Tensor methods for nonconvex optimization problems

Coralia Cartis (University of Oxford)

Joint with Nick Gould (Rutherford Appleton Laboratory, UK)
and Philippe Toint (University of Namur, Belgium)

One World Optimization Seminar Series, Online July 27, 2020

Standard methods for nonconvex optimization

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{where} \quad f \quad \text{is smooth.}$$

• f has gradient vector ∇f (first derivatives) and Hessian matrix $\nabla^2 f$ (second derivatives).

$\longrightarrow$ local minimizer x_* with $\nabla f(x_*) = 0$ (stationarity) and $\nabla^2 f(x_*) \succ 0$ (local convexity).

Derivative-based methods:

- user-given $x_0 \in \mathbb{R}^n$, generate iterates x_k, $k \geq 0$.
- $f(x_k + s) \approx \; m_k(s)$ simple model of f at x_k;
  m_k linear or quadratic Taylor approximation of f.
  $s_k \to \min_s \; m_k(s)$; $s_k \to x_{k+1} - x_k$
- terminate within ϵ of optimality (small gradient values).

Derivative-based local models

Choices of models

- linear : $m_k(s) = f(x_k) + \nabla f(x_k)^T s$

  $\longrightarrow$ s_k steepest descent direction.
- quadratic : $m_k(s) = f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s$

  $\longrightarrow$ s_k Newton-like direction.

Must safeguard s_k to ensure method converges globally, from an arbitrary starting point x_0, to first/second order critical points.

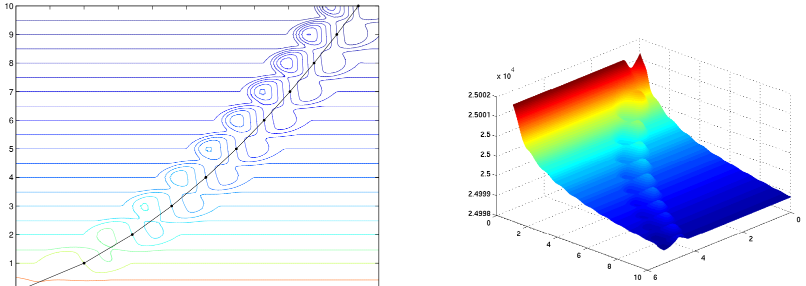
Adaptive 'globalization' strategies:

- Linesearch (Cauchy?, Armijo (1966))
- Trust region (Fletcher, Powell (1970s))
- Regularization (Levenberg-Marquardt ('44, '63), Griewank ('83), Nesterov & Polyak ('06), C, Gould & Toint ('11), ...)

Global efficiency of Newton's method

Newton's method: as slow as steepest descent

- may require $\lceil \epsilon^{-2} \rceil$ evaluations/iterations, same as steepest descent method



Globally Lipschitz continuous gradient and Hessian

But Regularized Newton (ie, ARC) has better/optimal complexity.

Worst-case evaluation complexity of methods

Global rates of convergence from any initial guess

Under sufficient smoothness assumptions on derivatives of f (Lipschitz continuity), for any $(\epsilon_1, \epsilon_2) > 0$, the algorithms generate $\|\nabla f(x_k)\| \leq \epsilon_1$ (and $\lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_2$) in at most $k_\epsilon^{\mathrm{alg}}$ iterations/evaluations:

1st, 2nd Criticality	SD	Newton/TR/LS	ARC	TR+/ LS+
$\|\nabla f(x_k)\|_2 \leq \epsilon_1$	$\mathcal{O}(\epsilon_1^{-2})$	$\mathcal{O}(\epsilon_1^{-2})$	$\mathcal{O}\left(\epsilon_1^{-\frac{3}{2}}\right)$	$\mathcal{O}\left(\epsilon_1^{-\frac{3}{2}}\right)$
$\lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_2$	–	$\mathcal{O}(\epsilon_2^{-3})$	$\mathcal{O}(\epsilon_2^{-3})$	$\mathcal{O}(\epsilon_2^{-3})$

