High-order optimality in nonlinear optimization: necessary conditions and a conceptual approach of evaluation complexity

Philippe Toint (with Coralia Cartis and Nick Gould)

Namur Center for Complex Systems (naXys), University of Namur, Belgium

(philippe.toint@fundp.ac.be)

Beijing, August 2016

Thanks

- Leverhulme Trust, UK
- Balliol College, Oxford
- Belgian Fund for Scientific Research (FNRS)
- University of Namur, Belgium
- ICNAAO 2016

The problem

We consider the convexly-constrained nonlinear programming problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ x \in \mathcal{F} & \end{array}$$

for $\mathcal{F}$ convex, non-empty, and $f : \mathbb{R}^n \to \mathbb{R}$ smooth.

Important special case: the (constrained) nonlinear least-squares problem

$$\text{minimize} \quad f(x) = \tfrac{1}{2}\|F(x)\|^2$$

for $x \in \mathbb{R}^n$ and $F : \mathbb{R}^n \to \mathbb{R}^m$ smooth.

High-order optimality?

Observation: Standard nonlinear optimization techniques stuck for more nonlinear problems

$\Rightarrow$ quadratic models too simple to capture strong nonlinear behaviour

$\Rightarrow$ use of higher-order polynomials (Taylor) models?

$\Rightarrow$ given high-order models, what about **high-order optimality???**

- What do we mean?
- Is it acheivable? At what cost?

Necessary optimality conditions: feasible arcs

Take into account:

- geometry of the feasible set
- potential decrease of the objective function

1) Geometry of the feasible set

Locally feasible arcs at x:

$$x(\alpha) = x + \alpha s_1 + \alpha^2 s_2 + \cdots + \alpha^q s_q + o(\alpha^q) \stackrel{\text{def}}{=} x + s(\alpha)$$

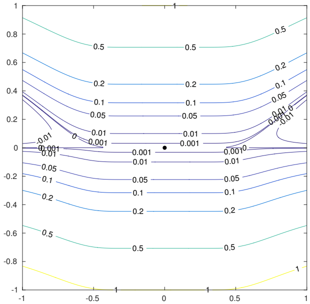
must be feasible for small enough $\alpha > 0$
(contraint qualification)

Necessary optimality conditions: objective decrease (1)

2) Decrease of the objective function (along feasible arcs)

- Some cases hopeless when using derivatives/Taylor series (Hancock)

$$\min_{x\in\mathbf{R}^2} f(x) = \begin{cases} x_2\left(x_2 - e^{-1/x_1^2}\right) & \text{if } x_1 \neq 0, \\ x_2^2 & \text{if } x_1 = 0, \end{cases}$$

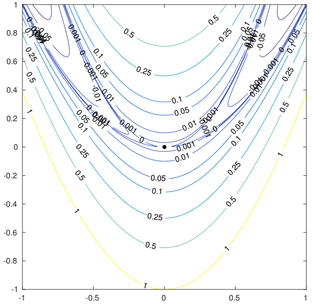


Necessary optimality conditions: objective decrease (2)

- Conditions along lines/subspaces not adequate!

Peano's example:

$$\min_{x \in \mathbf{R}^2} f(x) = x_2^2 - 3x_1^2 x_2 + 2x_1^4,$$



Local saddle point is minimum along every straight line!

Necessary optimality conditions (1)

Define the q-th order taylor series

$$T_{f,q}(x,s) = \sum_{j=0}^{q} \frac{1}{j!} \nabla_c^j f(x)[s]^j$$

A technical theorem stating necessary conditions (in words)

Suppose x is a local minimum of the convexly-constrained problem. Then, for every $q > 0$,

$$T_{f,q}(x, s(\alpha)) \geq 0$$

for all locally feasible $s(\alpha)$ such that

$$T_{f,j}(x, s(\alpha)) = 0 \quad j \in \{1, \ldots, q-1\}.$$

Define x to be q-th order critical

Necessary optimality conditions (2)

Note: $T_{f,j}(x, s(\alpha))$ is a polynomial in α with

coefficients depending on $s_1, \ldots, s_q$

(geometry of the feasible set)

k-th coeff for $T_{f,j}(x, s(\alpha))$:

$$c_{k,j}(x) = \frac{1}{k!}\left(\sum_{(\ell_1,\ldots,\ell_k)\in\mathcal{P}(j,k)} \nabla_x^k f(x_*)[s_{\ell_1}, \ldots, s_{\ell_k}] \right)$$

($\mathcal{P}(j,k)$ is a suitable set of multi-indices of size growing with j)

Verification essentially hopeless because of

- dependence of $c_{k,j}(x)$ on $s_{\ell_1}, \ldots, s_{\ell_k}$
- growing number of coefficients
- involves more than $\nabla_x^q f$ for $q \geq 4$!

