Multiresolution in image restoration

4- Globalization strategies

Elisa Riccietti, Nelly Pustelnik

Globalization strategies for smooth optimization

Given the problem

$$\min_{x \in \mathbb{R}^N} f(x)$$

with $f \in \mathcal{C}^1$, and a sequence

$$x^{[k+1]} = x^{[k]} + \gamma_k d^{[k]}$$

with $\gamma_k > 0$ and $d^{[k]}$ a descent direction.

In order to guarantee global convergence, i.e.,

$$(\forall x^{[0]} \in \mathbb{R}^N) \qquad \lim_{k \to \infty} \|\nabla f(x^{[k]})\| = 0$$

we need a globalization strategy to control the stepsize.

Globalization strategies

Two classes of globalization strategies:

- **Line-search**
- **Trust-region/adaptive regularization.**

Globalization strategies

Two classes of globalization strategies:

- **Line-search**

  $\rightarrow$ we first choose the step direction and then we determine its length.

- **Trust-region/adaptive regularization.**

Globalization strategies

Two classes of globalization strategies:

- **Line-search**

  $\rightarrow$ we first choose the step direction and then we determine its length.

- **Trust-region/adaptive regularization.**

  $\rightarrow$ we first choose the maximal length of the step (the trust-region radius or the regularization parameter) and then we determine the direction.

Outline

Line-search strategies

Trust-region strategies

Adaptive regularization

Line-search strategies

Given a descent direction $d^{[k]}$ (negative gradient, Newton's or quasi-Newton direction), choose the step-length γ_k in:

$$x^{[k+1]} = x^{[k]} + \gamma_k d^{[k]}.$$

Ideally, we should take γ_k such that

$$\boxed{\gamma_k = \arg\min_{\gamma \in \mathbb{R}^+} \varphi(\gamma) := f(x^{[k]} + \gamma d^{[k]})}$$

Need to solve a (scalar) optimization pb at each iteration: too expensive !

Alternative strategy

Want to find conditions on γ_k to ensure $\lim_{k\to\infty} \|\nabla f(x^{[k]})\| = 0$. Asking for the simple decrease of f is <u>not a sufficient condition</u> to get a convergent method.

Example 1 (Too long steps)

Let us consider $f(x) = x^2$, $x^{[0]} = 2$, $d^{[k]} = (-1)^{k+1}$ (these are descent directions), $\gamma_k = 2 + \frac{3}{2^{k+1}}$. The sequence $x^{[k]}$ is

$$x^{[k]} = (-1)^k \left(1 + \frac{1}{2^k}\right), \qquad k \in \mathbb{N}.$$

The simple decrease condition is satisfied because

$$f(x^{[k+1]}) = (x^{[k+1]})^2 = \left(1 + \frac{1}{2^{k+1}}\right)^2 < \left(1 + \frac{1}{2^k}\right)^2 = (x^{[k]})^2 = f(x^{[k]}),$$

but $x^{[k]}$ does not converge to the minimizer of $f(x)$, $x^* = 0$, because

$$x^{[2k]} = 1 + \frac{1}{2^{2k}} \xrightarrow[k \to +\infty]{} 1, \qquad x^{[2k+1]} = -\left(1 + \frac{1}{2^{2k+1}}\right) \xrightarrow[k \to +\infty]{} -1.$$

The simple decrease condition is not sufficient to guarantee a convergence of $\{x^{[k]}\}$ to $x^* = 0$ because we have chosen a sequence of γ_k with too large values.

Example 2 (Too short steps)

Let us consider $f(x) = x^2$, $x^{[0]} = 2$, $d^{[k]} = -1$ (these are descent directions), $\gamma_k = \frac{1}{2^{k+1}}$. The sequence $x^{[k]}$ is

$$x^{[k]} = 1 + \frac{1}{2^k}, \qquad k \in \mathbb{N}.$$

The simple decrease condition is satisfied as

$$f(x^{[k+1]}) = (x^{[k+1]})^2 = \left(1 + \frac{1}{2^{k+1}}\right)^2 < \left(1 + \frac{1}{2^k}\right)^2 = (x^{[k]})^2 = f(x^{[k]}),$$

but $x^{[k]}$ does not converge to $x^* = 0$ minimizer of $f(x)$ because

$$x^{[k]} = 1 + \frac{1}{2^k} \xrightarrow{k \to +\infty} 1.$$

The simple decrease condition is not sufficient to guarantee a convergence of $\{x^{[k]}\}$ to $x^* = 0$ because we have chosen a sequence of γ_k with too small values.