[TR+:Curtis et al,'17]

[LS+:Royer et al'18]

- $\mathcal{O}(\cdot)$ contains $f(x_0) - f_{\mathrm{low}}$, L_{grad} or L_{Hessian} and algorithm parameters.
- all bounds are sharp, ARC bound is optimal for second-order methods [C, Gould & Toint,'10,'11, '17; Carmon et al ('18)]

Adaptive cubic regularization: ARC (=AR2)

[Griewank ('81, TR); Nesterov & Polyak ('06); Weiser et al ('07); C, Gould & Toint ('11)]

[Dussault ('15); Birgin et al ('17)]

- cubic regularization model at x_k

$$m_k(s) = \underbrace{f(x_k) + \nabla f(x_k)[s] + \tfrac{1}{2}\nabla f^2(x_k)[s]^2}_{T_2(x_k,s)} + \tfrac{1}{3}\sigma_k \|s\|_2^3$$

where $\sigma_k > 0$ is a regularization weight. [$B_k \approx \nabla f^2(x_k)$ allowed]

- compute s_k : $m_k(s_k) < f(x_k)$, $\|\nabla_s m_k(s_k)\| \leq \theta_1 \|s_k\|_2^2$ and $\lambda_{\min}(\nabla_s^2 m_k(s_k)) \geq -\theta_2 \|s_k\|_2^1$ [no global model minimization required, but possible]
- compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - T_2(x_k, s_k)}$
- set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > \eta = 0.1 \\ x_k & \text{otherwise} \end{cases}$
- $\sigma_{k+1} = \dfrac{\sigma_k}{\gamma_1} = 2\sigma_k$ when $\rho_k < \eta$; else $\sigma_{k+1} = \max\{\gamma_2 \sigma_k, \sigma_{\min}\} = \max\{\tfrac{1}{2}\sigma_k, \sigma_{\min}\}$

Regularization methods with higher derivatives

Adaptive pth order regularization: ARp

[Birgin et al ('17), C, Gould, Toint('20)]

ARp proceeds similarly to ARC/AR2:

- pth order regularization model at x_k

$$m_k(s) = \underbrace{f(x_k) + \nabla f(x_k)[s] + \ldots + \frac{1}{p!}\nabla^p f(x^k)[s]^p}_{T_p(x_k,s)} + \frac{1}{(p+1)!}\sigma_k\|s\|_2^{p+1}$$

  where $\sigma_k > 0$ is a regularization weight.

- compute s_k : $m_k(s_k) < f(x_k)$, $\|\nabla_s m_k(s_k)\| \leq \theta_1\|s_k\|_2^p$ and $\lambda_{\min}(\nabla_s^2 m_k(s_k)) \geq -\theta_2\|s_k\|^{p-1}$ [no global model minimization required]
- compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - T_p(x_k, s_k)}$
- set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > \eta = 0.1 \\ x_k & \text{otherwise} \end{cases}$
- $\sigma_{k+1} = \dfrac{\sigma_k}{\gamma_1} = 2\sigma_k$ when $\rho_k < \eta$; else $\sigma_{k+1} = \max\{\gamma_2\sigma_k, \sigma_{\min}\} = \max\{\frac{1}{2}\sigma_k, \sigma_{\min}\}$

Worst-case complexity of ARp for 1st/2nd-order criticality

[Birgin et al ('17), C, Gould, Toint('20)]

Theorem: Let $p \geq 2$, $f \in C^p(\mathbb{R}^n)$, bounded below by $f_{\rm low}$ and with the pth derivative Lipschitz continuous. Then ARp requires at most

$$\left\lceil \kappa_{1,2} \cdot (f(x_0) - f_{\rm low}) \cdot \max\left[\epsilon_1^{-\frac{p+1}{p}}, \epsilon_2^{-\frac{p+1}{p-1}}\right] + \kappa_{1,2} \right\rceil$$

function and derivatives' evaluations/iterations to ensure $\|\nabla f(x_k)\| \leq \epsilon_1$ and $\lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_2$.