Necessary optimality conditions: an alternative

Consider using the Taylor's models themselves!

$$\phi_{f,j}^{\Delta}(x) \stackrel{\rm def}{=} f(x) - \mathop{\rm globmin}_{\substack{x+d\in\mathcal{F}\\ \|d\|\leq\Delta}} T_{f,j}(x,d),$$

Serious drawback: global minimization in small neighbourhood of x

But in the unconstrained case, for any $\Delta > 0$,

$$\phi_{f,1}^{\Delta}(x) = \|\nabla_x^1 f(x)\|$$

and, if $\phi_{f,1}^{\Delta}(x) = 0$,

$$\phi_{f,2}^{\Delta}(x) = \left|\min\left[0, \lambda_{\min}(\nabla_x^2 f(x))\right]\right|$$

Ensuring (approximate) necessary conditions

Suppose that

$$\lim_{\Delta \to 0} \frac{\phi_{f,j}^{\Delta}(x)}{\Delta^j} = 0 \quad \text{for} \quad j \in \{1, \ldots, q\}$$

then x is a q-th order critical point

Approximated by

x is a q-th order ϵ-approximate critical point iff, for $\epsilon > 0$ and $\Delta > 0$ small,

$$\phi_{f,j}^{\Delta}(x) \leq \epsilon \Delta^j \quad \text{for} \quad j \in \{1, \ldots, q\}.$$

Minimizing property of q-th order ϵ-approximate critical points

Suppose that x is a q-th order ϵ-approximate critical point and that $\nabla_x^q f$ is Lipschitz continous (in tensor norm) with constant $L_{f,q}$. Then

$$f(x+d) \geq f(x) - 2\epsilon\Delta^q$$

for all $x + d \in \mathcal{F}$ such that

$$\|d\| \leq \min \left(\frac{p!\,\epsilon\Delta^q}{L_{f,p}} \right)^{\frac{1}{q+1}}.$$

(f cannot decrease much in a neighbourhood whose size increase with the order q $\Rightarrow$ stronger than simple effect of Lipschitz continuity)

An algorithmic approach to complexity

- Makes sense to search for x such that

$$\phi^{\Delta}_{f,j}(x) \leq \epsilon \Delta^j \quad \text{for} \quad j \in \{1, \ldots, q\}.$$

- Once $\phi^{\Delta}_{f,j}(x)$ is computed, exploit d_ϕ the argument of the global min!
- Imbed in a standard trust-region algorithm

A simple trust-region algorithm

A trust-region algorithm.

Step 0: Initialization. Given: $q > 1$, $\epsilon \in (0,1]$, x_0, $\Delta_1 \in [\epsilon, 1]$ as well as $\Delta_{\max} \in [\Delta_1, 1]$, $\gamma_1 \le \gamma_2 < 1 \le \gamma_3$ and $0 < \eta_1 \le \eta_2 < 1$. Compute $x_1 = P_{\mathcal{F}}[x_0]$, evaluate $f(x_1)$ and set $k = 1$.

Step 1: Step computation. For $j = 1, \ldots, q$, (i) evaluate $\nabla^j f(x_k)$ and $\phi_{f,j}^{\Delta_k}(x_k)$ (ii) if $\phi_{f,j}^{\Delta_k}(x_k) > \epsilon \Delta_k^j$, go to Step 3 with $s_k = d_\phi$,

Step 2: Termination. Terminate with $x_\epsilon = x_k$ and $\Delta_\epsilon = \Delta_k$.

Step 3: Accept the new iterate. Compute $f(x_k + s_k)$ and

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

If $\rho_k \ge \eta_1$, set $x_{k+1} = x_k + s_k$. Otherwise set $x_{k+1} = x_k$.

Step 4: Update the trust-region radius. Set

$$\Delta_{k+1} \in \begin{cases} [\gamma_1 \Delta_k, \gamma_2 \Delta_k] & \text{if } \rho_k < \eta_1, \\ [\gamma_2 \Delta_k, \Delta_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\Delta_k, \min(\Delta_{\max}, \gamma_3 \Delta_k)] & \text{if } \rho_k \ge \eta_2, \end{cases}$$

increment k by one and go to Step 1.

