Regularization methods with Nonsmooth Norms

Philippe Toint with S. Gratton (and S. Jerad)

University of Namur - naXys
philippe.toint@unamur.be

PolyU seminar, Hong Kong, December 2021

The problem (again)

We consider the unconstrained nonlinear programming problem:

$$\text{minimize} \quad f(x)$$

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

For now, focus on the

unconstrained case

We consider algorithms

- that use derivatives for the step computation
- rely on function evaluations for the step size control

Main idea behind AR algorithms (Griewank)

Suppose f is p times continuously differentiable. We define

$$T_{f,p}(x,s) \stackrel{\rm def}{=} f(x) + \sum_{\ell=1}^{p} \frac{1}{\ell!} \nabla_x^\ell f(x)[s]^\ell.$$

By standard calculus, $f(x+s) =$

$$T_{f,p}(x,s) + \frac{1}{(p-1)!} \int_0^1 (1-\xi)^{p-1} \left(\nabla_x^p f(x+\xi s) - \nabla_x^p f(x)\right) [s]^p \, d\xi.$$

If $\nabla_x^p f(x)$ is is globally Lipschitz continuous in the $\|\cdot\|_r$ norm,

$$\|\nabla_x^p f(x) - \nabla_x^p f(y)\|_{r,p} \le L_{f,p} \|x-y\|_r \quad \text{for all} \quad x, y \in \mathbb{R}^n.$$

$$f(x+s) \le \underbrace{T_{f,p}(x,s) + \frac{L_{f,p}}{(p+1)!} \|s\|_r^{p+1}}_{m(s)}$$

$\Longrightarrow$ reducing m from $s=0$ improves $f(x)$

Approximate model minimization (I)

Lipschitz constant is not known: use an adaptative model

$$m_k(s) \stackrel{\rm def}{=} T_{f,p}(x_k,s) + \frac{\sigma_k}{(p+1)!}\|s\|_2^{p+1}$$

1. In the case of a norm which is differentiable (except at the origin), we usually require strict decrease and approximate first order stationarity

$$\| \ \nabla_s^1 T_{f,p}(x_k,s_k) + \frac{\sigma_k}{p!}\|s_k\|_r^p \nabla_s^1 \|s_k\|_2 \ \|_2 \leq \kappa_{\rm stop}\|s_k\|_2^p$$

2. This condition is satisfied in the neighborhood of regularized model critical points
3. Fro example, Krylov subspace methods can be used to provide steps
4. The regularization parameter σ_k is adapted using the predicted versus achieved decrease principle

Algorithm 0.1: First-Order Adaptive Regularization

Step 0: Initialization: $x_0 \in \mathbb{R}^n$, σ_0 and set $k = 0$

Step 1: Termination if $\|\nabla_x^1 f(x_k)\|_{r,1} \le \epsilon_1$.

Step 2: Step calculation: Compute a step s_k such that

$m_k(s_k) \le m_k(0)$ and

$$\| \; \nabla_s^1 T_{f,p}(x_k, s_k) + \frac{\sigma_k}{p!} \|s_k\|_2^p \, \nabla_s^1 \|s_k\|_2 \; \|_2 \le \kappa_{\rm stop} \|s_k\|_2^p$$

Step 3: Step acceptance: Compute $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_{f,p}(x_k, 0) - T_{f,p}(x_k, s_k)}$, and set

$$x_{k+1} = \begin{cases} x_k + s_k & \text{if } \; \rho_k \ge \eta_2 \\ x_k & \text{otherwise} \end{cases}$$

Step 4: Regularization parameter update.

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

How many function evaluations and iterations to reach $\|\nabla f(x_k)\|_2 \leq \epsilon_1$?

On has the upper-bound : $\sigma_k \leq \sigma_{\max} \stackrel{\rm def}{=} \gamma_3 \max\left[\sigma_0, \frac{L_{f,p}}{(1-\eta_2)}\right]$.

