*) = 0$$

[does not include cost of solving the subproblem]

$\longrightarrow$ As far as we can go:

[computing second-order critical points of (NLO) is at least NP-hard; cf Vavasis at al.]

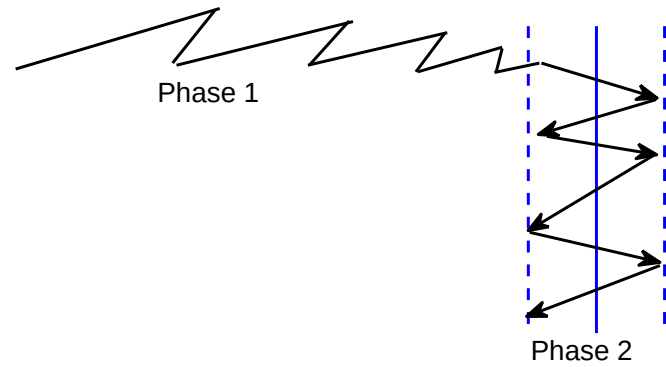
Methods with same worst-case evaluation complexity as in the unconstrained case: $\mathcal{O}(\epsilon^{-2})$ for first-order techniques. [sharp]

# A first-order algorithm for constrained problems

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c(x) = 0$$

Idea for a first-order algorithm:

- get feasible (if possible) by minimizing $\|c(x)\|$.
- track the trajectory




$$\mathcal{T}(t) = \{x \in \mathbb{R}^n : c(x) = 0 \text{ and } f(x) = t\},$$

for decreasing values of $t$ from some $t_0$ (corresponding to the first feasible iterate).

# A first-order algorithm for constrained problems...

A Short-Step Steepest-Descent (ShS - SD) algorithm:

Feasibility: apply non-smooth composite algorithm to

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \|c(x)\|$$

$\Longrightarrow$ at most $\mathcal{O}(\epsilon^{-2})$ function evaluations. $[h = \|\cdot\|, r(x) = c(x)]$

# A first-order algorithm for constrained problems...

A Short-Step Steepest-Descent (ShS - SD) algorithm:

Feasibility: apply non-smooth composite algorithm to

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad \|c(x)\|$$

$\Longrightarrow$ at most $\mathcal{O}(\epsilon^{-2})$ function evaluations. $[h = \|\cdot\|, r(x) = c(x)]$

Tracking/target-following: successively

- apply one (successful) step of non-smooth composite algorithm to $[r(x) = (c(x),\ f(x) - t)^T]$

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad \Phi(x,t) = \|c(x)\| + |f(x) - t| \qquad \Longrightarrow \quad x_+$$

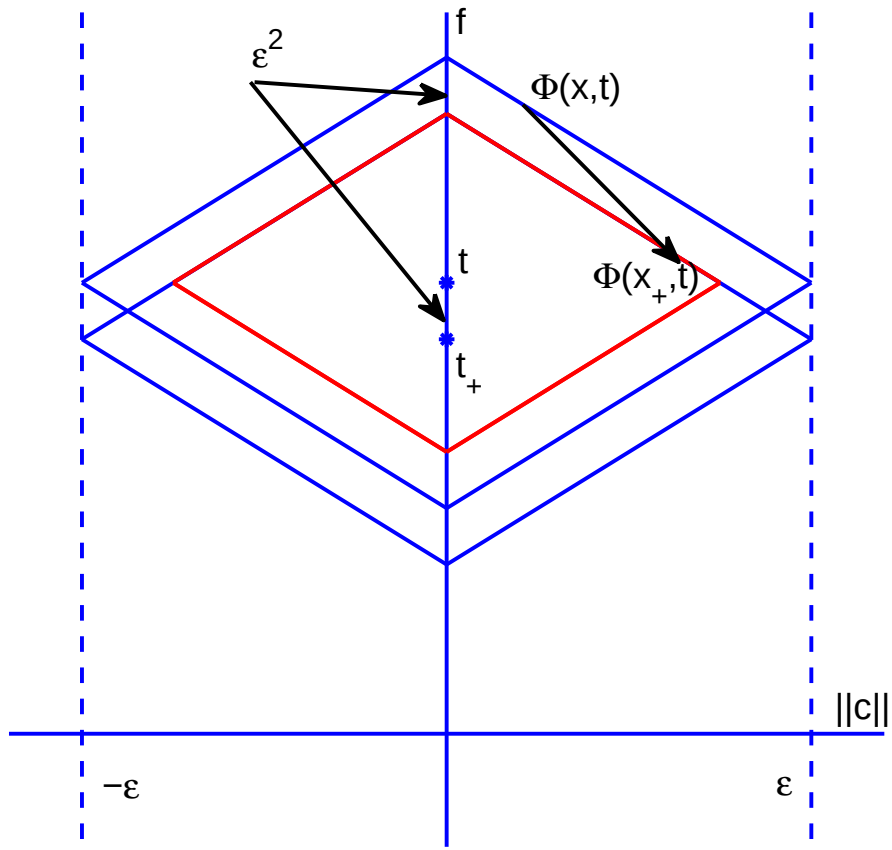
- decrease $t \to t_+$ to ensure $\Phi(x_+, t_+) = \epsilon$

$$\Longrightarrow t - t_+ \geq \Phi(x,t) - \Phi(x_+,t) \geq \kappa \cdot \epsilon^2$$

$\Longrightarrow$ at most $\mathcal{O}(\epsilon^{-2})$ problem evaluations.