Evaluation complexity (1)

- No evaluation of f or derivative in the computation of $\phi_{f,j}^{\Delta_k}(x_k)$!
- Evaluation complexity can be evaluated:

> Suppose that $\nabla_x^j f$ is Lipschitz continous (in tensor norm) for $j \in \{1, \ldots, q\}$. Then the TR algorithm above needs at most
>
> $$O(\epsilon^{-(q+1)})$$
>
> evaluations of f and its first q derivative tensors to find a q-th order ϵ-approximate critical point

- But also

> This bound is essentially sharp

$$(\forall \delta > 0 \; \exists f(x) \; \forall \epsilon \text{ TR algo needs } O(\epsilon^{-\frac{q+1}{1+(q+1)\delta}}) \text{ evals})$$

> First theoretical result for arbitrary optimality order!

Evaluation complexity (2)

- In general: a conceptual algorithm!
- globmin effort limited by choosing $\Delta_{\max}$ not too large
- Maybe semi-realistic if derivative tensors are small an structured
- At all iterations, $\Delta_k \geq \kappa\epsilon$. Allows $\Delta_k \searrow 0$ when $\epsilon \searrow 0$

Complexity of convexly-constrained problems

Where do we stand?

$\vdots$	–	–	–	–		?
q	–	–	–	$O(\epsilon^{-(q+1)})$	?	?
$\vdots$	–	–		?	?	?
2		$O(\epsilon^{-3})$	?	?	?	?
1	$O(\epsilon^{-2})$	$O(\epsilon^{-3/2})$	$\dots$	$\dots$	$O(\epsilon^{-(p+1)/p})$	$\dots$
$\uparrow q/p \rightarrow$	1	2	$\dots$	$\dots$	p	$\dots$

Complexity of optimality order q as a function of model degree p

Trust-region algo

Regularization algo (BGMST)

A special case: fist-order optimality for nonlinear least-squares

Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) = \tfrac{1}{2}\|r(x)\|^2 \\ x \in \mathcal{F} & \end{array}$$

- Apply an $O(\epsilon^{-\pi})$ method for convex constraints ($\pi = 2$ or $\pi = (p+1)/p$)
- New termination test;

$$\|r(x)\| \le \epsilon_{\rm P} \quad \text{OR} \quad \phi^{\Delta}_{\|r\|,1}(x) \le \epsilon_{\rm D}\Delta^j$$

(zero residual vs. nonzero residual)

$$\text{Evaluation complexity} = O\left(\epsilon_{\rm P}^{1-\pi}\epsilon_{\rm D}^{-\pi}\right)$$

TR algo $\Rightarrow O(\epsilon_{\rm P}^{-1}\epsilon_{\rm D}^{-2})$ Reg algo $\Rightarrow O(\epsilon_{\rm P}^{-1/p}\epsilon_{\rm D}^{-(p+1)/p})$

The equality-constrained case

Consider now the EC-NLO (general with slack variables formulation):

$$\begin{array}{ll} \text{minimize}_x & f(x) \\ \text{such that} & c(x) = 0 \quad \text{and} \quad x \in \mathcal{F} \end{array}$$

Suppose x is a local minimum of the EC-NLO problem. Then, for every $q > 0$ and $\Lambda(x,y) = f(x) + y^T c(x)$,

$$T_{\Lambda,q}(x, s(\alpha)) \geq 0$$

for all locally feasible $s(\alpha)$ such that

$$T_{\Lambda,j}(x, s(\alpha)) = 0 \quad j \in \{1, \ldots, q-1\}$$

and

$$T_{c,j}(x, s(\alpha)) = 0 \quad j \in \{1, \ldots, q\}$$

$\Rightarrow$ even more complicated to handle!

Necessary conditions for EC-NLO

Verification essentially (even more) hopeless because of

- dependence of $c_{k,j}(x)$ on $s_{\ell_1}, \dots, s_{\ell_k}$
- growing number of coefficients
- involves more than $\nabla_x^q f$ for $q \geq 3$!