Simple model decrease yields:
$\Delta T_{f,p}(x_k, s_k) \stackrel{\rm def}{=} T_{f,p}(x_k, 0) - T_{f,p}(x_k, s) \geq \frac{\sigma_k}{(p+1)!}\|s_k\|_2^{p+1}$.
Assume now that $\sigma_k \geq L_{f,p}/(1-\eta_2)$. From

$$|\rho_k - 1| \leq \frac{(p+1)!|f(x_k+s_k) - T_{f,p}(x_k, s_k)|}{\sigma_k\|s_k\|_2^{p+1}} \leq \frac{L_{f,p}}{\sigma_k}$$

we get $\rho_k \geq \eta_2$ and thus $\sigma_{k+1} \leq \sigma_k$.

Convergence to first order critical points (III)

At a successsful step k before termination, $\|s_k{}^p\|_2 \geq \frac{p!}{L_{f,p}+\kappa_{\rm stop}\,p!+\sigma_{\max}}\,\epsilon_1.$

$$
\begin{aligned}
\epsilon_1 &< \|\nabla_x^1 f(x_{k+1})\|_2 \\
&\leq \|\nabla_x^1 f(x_{k+1}) - \nabla_x^1 T_{f,p}(x_k, s_k)\|_2 + \\
&\quad \|\nabla_x^1 T_{f,p}(x_k, s_k) + \tfrac{\sigma_k}{p!}\|s_k\|_2^{p-1}\, s_k\|_2 + \frac{\sigma_k}{p!}\|s_k\|_2^p \\
&\leq \frac{L_{f,p}}{p!}\|s_k\|_2^p + \kappa_{\rm stop}\|s_k\|_2^p + \frac{\sigma_k}{p!}\|s_k\|_2^p.
\end{aligned}
$$

Convergence to first order critical points (IV)

$\|\nabla f(x_k)\|_2 \le \varepsilon_1$ in at most $\frac{f(x_0)-f}{\kappa}\,\epsilon_1^{-\frac{p+1}{p}}$ successsful iterations.
This yields the $O(\epsilon_1^{-\frac{p+1}{p}})$ complexity result.

$$
\begin{aligned}
f(x_0) - f_{low} &\ge f(x_0) - f(x_{k+1}) \ge \textstyle\sum_{i\in\mathcal{S}_k} f(x_i) - f(x_{i+1}) \\
&\ge \eta_1 \textstyle\sum_{i\in\mathcal{S}_k} \frac{\sigma_k}{(p+1)!}\|s_k\|_r^{p+1} \\
&\ge |\mathcal{S}_k|\frac{\eta_1\sigma_{\min}}{(p+1)!}\left(\frac{L_{f,p}+\kappa_{\rm stop}p!+\sigma_{\max}}{p!}\right)^{\frac{p+1}{p}} \epsilon_1^{\frac{p+1}{p}}
\end{aligned}
$$

Moreover,

$$
k \le |\mathcal{S}_k|\left(1+\frac{|\log\gamma_1|}{\log\gamma_2}\right) + \frac{1}{\log\gamma_2}\log\left(\frac{\sigma_{\max}}{\sigma_0}\right).
$$

Some references

C. Cartis, N. Gould and Ph. L. Toint,
"Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints", SIOPT, vol. 30(1), pp. 513-541, 2020

C. Cartis and N. I. M. Gould and Ph. L. Toint,
"Strong Evaluation Complexity Bounds for Arbitrary-Order Optimization of Nonconvex Nonsmooth Composite Functions", arXiv:2001.10802, 2020.

S. Bellavia, G. Gurioli, B. Morini and Ph. L. Toint,
"Deterministic and stochastic inexact regularization algorithms for nonconvex optimization with optimal complexity", SIOPT, vol. 29(4), pp. 2881-2915, 2019.