1st, 2nd Criticality	p=2	p=3	p=4	...p
$\|\nabla f(x_k)\|_2 \leq \epsilon_1$	$\mathcal{O}(\epsilon_1^{-3/2})$	$\mathcal{O}(\epsilon_1^{-4/3})$	$\mathcal{O}\left(\epsilon_1^{-5/4}\right)$	$\mathcal{O}\left(\epsilon_1^{-(p+1)/p}\right)$
$\lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_2$	$\mathcal{O}(\epsilon_2^{-3})$	$\mathcal{O}(\epsilon_2^{-2})$	$\mathcal{O}(\epsilon_2^{-5/3})$	$\mathcal{O}(\epsilon_2^{-(p+1)/(p-1)})$

All bounds are sharp, and ARp 1st-order bound is optimal for pth order mthds.

[C, Gould & Toint,'20 Carmon et al ('18)]

Worst-case complexity of ARp for 1st/2nd-order criticality

Sketch of Proof (Theorem): [Birgin et al ('17), C, Gould, Toint('20)]

- Sufficient decrease on successful steps

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \eta[f(x_k) - T_p(x_k, s_k)] \\ &= f(x_k) - m_k(s_k) + \tfrac{\sigma_k}{(p+1)!}\|s_k\|^{p+1} \\ &\geq \tfrac{\sigma_{\min}}{(p+1)!}\|s_k\|^{p+1} \\ &\geq c\min\{\epsilon_1^{(p+1)/p}, \epsilon_2^{(p+1)/(p-1)}\} \qquad (*) \end{aligned}$$

- Long steps: first-order

$$\|s_k\| \geq c_1 \left(\frac{\nabla f(x_k + s_k)}{L + \theta_1 + \sigma_{\max}}\right)^{1/p} \geq c_1 \epsilon_1^{1/p}$$

and second-order

$$\|s_k\| \geq c_2 \left(\frac{\lambda_{\min}(\nabla^2 f(x_k + s_k))}{L + \theta_2 + \sigma_{\max}}\right)^{1/(p-1)} \geq c_2 \epsilon_2^{1/(p-1)}$$

where $\sigma_k \leq \sigma_{\max} = C \cdot L$. Summing up (*) over successful iterations + counting unsuccessful iterations.

ARp for 3rd-order criticality

In the model minimization, require also the 3rd order approximate condition:

$$\max_{d \in \mathcal{M}_{k+1}} \left| \nabla_s^3 m_k(s_k)[d]^3 \right| \leq \|s_k\|^{p-2},$$

whenever

$$\mathcal{M}_{k+1} = \left\{ d \mid \|d\| = 1 \text{ and } |\nabla_s^2 m_k(s_k)[d]^2| \leq \|s_k\|^{p-1} \right\} \neq \emptyset.$$

Then under same conditions as Theorem, ARp takes at most

$$\left\lceil \kappa_{1,2,3} \cdot (f(x_0) - f_{\rm low}) \cdot \max\left[\epsilon_1^{-\frac{p+1}{p}}, \epsilon_2^{-\frac{p+1}{p-1}}, \epsilon_3^{-\frac{p+1}{p-2}} \right] + \kappa_{1,2,3} \right\rceil$$

function and derivatives' evaluations/iterations to ensure

$$\|\nabla f(x_k)\| \leq \epsilon_1, \; \lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_2$$

and $\left| \nabla^3 f(x_k)[d]^3 \right| \leq \epsilon_3$, $|\nabla^2 f(x_k)[d]^2| \leq \epsilon_2$, for all $d \in \mathcal{M}_k$.

- $\mathcal{M}_k$ includes approximate objective's Hessian null space if subproblem is solved to local ϵ accuracy.

Regularization methods for high order optimality

Beyond 3rd order: high(er)-order optimality conditions

[C, Gould, Toint('18, J FoCM)]

Let x_* be a local minimizer of $f \in C^q(\mathbb{R}^n)$. Consider (feasible) descent arcs $x(\alpha) = x_* + \sum_{i=1}^q \alpha^i s_i + o(\alpha^q)$ where $\alpha > 0$. Derive necessary (and sometimes sufficient) optimality conditions.