Ideas for a first-order algorithm:

1. get $\|c(x)\| \leq \epsilon$ (if possible) by minimizing $\|c(x)\|^2$ such that $x \in \mathcal{F}$ (getting $\|J(x)^T c(x)\|$ small unsuitable!)
2. track the "trajectory"

$$\mathcal{T}(t) \stackrel{\text{def}}{=} \{x \in \mathbf{R}^n \mid c(x) = 0 \quad \text{and} \quad f(x) = t\}$$

for values of t decreasing from f(first feasible iterate) while preserving $x \in \mathcal{F}$

First-order complexity for EC-NLO

Sketch of a two-phases algorithm:

feasibility: apply a $O(\epsilon^{-\pi})$ method for convex constraints (with specific termination test) to

$$\min_x \nu(x) \stackrel{\text{def}}{=} \|c(x)\|^2 \quad \text{such that} \quad x \in \mathcal{F}$$

> at most $O(\max[\epsilon_P^{-1}, \epsilon_P^{1-\pi}\epsilon_D^{-\pi}])$ evaluations

tracking: successively

- apply a $O(\epsilon^{-\pi})$ method for convex constraints (with specific termination test) to

$$\min_x \mu(x,t) \stackrel{\text{def}}{=} \|c(x)\|^2 + (f(x) - t)^2 \quad \text{such that} \quad x \in \mathcal{F}$$

- decrease t (proportionally to the decrease in $\phi(x)$)

> at most $O(\max[\epsilon_P^{-1}, \epsilon_P^{1-\pi}\epsilon_D^{-\pi}])$ evaluations

First-order complexity for EC-NLO

Under the "conditions stated above", the above algorithm takes at most

$$''O''(\epsilon_{\rm P}^{1-\pi}\epsilon_{\rm D}^{-\pi}) \text{ evaluations}$$

to find an iterate x_k with either

$$\|c(x_k)\| \le \delta\epsilon_P \quad \text{and} \quad \phi^{\Delta}_{\Lambda,1} \le \|(y,1)\|\epsilon_D\Delta$$

for some Lagrange multiplier y, or

$$\|c(x_k)\| > \delta\epsilon \quad \text{and} \quad \phi^{\Delta}_{||c||,1} \le \epsilon\Delta.$$

Higher order complexity for EC-NLO? (1)

The above approach for $q = 1$ hinges on

$$\nabla_x^1 \Lambda(x,y) = \frac{1}{f(x) - t} \nabla_x^1 \mu(x,t)$$

Hopeful for $q = 2$ since

$$\nabla_x^2 \Lambda(x,y)[d]^2 = \frac{1}{f(x) - t} \nabla_x^2 \mu(x,t)[d]^2$$

for all

$$d \in \text{span}\left\{\nabla_x^1 f(x)\right\}^\perp \cap \text{span}\left\{\nabla_x^1 c(x)\right\}^\perp \stackrel{\text{def}}{=} \mathcal{M}(x)$$

More difficult but maybe not imposible for $q = 3$ as

$$\nabla_x^3 \Lambda(x,y)[d]^3 = \frac{1}{f(x) - t} \nabla_x^3 \mu(x,t)[d]^3$$

for all

$$d \in \mathcal{M}(x) \cap [\text{a complicated set depending } \{\nabla_x^1 f\}, \{\nabla_x^2 f\}, \{\nabla_x^1 c\}, \{\nabla_x^2 c_i\}]$$

Higher order complexity for EC-NLO? (2)

But impossible for $q = 4$ (and above) because

$$\begin{aligned}\nabla_x^4 \Lambda(x,y) \;=\; & \frac{1}{f(x)-t}\nabla_x^4 \mu(x,t) \\ & -4\left[\nabla_x^3 f(x)\otimes \nabla_x^1 f(x) + \textstyle\sum_{i=1}^m \nabla_x^3 c_i(x)\otimes \nabla_x^1 c_i(x)\right] \\ & -3\left[\nabla_x^2 f(x)\otimes \nabla_x^2 f(x) + \textstyle\sum_{i=1}^m \nabla_x^2 c_i(x)\otimes \nabla_x^2 c_i(x)\right]\end{aligned}$$

A possibly important consequence:

Every approach based on quadratic (or more general strictly increasing) penalization is probably doomed for $q \geq 4$!

$\Rightarrow$ Need for a completely fresh point of view!

Conclusions

- Complexity analysis for general q-th order critical points

$$O(\epsilon^{-(q+1)}) \;\; \text{(unconstrained, convex constraints)}$$

- Complexity analysis for fisrt-order critical points

$$O(\epsilon_{\rm P}^{1-\pi}\epsilon_{\rm D}^{-\pi}) \;\; \text{(equality and general constraints)}$$

- Jarre's example $\Rightarrow$ global optimization much harder
- Many questions remaining:
  - high-order optimality with high-degree model?
  - beyond first-order for EC-NLO?

Many thanks for your attention. . .