Armijo rule

Armijo rule (A) requires that

$$f(x^{[k]} + \gamma_k d^{[k]}) \leq f(x^{[k]}) + \gamma_k c_1 \nabla f(x^{[k]})^T d^{[k]}, \quad c_1 \in (0,1), \tag{A}$$

usually $c_1 = 10^{-4}$.

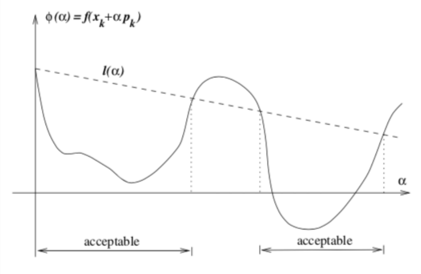
(A) is stronger than just asking the simple decrease $f(x^{[k+1]}) \leq f(x^{[k]})$ because $\nabla f(x^{[k]})^T d^{[k]} < 0$.

Armijo rule - geometrical meaning

Let

$$\varphi(\gamma) = f(x^{[k]} + \gamma d^{[k]}),$$
$$\ell(\gamma) = f(x^{[k]}) + \gamma c_1 \nabla f(x^{[k]})^T d^{[k]}.$$

Armijo rule requires $\varphi(\gamma) \leq \ell(\gamma)$.



Parameters γ that satisfy (A)

Wolfe rule: avoids small steps

Wolfe rule (W)

$$\nabla f(x^{[k]} + \gamma_k d^{[k]})^T d^{[k]} \geq c_2 \nabla f(x^{[k]})^T d^{[k]} \quad c_2 \in (c_1, 1). \tag{W}$$

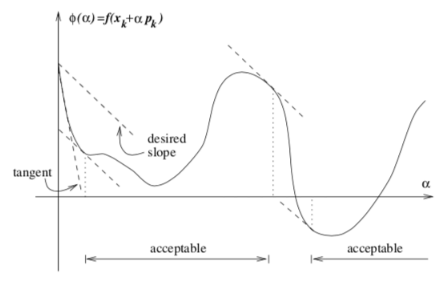
- The first term $\nabla f(x^{[k]} + \gamma_k d^{[k]})^T d^{[k]} = \varphi'(\gamma_k)$ is the slope of $\varphi(\gamma)$.
- The condition requires this slope to be greater than the negative slope $c_2 \nabla f(x^{[k]})^T d^{[k]}$ (desired slope).
- This ensures that the slope has been reduced sufficiently.

Wolfe rule - geometrical meaning

Because $c_1 < c_2 < 1$ and $\nabla f(x^{[k]})^T d^{[k]}$ is negative it holds

$$c_1 \nabla f(x^{[k]})^T d^{[k]} > c_2 \nabla f(x^{[k]})^T d^{[k]} > \nabla f(x^{[k]})^T d^{[k]},$$

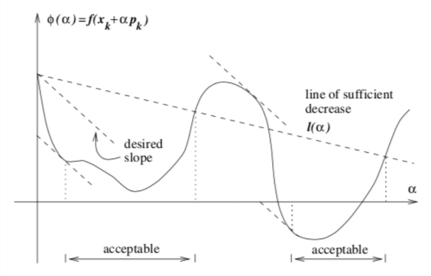
i.e., the desired slope is between those of $\ell(\gamma)$ and $\varphi(\gamma)$.



Parameters γ that satisfy (W)

Wolfe's Lemma

Let $f : \mathbb{R}^N \to \mathbb{R}$, $f \in \mathcal{C}^1$ and bounded below in $\{x^{[k]} + \gamma d^{[k]} \mid \gamma > 0\}$, with $d^{[k]}$ a descent direction for f in $x^{[k]}$, and let $c_1, c_2 : 0 < c_1 < c_2 < 1$.
It exists $I \subseteq (0, +\infty)$ non empty such that every $\gamma \in I$ satisfies (A) + (W).