[Hancock, Peano example of non-Taylor based arcs along which descent happens!]

For $j \in \{1, \ldots, q\}$, the inequality

$$\sum_{k=1}^{j} \frac{1}{k!} \left(\sum_{(\ell_1,\ldots,\ell_k)\in\mathcal{P}(j,k)} \nabla_x^k f(x_*)[s_{\ell_1}, \ldots, s_{\ell_k}] \right) \geq 0$$

holds for all $(s_1, \ldots, s_j)$ such that, for $i \in \{1, \ldots, j-1\}$,

$$\sum_{k=1}^{i} \frac{1}{k!} \left(\sum_{(\ell_1,\ldots,\ell_k)\in\mathcal{P}(i,k)} \nabla_x^k f(x_*)[s_{\ell_1}, \ldots, s_{\ell_k}] \right) = 0,$$

where the index sets $\mathcal{P}(j,k) = \{(\ell_1, \ldots, \ell_k) \in \{1, \ldots, j\}^k \mid \sum_{i=1}^k \ell_i = j\}$.

Beyond 3rd order: high(er)-order optimality conditions

[C, Gould, Toint('18, J FoCM)]

- Convex constraints (and suitable constraint qualifications) can be incorporated.
- Usual first, second and third order optimality conditions can be derived.
- But, starting at fourth-order and beyond, necessary conditions above involve a mixture of derivatives of different orders and cannot/should not be separated/disentangled.

Example: Peano variant: $\min_{x\in\mathbb{R}^2} f(x) = x_2^2 - \kappa_1 x_1^2 x_2 + \kappa_2 x_1^4$,

where κ_1 and κ_2 are specified parameters.

Fourth-order condition (κ_1 large):

$\ker^1[\nabla_x^1 f(0)] = \Re^2$, $\ker^2[\nabla_x^2 f(0)] = e_1$, $\ker^3[\nabla_x^3 f(0)] = e_1 \cup e_2$.

$$\tfrac{1}{2}\nabla_x^2 f(0)[s_2]^2 + \tfrac{1}{2}\nabla_x^3 f(0)[s_1, s_1, s_2] + \tfrac{1}{24}\nabla_x^4 f(0)[s_1]^4 \geq 0$$

implies the much weaker $\nabla_x^4 f(x_*)[s_1]^4 \geq 0$ on $\cap_{i=1}^3 \ker^i[\nabla_x^i f(x_*)]$.

Beyond 3rd order: high(er)-order optimality conditions

[C, Gould, Toint('20, arXiv)]

Challenge: find a (necessary) optimality measure for qth order criticality for f that is sufficiently accurate and useful in ARp ?
For $j \in \{1, \ldots, q\}$, a jth order criticality measure for f is: for some $\delta \in (0, 1]$, let

$$\phi^{\delta}_{f,j}(x) = f(x) - \text{globmin}_{\|d\| \leq \delta} T_j(x, d).$$

$\longrightarrow$ a robust notion of criticality.

- $\phi^{\delta}_{f,j}(x)$ is continuous in x and δ for all orders q.
- $\phi^{\delta}_{f,1}(x) = \|\nabla f(x)\| \delta$
- $\phi^{\delta}_{f,2}(x) = \max\{0, -\lambda_{\min}(\nabla^2 f(x))\} \delta^2$.

If x is a local minimizer of f, then for $j \in \{1, \ldots, q\}$,

$$\lim_{\delta \to 0} \frac{\phi^{\delta}_{f,j}(x)}{\delta^j} = 0,$$

and this limit also implies the involved necessary conditions before.