# A first-order algorithm for constrained problems...

An iteration of the ShS - SD algorithm



# A complexity result for constrained problems

Assume that the objective's gradient $g$ and the constraints' Jacobian $J$ are globally Lipschitz continuous; $f$ bounded below and above in a neighbourhood of feasibility. Then the ShS - SD algorithm takes at most

$$\mathcal{O}\left(\epsilon^{-2}\right)^{(*)} \text{ problem evaluations}$$

to find an iterate $x_k$ with either

$$\|c(x_k)\| \leq \epsilon \text{ and } \|g(x_k) + J(x_k)^T y\| \leq \epsilon,$$

or

$$\|c(x_k)\| > \epsilon \text{ and } \|J(x_k)^T z\| \leq \epsilon,$$

for some $y$ and $z$.

$(*)$ same as evaluation complexity of unconstrained problems

# A complexity result for constrained problems

Assume that the objective's gradient $g$ and the constraints' Jacobian $J$ are globally Lipschitz continuous; $f$ bounded below and above in a neighbourhood of feasibility. Then the ShS - SD algorithm takes at most

$$\mathcal{O}\left(\epsilon^{-2}\right)^{(*)} \text{ problem evaluations}$$

to find an iterate $x_k$ with either

$$\|c(x_k)\| \leq \epsilon \text{ and } \|g(x_k) + J(x_k)^T y\| \leq \epsilon,$$

or

$$\|c(x_k)\| > \epsilon \text{ and } \|J(x_k)^T z\| \leq \epsilon,$$

for some $y$ and $z$.

$(*)$ same as evaluation complexity of unconstrained problems

- also applies to inequality-constrained problems: replace $\|c(x)\|$ by $\|\min(c(x), 0)\|$.

# Complexity of other methods for constrained pbs

A first-order exact penalty method for EC-NLO:

To generate an approximate (within $\epsilon$) KKT point or infeasible point of criticality measure, requires:

- $\mathcal{O}(\epsilon^{-2})$ problem-evaluations when the penalty parameter is bounded (independent of $\epsilon$)
- $\mathcal{O}(\epsilon^{-5})$ problem-evaluations, otherwise.

$\longrightarrow$ Apply non-smooth composite algorithm to minimizing penalty function

$$\Phi_\rho(x) = f(x) + \rho\|c(x)\|.$$

$\longrightarrow$ Update $\rho$ by 'steering' procedure (to control infeasibility)

- requires $f$ (so that $\Phi_\rho$) bounded below over entire $\mathbb{R}^n$ !

# Work-in-progress and conclusions

For non-smooth composite minimization:

- Approximate subproblem solution?

  $\Longrightarrow$ 'Cauchy points'? accuracy? (work in progress)

- Higher-order methods to find first-order critical points?

  $\Longrightarrow$ Higher-order models for the smooth part

- Numerical experiments ...

# Work-in-progress and conclusions

For non-smooth composite minimization:

- Approximate subproblem solution?

  $\Longrightarrow$ 'Cauchy points'? accuracy? (work in progress)

- Higher-order methods to find first-order critical points?

  $\Longrightarrow$ Higher-order models for the smooth part

- Numerical experiments ...

For the smooth constrained case:

- From a general worst-case complexity viewpoint, best to stay close to/onto the manifold of feasible points

  $\longrightarrow$ at variance with practical methods!

- Non-smooth formulations and techniques help us solve smooth problems more efficiently.

# Work-in-progress and conclusions ...

SELECTED REFERENCES:


- Yu. Nesterov, Gradient methods for minimizing composite objective function. CORE Discussion Paper 76, Université Catholique de Louvain, Belgium, 2007.

- Yu. Nesterov, Modified Gauss-Newton scheme with worst case guarantees for global performance. Optimization Methods and Software, 22:469–483, 2007.

- CGT, On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM Journal on Optimization 21(4):1721–1739, 2011.

- CGT, On the complexity of finding first-order critical points in constrained nonlinear programming. ERGO TR 11-005, School of Mathematics, University of Edinburgh, 2011.

- Bellavia, C, G, Morini and T, Convergence of a regularized Euclidean residual algorithm for nonlinear least-squares. SIAM Journal on Numerical Analysis 48(1):1–19, 2010.

- CGT, How much patience do you have? A worst-case perspective on smooth nonconvex optimization. OPTIMA 88, May 2012.