S. Bellavia, G. Gurioli, B. Morini and Ph. L. Toint,
"High-order Evaluation Complexity of a Stochastic Adaptive Regularization Algorithm for Nonconvex Optimization Using Inexact Function Evaluations and Randomly Perturbed Derivatives", arXiv:2005.04639, 2020.

C. Cartis, N. Gould and Ph. L. Toint,
"Second-order optimality and beyond: characterization and evaluation complexity in convexly-constrained nonlinear optimization", FoCM, vol. 18(5), pp. 1083-1107, 2018.

Also see http://perso.fundp.ac.be/~phtoint/toint.html

However...

Importantly the abstract space $\mathbb{R}^n$ is often multivariate in practical applications. The vector $x \in \mathbb{R}^n$ may contain

- quantities of different physical nature (temperature, pressure)
- quantites of different mathematical nature (a Gaussian is represented by mean and covariance)
- use of norm equivalence induces oversolving (Minimum surface and Sobolev norm)

The problem may be a discretization of a continuous problem.

Using appropriate norms $\| \bullet \|_r$ in our algorithms is a must

$\Rightarrow$ Adapt regularization algorithms to this requirement

One may then wonder. . .

Is it possible to derive regularization algorithms in which the regularization term involves a nondifferentiable norm to get first- or second-order critical points???

What do we mean by critical points of order larger than 1 ???

Can we design a regularization method that would be adapted to infinite dimension problems ???

Not an obvious question!

Approximate model minimization (I)

Lipschitz constant is not known: use an adaptive model

$$m_k(s) \stackrel{\text{def}}{=} T_{f,p}(x_k, s) + \frac{\sigma_k}{(p+1)!}\|s\|_r^{p+1}$$

1. For a differentiable norm (except at the origin), we usually require strict decrease and approximate first order stationarity

$$\left\| \; \nabla_s^1 T_{f,p}(x_k, s_k) + \frac{\sigma_k}{p!}\|s_k\|_r^p \nabla_s^1 \|s_k\|_r \; \right\|_{r,1} \leq \kappa_{\text{stop}} \|s_k\|_r^p$$

2. For nonsmooth norms, Clarke stationarity writes

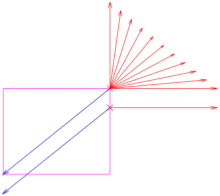
$$\exists \xi \in \mathbf{R}^n, \; \nabla_s^1 T_{f,p}(x_k, s_k^*) = -\frac{\sigma_k}{p!}\|s_k^*\|_r^p \; \xi$$

3. The quantity ξ is a dual vector of s_k^*:

$${s_k^*}^\top \xi = \|s_k^*\|_r \|\xi\|_{r,1} \text{ and } \|\xi\|_{r,1} = 1$$

Approximate model minimization (II)

Regularized quadratic $\frac{1}{2}s^\top s + s^\top g + \frac{10}{6}\|s\|_\infty^3$



Stationnarity:

- $s + g = -5\|s\|_\infty^2 \xi$, where
- $s^\top \xi = \|s\|_\infty \|\xi\|_1$ and $\|\xi\|_1 = 1$

$\Longrightarrow$ it seems reasonnable to request for some $\theta_1 \geq 1$

$$\|\nabla_s^1 T_{f,p}(x_k, s_k)\|_{r,1} \leq \theta_1 \frac{\sigma_k}{p!} \|s\|_r^p$$

The algorithm

Algorithm 0.2: First-Order Adaptive Regularization with General Norm (AR1pGN)

Step 0: Initialization: $x_0 \in \mathbb{R}^n$, σ_0, $\theta_1 > 1$ and set $k = 0$.

Step 1: Termination if $\|\nabla_x^1 f(x_k)\|_{r,1} \le \epsilon_1$.