ARqp: a high order regularization and criticality framework

[C, Gould, Toint('20, arXiv)]

- Let $q \leq p$. The pth order regularization model at x_k

$$m_k(s) = T_p(x_k, s) + \frac{1}{(p+1)!}\sigma_k \|s\|_2^{p+1}.$$

- compute (s_k, δ_s): $m_k(s_k) < f(x_k)$,

$$\phi_{m_k,j}^{\delta_s}(s_k) \leq \theta \epsilon_j \delta_s^j, \quad j \in \{1, \ldots, q\}.$$

- compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - T_p(x_k, s_k)}$
- set $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \delta_s$ if $\rho_k > \eta = 0.1$; else $x_{k+1} = x_k$ and $\delta_{k+1} = \delta_k$.
- $\sigma_{k+1} = \dfrac{\sigma_k}{\gamma_1} = 2\sigma_k$ when $\rho_k < \eta$; else $\sigma_{k+1} = \max\{\gamma_2 \sigma_k, \sigma_{\min}\} = \max\{\frac{1}{2}\sigma_k, \sigma_{\min}\}$

ARqp: a high order regularization and criticality framework

[C, Gould, Toint('20, arXiv)]

Theorem: Let $p \geq q \geq 1$, $f \in C^p(\mathbb{R}^n)$, bounded below by $f_{\rm low}$ and with derivatives $\nabla^j f$ Lipschitz continuous for $j \in \{1, \ldots, p\}$. Terminate ARqp when

$$\phi_{f,j}^{\delta_k}(x_k) \leq \epsilon_j \delta_k^j \quad \text{for all } j \in \{1, \ldots, q\}$$

for some δ_k that is either 1 ($q = 1, 2$) or at least $C\epsilon = C(\epsilon_i)_{i=\overline{1,q}}$ [achievable for ARqp]. Until termination, ARqp requires at most

- $q = 1, 2$: $\quad \left\lceil \kappa_{1,2} \cdot (f(x_0) - f_{\rm low}) \cdot \max_{j=\overline{1,q}} \epsilon_j^{-\frac{p+1}{p-j+1}} + \kappa_{1,2} \right\rceil$

  [same as ARp]
- $q > 2$: $\quad \left\lceil \kappa_q \cdot (f(x_0) - f_{\rm low}) \cdot \max_{j=\overline{1,q}} \epsilon_j^{-\frac{q(p+1)}{p}} + \kappa_q \right\rceil$

function and derivatives' evaluations/iterations.

All bounds are sharp [C, Gould, Toint,'20]

ARqp: a high order regularization and criticality framework

[C, Gould, Toint('20, arXiv)]

Sketch of Proof (Theorem): Same ingredients as for ARp complexity proof:

Sufficient decrease on successful steps

$$f(x_k) - f(x_{k+1}) \geq \frac{\sigma_{\min}}{(p+1)!} \|s_k\|^{p+1}$$

Long steps: much more challenging when $q > 2$!

$$\|s_k\| \geq c_q \left(\frac{1-\theta}{L+\sigma_{\max}} \right)^{1/p} \epsilon_j^{j/p}$$

for some $j \in \{1, \ldots, q\}$, where $\sigma_k \leq \sigma_{\max} = C \cdot L$.

Lower bound on s_k: $(1-\theta)\epsilon_j \delta_k^j \leq (L+\sigma_{\max}) \sum_{l=1}^{j} \delta_k^l \|s_k\|^{p-l+1}$

Summing up (*) over successful iterations + counting unsuccessful iterations.

Higher order methods

A few remarks...

- ARqp with weaker optimality condition: $\phi_{f,j}^{\delta_k} \leq \epsilon_j \delta_k$, $j = \overline{1,q}$, satisfies complexity bound $\mathcal{O}\left(\max_{j=\overline{1,q}} \epsilon_j^{-\frac{p+1}{p-j+1}}\right)$.
- TRq (Trust-region detecting qth order criticality) satisfies the weaker complexity bound: $\mathcal{O}(\max_{j=\overline{1,q}} \epsilon_j^{-(q+1)})$.
- Convex constraints can be incorporated into ARp and ARqp without affecting the complexity.

Universal regularization methods

Universal ARp for first order criticality

[C, Gould, Toint ('19)]

Universal ARp (U-ARp) employs regularized local models

$$m_k(s) = T_p(x_k, s) + \frac{\sigma_k}{r}\|s\|_2^r,$$

where $r > p \geq 1$, r real, and $T_p(x_k, s)$ as in ARp.