Parameters γ that satisfy (A) e (W)

Convergence of line-search methods

Zoutendijk's theorem
Let $\Omega = \{x \in \mathbb{R}^N \mid f(x) \leq f(x^{[0]})\}$, $f \in \mathcal{C}^1(\Omega)$ and lower bounded on Ω, $d^{[k]}$ a descent direction for f, and assume that γ_k satisfies (A) and (W) and that $\nabla f(x)$ is Lipschitz continuous in Ω.
Let $\vartheta^{[k]}$ be the angle between $-\nabla f(x^{[k]})$ and $d^{[k]}$, i.e., the angle such that

$$\cos(\vartheta^{[k]}) = -\frac{\nabla f(x^{[k]})^T d^{[k]}}{\|\nabla f(x^{[k]})\| \|d^{[k]}\|}.$$

The numerical series

$$\sum_{j=0}^{+\infty} \cos^2(\vartheta^{[j]}) \|\nabla f(x^{[j]})\|^2$$

is convergent.

Global convergence

- The fact that the series $\sum_{j=0}^{+\infty} \cos^2(\vartheta_j)\|\nabla f(x_j)\|^2$ converges implies that

$$\lim_{k\to+\infty} \cos^2(\vartheta^{[k]})\|\nabla f(x^{[k]})\|^2 = 0.$$

If the sequence $\{\cos(\vartheta^{[k]})\}_k$ is bounded away from zero, we can deduce that

$$\boxed{\lim_{k\to+\infty} \|\nabla f(x^{[k]})\| = 0.}$$

- In this case every accumulation point of $\{x^{[k]}\}$ (if it exists) is a stationary point. Indeed, let $\widetilde{x}$ be an accumulation point of $\{x^{[k]}\}$, i.e., let $\widetilde{x}$ be the limit point of a subsequence $\{x_{k_j}\}$ of $\{x^{[k]}\}$. Then

$$\nabla f(\widetilde{x}) = \nabla f\left(\lim_{k_j\to+\infty} x^{[k_j]}\right) = \lim_{k_j\to+\infty} \nabla f(x^{[k_j]}) \underbrace{=}_{\lim_{k\to+\infty}\nabla f(x^{[k]})=0} 0.$$

Condition on the directions

- It can happen that $\lim_{k\to+\infty}\cos(\vartheta^{[k]}) = 0$.
- This happens if

$$\lim_{k\to+\infty} \nabla f(x^{[k]})^T d^{[k]} = 0,$$

  that is if $\nabla f(x^{[k]})$ and $d^{[k]}$ tend to be orthogonal. Formally $\nabla f(x^{[k]})^T d^{[k]}$ is negative, and $d^{[k]}$ remains a descent direction, but actually for k large $\nabla f(x^{[k]})^T d^{[k]}$ is close to 0, i.e., $d^{[k]}$ is a "poor" descent direction and along this direction the values of f are almost constant.
- This is a situation that can be avoided, choosing a descent direction $d^{[k]}$ such that $\cos(\vartheta^{[k]}) > M$ for some $M > 0$ and for all k.

Gradient method

With the steepest descent method, as $d^{[k]} = -\nabla f(x^{[k]})$, we have

$$\cos(\vartheta^{[k]}) = -\frac{\nabla f(x^{[k]})^T d^{[k]}}{\|\nabla f(x^{[k]})\| \|d^{[k]}\|} = \frac{\nabla f(x^{[k]})^T \nabla f(x^{[k]})}{\|\nabla f(x^{[k]})\| \|\nabla f(x^{[k]})\|} = 1.$$

Then, under the assumptions of Zoutendijk's Theorem, it holds $\lim_{k \to +\infty} \|\nabla f(x^{[k]})\| = 0$.