Step 2: Step calculation: Compute a step s_k such that

$$m_k(s_k) \le m_k(0) \text{ and } \|\nabla_s^1 T_{f,p}(x_k, s_k)\|_{r,1} \le \theta_1 \frac{\sigma_k}{p!}\|s_k\|_r^p$$

Step 3: Step acceptance: Compute $\rho_k = \frac{f(x_k) - f(x_k+s_k)}{T_{f,p}(x_k,0) - T_{f,p}(x_k,s_k)}$, and set

$$x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k \ge \eta_2 \\ x_k & \text{otherwise} \end{cases}$$

Step 4: Regularization parameter update.

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

Step 5: Next iteration Increment k by one and go to Step 1

Convergence to first order critical points (I)

How many function evaluations and iterations to reach $\|\nabla f(x_k)\|_{r,1} \le \epsilon_1$?

On has the upper-bound : $\sigma_k \le \sigma_{\max} \stackrel{\rm def}{=} \gamma_3 \max\left[\sigma_0, \frac{L_{f,p}}{(1-\eta_2)}\right]$.

Simple model decrease yields:

$\Delta T_{f,p}(x_k, s_k) \stackrel{\rm def}{=} T_{f,p}(x_k, 0) - T_{f,p}(x_k, s) \ge \frac{\sigma_k}{(p+1)!}\|s_k\|_r^{p+1}$.

Assume now that $\sigma_k \ge L_{f,p}/(1-\eta_2)$. From

$$|\rho_k - 1| \le \frac{(p+1)!|f(x_k+s_k) - T_{f,p}(x_k, s_k)|}{\sigma_k\|s_k\|_r^{p+1}} \le \frac{L_{f,p}}{\sigma_k}$$

we get $\rho_k \ge \eta_2$ and thus $\sigma_{k+1} \le \sigma_k$.

At a successsful step k before termination, $\|s_k\|_r^p \geq \frac{p!}{L_{f,p}+\theta_1\sigma_{\max}}\,\epsilon_1$.

$$
\begin{aligned}
\epsilon_1 &< \|\nabla_x^1 f(x_{k+1})\|_{r,1} \\
&\leq \|\nabla_x^1 f(x_{k+1}) - \nabla_x^1 T_{f,p}(x_k, s_k)\|_{r,1} + \|\nabla_x^1 T_{f,p}(x_k, s_k)\|_{r,1} \\
&\leq \frac{L_{f,p}}{p!}\|s_k\|_r^p + \theta_1 \frac{\sigma_k}{p!}\|s_k\|_r^p.
\end{aligned}
$$

SIMPLER!!

$\|\nabla f(x_k)\|_{r,1} \leq \varepsilon$ in at most $\frac{f(x_0)-f}{\kappa}\,\epsilon_1^{-\frac{p+1}{p}}$ successsful iterations.
This yields the $O(\epsilon_1^{-\frac{p+1}{p}})$ complexity result.

$$
\begin{aligned}
f(x_0) - f_{low} &\geq f(x_0) - f(x_{k+1}) \geq \sum_{i \in \mathcal{S}_k} f(x_i) - f(x_{i+1}) \\
&\geq \eta_1 \sum_{i \in \mathcal{S}_k} \frac{\sigma_k}{(p+1)!} \|s_k\|_r^{p+1} \\
&\geq |\mathcal{S}_k| \frac{\eta_1 \sigma_{\min}}{(p+1)!} \left(\frac{L_{f,p} + \theta_1 \sigma_{\max}}{p!} \right)^{\frac{p+1}{p}} \epsilon_1^{\frac{p+1}{p}}
\end{aligned}
$$

The model is not (2nd order) differentiable, but..

1. Let $p = 2$ and $\phi(s) = f_0 + \langle g, s\rangle + \frac{1}{2}\langle Hs, s\rangle$ be a quadratic polynomial in $s \in \mathbb{R}^n$.
2. Let s_* be a global minimizer of $m(s) = \phi(s) + \frac{1}{6}\sigma\|s\|_r^3$ and define $\lambda_r[H] = \min_{v\neq 0} \frac{\langle Hv,v\rangle}{\|v\|_r^2}$.
3. Let λ_a and u_a satisfy $\|u_a\|_r = 1, \quad \lambda_a \stackrel{\rm def}{=} \langle Hu_a, u_a\rangle$

$$\text{and} \quad \Big(\lambda_a \in \Big[\lambda_r[H], \tau\lambda_r[H]\Big] \quad \text{if} \quad \lambda_r[H] < 0\Big),$$

.