U-ARp proceeds similarly to ARp:

- compute s_k: $m_k(s_k) < f(x_k)$ and $\|\nabla_s m_k(s_k)\| \leq \theta \|s_k\|^{r-1}$
- $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - T_p(x_k, s_k)}$
- update σ_k

But U-ARp has an additional crucial ingredient: if $\rho_k \geq \eta$ [i.e., k successful], check whether

$$\sigma_k \|s_k\|^{r-1} \geq \alpha \epsilon_1 \qquad (*)$$

where $\alpha \in (0, \frac{1}{3}]$ is a user-chosen constant.

U-ARp allows $x_{k+1} = x_k + s_k$ (and σ_k decrease) only when both $\rho_k \geq \eta$ and (*) hold. Else, σ_k is increased.

Beyond Lipschitz continuity, towards non-smoothness

$f \in C^{p,\beta_p}(\mathbb{R}^n)$: $f \in C^p(\mathbb{R}^n)$ and $\nabla^p f$ is Hölder continuous on the path of the iterates (and trial points), namely,

$$\|\nabla^p f(y) - \nabla^p f(x_k)\| \leq L\|y - x_k\|^{\beta_p}$$

holds for all $y \in [x_k, x_k + s_k]$, $k \geq 0$.
$L_p > 0$ and $\beta_p \in [0, 1]$ for any $p \geq 1$.

- $\beta_p = 0$: $\nabla^p f$ uniformly bounded.
- $\beta_p \in (0, 1)$: $\nabla^p f$ continuous but not differentiable.
- $\beta_p = 1$: $\nabla^p f$ Lipschitz continuous (and differentiable).
- $\beta_p > 1$: f reduces to polynomials.

$\longrightarrow$ Hölder continuity : a bridging case between smooth and non-smooth problems [Nemirovskii & Yudin ('83), Nesterov ('13), Devolder ('13), Grapiglia & Nesterov ('16)]

Worst-case complexity of UARp

Let $r \geq p \geq 1$, r real and p integer.
Let $f \in C^{p,\beta_p}(\mathbb{R}^n)$.
If $r \geq p + \beta_p$ [e.g., $r = p + 1$], then U-ARp requires at most

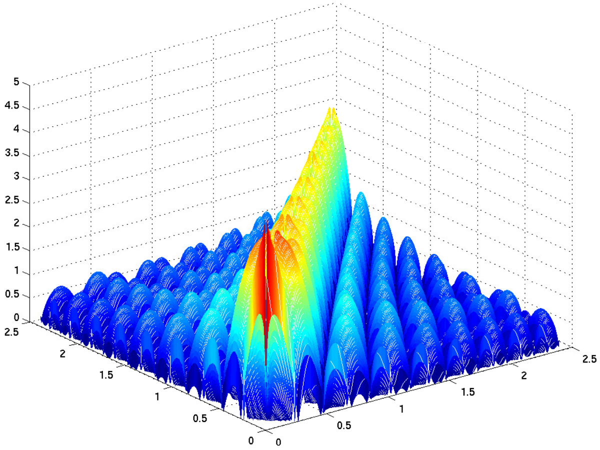
$$\left\lceil \kappa_1 \cdot (f(x_0) - f_{\text{low}}) \cdot \epsilon_1^{-\frac{p+\beta_p}{p+\beta_p-1}} \right\rceil$$

function/derivative evaluations and iterations to ensure $\|\nabla f(x_k)\| \leq \epsilon_1$.

$r \geq p + \beta_p$ [e.g., $r = p + 1$]: the bound is 'universal', adapting to landscape smoothness without knowing β_p/smoothness of f, independent of r.

Smooth or nonsmooth?

Sharpness example: the ragged landscape of a $f \in C^{1,\beta_1}$



Ratio of $|\nabla f(x) - \nabla f(y)|/|x-y|^{\beta}$