A descent lemma

Let $m(s) = f_0 + \langle g, s\rangle + \frac{1}{2}\langle Hs, s\rangle + \frac{1}{6}\sigma\|s\|_r^3$ Consider $s \neq 0$ and let λ_a and u_a.

1. Choose the sign of u_a to ensure that $\langle g + Hs, u_a\rangle \leq 0$ and assume that $\lambda_a + \sigma\|s\|_r < 0$.
2. Then there exists an $\alpha > 0$ such that

$$m(s) - m(s + \alpha\|s\|_r u_a) \geq \frac{3(\lambda_a + \sigma\|s\|_r)}{4\sigma^2}\left[\psi(s)\sigma^2\|s\|_r^2 - \frac{3}{4}(\lambda_a + \sigma\|s\|_r)^2\right],$$

where

$$\psi(s) \stackrel{\rm def}{=} \max\left[0, 1 + 2\frac{\langle g + Hs, u_a\rangle}{\sigma\|s\|_r^2}\right].$$

The model is not (2nd order) differentiable, but..

If s_* is a minimizer of the regularized polynomial,

$$\lambda_a + \omega(s_*)\sigma\|s_*\|_r \geq 0,$$

where

$$\omega(s) \stackrel{\text{def}}{=} \begin{cases} 1 + \frac{2\sqrt{\psi(s)}}{\sqrt{3}} \leq 1 + \frac{2}{\sqrt{3}} \stackrel{\text{def}}{=} \kappa_\omega & \text{if } s \neq 0, \\ 1 & \text{otherwise,} \end{cases}.$$

Suppose $\lambda_a + \sigma\|s_*\|_r < 0$. The descent lemma shows that

$$\psi(s_*)\sigma^2\|s_*\|_r^2 > \frac{3}{4}(\lambda_a + \sigma\|s_*\|_r)^2.$$

which implies

$$\sqrt{\psi(s_*)}\sigma\|s_*\|_r > \sqrt{\frac{3}{4}}\left|\lambda_a + \sigma\|s_*\|_r\right| > -\frac{\sqrt{3}}{2}\Big(\lambda_a + \sigma\|s_*\|_r\Big).$$

Algorithm 0.3: First-Order Adaptive Regularization with General Norm (AR1pGN)

Step 0: Initialization: $x_0 \in \mathbb{R}^n$, σ_0 and set $k = 0$

Step 1: Termination Compute $\lambda_{a,k}$ and $u_{a,k}$ approx. curvature of H_k.

$$\text{Terminate if } \|g_k\|_{r,1} \leq \epsilon_1 \quad \text{and} \quad \lambda_{a,k} \geq -\tau\epsilon_2.$$

Step 2: Step calculation: Compute a step s_k such that

$$m_k(s_k) \leq m_k(0),\ \|\nabla_s^1 T_{f,p}(x_k, s_k)\|_{r,1} \leq \theta_1 \frac{\sigma_k}{p!}\|s\|_r^p, \text{ and}$$

$$\lambda_{a,k} + \theta_2 \omega_*(s_k)\sigma_k \|s_k\|_r \geq 0.$$

Step 3: Step acceptance: Compute $\rho_k = \frac{f(x_k) - f(x_k+s_k)}{T_{f,p}(x_k,0) - T_{f,p}(x_k,s_k)}$, and set

$$x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k \geq \eta_2 \\ x_k & \text{otherwise} \end{cases}$$

Step 4: Regularization parameter update. As before

Convergence to "second-order" critical points for p=2 (I)

Let $k \in \mathcal{S}$. Suppose $\lambda_{a,k+1} < -\tau\epsilon_2$. Then

$$\|s_k\|_r \geq \frac{\tau}{L_{f,2} + \tau^{-1}\theta_2\kappa_\omega\sigma_{\max}}\,\epsilon_2.$$

$$\begin{aligned}
\lambda_{a,k+1} \geq \lambda_r[H_{k+1}] &= \min_{\|d\|_r=1} [\langle H_{k+1}d, d\rangle - \langle H_k d, d\rangle + \langle H_k d, d\rangle] \\
&\geq \min_{\|d\|_r=1} [\langle H_{k+1}d, d\rangle - \langle H_k d, d\rangle] + \min_{\|d\|_r=1} \langle H_k d, d\rangle \\
&\geq -\|H_{k+1} - H_k\|_{r,2} + \lambda_r[H_k] \\
&\geq -\|H_{k+1} - H_k\|_{r,2} + \tau^{-1}\lambda_{a,k} \\
&\geq -\left(L_{f,2}\|s_k\|_r + \tau^{-1}\theta_2\omega_*(s_k)\sigma_k\|s_k\|_r\right)
\end{aligned}$$

Convergence to "second" order critical points (p=2)

Then the AR2GN algorithm requires at most

$$\left(\frac{\kappa_{\text{AR2GN}}}{\eta_1\sigma_{\min}}\right)\frac{f(x_0)-f_{\text{low}}}{\min\left[\epsilon_1^{3/2},\epsilon_2^3\right]} \text{ iterations}$$

to reach a point where

$$\|g(x_\epsilon)\|_{r,1} \le \epsilon_1 \quad \text{and} \quad \lambda_r[H(x_\epsilon)] \ge -\epsilon_2.$$

At a successful step, k before termination,
$\|s_k\|_r^2 \ge \frac{2!}{L_{f,2}+\theta_1\sigma_{\max}}\,\epsilon_1$ and $\|s_k\|_r \ge \frac{\tau}{L_{f,2}+\tau^{-1}\theta_2\kappa_\omega\sigma_{\max}}\,\epsilon_2$, which together with the decrease $m_k(0)-m_k(s_k) \ge \frac{\sigma_k}{3!}\|s_k\|_r^3$ implies the result.

Summary

- Global convergence or AR algorithms for any “regularization” norm
- Relies on approximate minimum of the model:

$$m_k(s_k) \le m_k(0), \ \|\nabla_s^1 T_{f,p}(x_k, s_k)\|_{r,1} \le \theta_1 \frac{\sigma_k}{2!}\|s\|_r^2,$$

and

$$\lambda_{a,k} + \theta_2 \omega_*(s_k)\sigma_k\|s_k\|_r \ge 0.$$

How to compute such a step in practice?

Define $g_k = \nabla_s^1 T_{f,p}(x_k, s_k)$.

A descent algorithm for general norms

- We consider $m(s) = f_0 + \langle g, s\rangle + \frac{1}{2}\langle Hs, s\rangle + \frac{1}{6}\sigma\|s\|_r^3$,
- and want to minimize "enough" m to get s_k such that

$$m(s_k) \le m(0), \qquad \|g + Hs_k\|_{r,1} \le \theta_1 \frac{\sigma}{2!}\|s_k\|_r^2,$$

and

$$\lambda_{a,k} + \theta_2\omega_*(s_k)\sigma_k\|s_k\|_r \ge 0.$$

Remember

$$\|u_a\|_r = 1, \quad \lambda_a \stackrel{\text{def}}{=} \langle Hu_a, u_a\rangle \quad \text{and} \quad \left(\lambda_a \in \Big[\lambda_r[H], \tau\lambda_r[H]\Big] \quad \text{if} \quad \lambda_r[H] < 0\right)$$

Let $s_k \in \mathbb{R}^n$ such that $m(s_k) \leq m(0)$ and

$$\|g_k\|_{r,1} > \tfrac{1}{2}\sigma\|s_k\|_r^2.$$

Then

$$m(s_k) - m(s_k^C) \geq \tfrac{1}{2}\min\left[\frac{\left|\|g_k\|_{r,1} - \frac{1}{2}\sigma\|s_k\|_r^2\right|^2}{1 + \frac{3}{2}\left(\|H\|_{r,2} + \sigma\|s_k\|_r\right)}, \frac{\left|\|g_k\|_{r,1} - \frac{1}{2}\sigma\|s_k\|_r^2\right|^{\frac{3}{2}}}{16\sqrt{\sigma}}\right],$$

where $g_k = g + Hs_k$ and

$$s_k^C = s_k + \alpha_k^C d_k, \quad \text{with} \quad d_k = \arg\min_{\|v\|_r=1}\langle g_k, v\rangle \quad \text{and} \quad \alpha_k^C = \arg\min_{\alpha>0} m(s_k + \alpha d_k).$$

Step computation. Retraction step

Let $s_k \in \mathbb{R}^n$ such that $m(s_k) \leq m(0)$ and

$$\|g_k\|_{r,1} < \tfrac{1}{2}\sigma\|s_k\|_r^2.$$

Then

$$m(s_k) - m(s_k^R) \geq \tfrac{1}{2}\min\left[\frac{\left|\|g_k\|_{r,1} - \tfrac{1}{2}\sigma\|s_k\|_r^2\right|^2}{1+\tfrac{3}{2}\left(\|H\|_{r,2}+\sigma\|s_k\|_r\right)}, \frac{\left|\|g_k\|_{r,1} - \tfrac{1}{2}\sigma\|s_k\|_r^2\right|^{\frac{3}{2}}}{16\sqrt{\sigma}}\right],$$

where $g_k = g + Hs_k$ and

$$s_k^R = (1-\alpha_k^R)s_k \quad \text{with} \quad \alpha_k^R = \arg\min_{\alpha>0} m(s_k - \alpha s_k).$$

Suppose that $\lambda_a < 0$ where λ_a and u_a as before.
For $k \geq 0$, define $s_k^E = s_k + \alpha_k^E u_k$, where

$$u_k = -\mathrm{sign}\big(\langle g_k, u_a\rangle\big)\, u_a \;\text{ and }\; \alpha_k^E = \arg\min_{\alpha>0} m(s_k + \alpha u_k).$$

At any iteration $k \geq 1$ such that $\lambda_a + \omega_*(s_k)\sigma\|s_k\|_r < 0$, one has that

$$m(s_k) - m(s_k^E) \geq \frac{9\|s_k\|_r^2}{16\sigma^2}\left|\lambda_a + \sigma\omega(s_k)\|s_k\|_r\right|^3 > 0.$$

Remains to find bounds for $\|s_k\|_r$.

Suppose that, for some s_k and some $\beta \geq 0$, $m(s_0) - m(s_k) \geq \beta$.

$$\|s_k\|_r \leq \frac{\frac{1}{2}\|H\|_{r,2} + \sqrt{\|H\|_{r,2}^2 + \frac{2}{3}\sigma\|g\|_{r,1}}}{\frac{1}{3}\sigma} \stackrel{\rm def}{=} \kappa_{s,\rm upp},$$

$$\|s_k\|_r \geq \begin{cases} \dfrac{\sqrt{\|g\|_{r,1}^2 + 2\beta|\lambda_r[H]|} - \|g\|_{r,1}}{|\lambda_r[H]|} & \text{if } \lambda_r[H] < 0 \\ \dfrac{\beta}{\|g\|_{r,1}} & \text{otherwise.} \end{cases}$$

From

$$\tfrac{1}{6}\sigma\|s_k\|_r^3 \leq |\langle g, s_k\rangle| + \tfrac{1}{2}|\langle Hs_k, s_k\rangle| \leq \|g\|_{r,1}\|s_k\|_r + \tfrac{1}{2}\|H\|_{r,2}\|s_k\|_r^2,$$

and

$$-\|g\|_{r,1}\|s_k\|_r + \tfrac{1}{2}\min\left[0, \lambda_r[H]\right]\|s_k\|_r^2 \leq \langle g, s_k\rangle + \tfrac{1}{2}\langle Hs_k, s_k\rangle \leq m(s_k) - m(0)$$

Algorithm 0.4: An algorithm for minimization of a regularized quadratic (RQMIN)

Step 0: Initialization If unavailable, compute λ_a and u_a. Set $k = 0$, $s_0 = 0$ and $g_0 = g$.

Step 1: Check for termination. Terminate if

$$\left| \|g_k\|_{r,1} - \tfrac{1}{2}\sigma\|s_k\|_r^2 \right| \leq \epsilon_1 \quad \text{and} \quad \lambda_a + \omega(s_k)\sigma\|s_k\|_r \geq 0.$$

Step 2: Negative gradient step. If $\|g_k\|_{r,1} > \frac{1}{2}\sigma\|s_k\|_r^2$, compute s_k^C and $m_{k,1} = m(s_k^C)$.

Step 3: Retraction step. If $\|g_k\|_{r,1} < \frac{1}{2}\sigma\|s_k\|_r^2$, compute s_k^C and $m_{k,1} = m(s_k^C)$.

Step 4: Eigenvalue step. If $\lambda_a + \omega(s_k)\sigma\|s_k\|_r < -\epsilon_2\sigma\|s_k\|_r$, compute s_k^E and $m_{k,2} = m(s_k^E)$. Else, set $m_{k,2} = m(s_k)$.

Step 5: Move. Set $s_{k+1} = \begin{cases} s_k^C & \text{if } m_{k,1} \leq m_{k,2}, \\ s_k^E & \text{otherwise,} \end{cases}$ and

$$g_{k+1} = g_k + H(s_{k+1} - s_k).$$

Complexity of RQMIN.

Given $\theta_1 > 1$ and $\theta_2 > 1$, there exist a constant $\kappa_{\text{RQMIN1}} > 0$ independent of k such that the RQMIN algorithm requires at most

$$\kappa_{\text{RQMIN1}} \max\left[(\theta_1 - 1)^{-2}, (\theta_1 - 1)^{-\frac{3}{2}}, (\theta_2 - 1)^{-1}\right]$$

iterations to produce an iterate s_k such that

$$\begin{aligned} m_k(s_k) &\leq m_k(0) \\ \|\nabla_s^1 T_{f,p}(x_k, s_k)\|_{r,1} &\leq \theta_1 \frac{\sigma_k}{p!} \|s\|_r^p \\ \lambda_{a,k} &\geq -\theta_2 \omega_*(s_k) \sigma_k \|s_k\|_r \end{aligned}$$

Conclusions and perspectives

Summary:

- Optimizing function in their "natural" norm to avoid oversolving
- General norm adaptive regularization yields simple convergence theory, and similar complexity results as for the Euclidean norm
- First order and "eigen" algorithms can be defined to compute the step

Perspectives:

- Banach spaces
- Probabilistic error specification
- Numerical results

Thank your for your attention!

References

- S. Gratton, Ph. L. Toint,
  Adaptive Regularization Minimization Algorithms with Non-Smooth Norms and Euclidean Curvature,
  arXiv:2105.07765
- S. Gratton, S. Jerad, Ph. L. Toint,
  Holder Gradient Descent and Adaptive Regularization Methods in Banach Spaces for First-Order Points,
  arXiv:2104.02564
- S. Gratton, E. Simon, Ph. L. Toint,
  Minimization of nonsmooth nonconvex functions using inexact evaluations and its worst-case complexity,
  Mathematical Programming, vol. 87(1-2), pp. 1–24, 2020.