Line-search algorithm

0. Given $x^{[0]}, f, \varepsilon$
1. For $k = 0, 1, \ldots$
   1.1 Choose $d^{[k]}$ descent direction (may be the steepest descent direction, Newton's direction, quasi-Newton direction..)
   1.2 Find γ_k that satisfies (A) and (W)
   1.3 Set $x^{[k+1]} = x^{[k]} + \gamma_k d^{[k]}$
   1.4 If $\|\nabla f(x^{[k+1]})\| \leq \varepsilon$ return $x^{[k+1]}$ (approximation of x^*) and stop
2. Return "failure"

This algorithm is well-defined because the sequence $\nabla f(x^{[k]})$ converges to 0 for Zoutendijk's Theorem and because from Wolfe's Lemma it exists γ_k that satisfies (A) and (W)

How to actually compute such an γ_k?

The backtracking technique

0. Given $x^{[k]}, \overline{\gamma}, d^{[k]}, b_{\max}, c_1 \in (0,1), \alpha \in (0,1)$

1. $\gamma_k = \overline{\gamma}$
2. For $b = 0, 1, \ldots, b_{\max}$
    1. If $f(x^{[k]} + \gamma_k d^{[k]}) \leq f(x^{[k]}) + \gamma_k c_1 \nabla f(x^{[k]})^T d^{[k]}$, i.e. if γ_k satisfies (A), then set $\mathrm{IND} = 1$ and stop,
    otherwise set $\gamma_k = \alpha \gamma_k$
3. Set $\mathrm{IND} = -1$

- This can be used at each iteration of the line-search algorithm (at step 1.2).
- Each iteration of backtracking algorithm costs a function evaluation, so the algorithm costs at most $b_{\max}$ function evaluations. Usually $b_{\max} = 20$.
- Backtracking checks just (A) because too expensive to check (W).

Choice of $\overline{\gamma}$

Depends on $d^{[k]}$:

- Gradient direction: assume that the first-order change in the function at iteration k will be the same as that obtained at the previous iteration. We then impose $\overline{\gamma}\nabla f(x^{[k]})^T d^{[k]} = \gamma_{k-1}\nabla f(x^{[k-1]})^T d^{[k-1]}$ so that

$$\overline{\gamma} = \gamma_{k-1}\frac{\nabla f(x^{[k-1]})^T d^{[k-1]}}{\nabla f(x^{[k]})^T d^{[k]}}$$

- Barzilai-Borwein (BB) choice

$$\overline{\gamma} = \frac{(s^{[k-1]})^T s^{[k-1]}}{(s^{[k-1]})^T y^{[k-1]}}, \qquad s^{[k-1]} = x^{[k]} - x^{[k-1]}, \quad y^{[k-1]} = \nabla f(x^{[k]}) - \nabla f(x^{[k-1]}),$$

  that takes inspiration from the quasi-Newton methods.

- Newton's or quasi-Newton's direction is chosen, we choose $\overline{\gamma} = 1$

Line-search with backtracking technique

0. Given $x^{[0]}, f, k_{\max}, \mathrm{toll}, b_{\max}, c_1 \in (0,1), \alpha \in (0,1)$
1. For $k = 0, 1, \ldots, k_{\max}$
    1. Choose $d^{[k]}$ descent direction for f in $x^{[k]}$
    2. Use the backtracking algorithm to find γ_k
    3. If in the backtracking algorithm $\mathrm{IND} = 0$ then set $x^{[k+1]} = x^{[k]} + \gamma_k d^{[k]}$, otherwise stop and set $\mathrm{FLAG} = -1$ (failure)
    4. If $\|\nabla f(x^{[k+1]})\| \leq \mathrm{toll}$ then stop, return $x^{[k+1]}$ and $\mathrm{FLAG} = 1$
2. Set $\mathrm{FLAG} = -2$ (failure)

The line-search algorithm outputs the flag IND:

- $\mathrm{FLAG} = 1$ an approximation has been computed with the desired accuracy;
- $\mathrm{FLAG} = -1$ the backtracking technique failed;
- $\mathrm{FLAG} = -2$ the stopping criterion was not satisfied within the maximum number of iterations (failure of line-search method)

Outline

Trust-region strategy

Quadratic model

$T(x^{[k]}, d)$ that approximates $f(x)$ in a neighbourhood of the current position $x^{[k]}$:

$$(\forall d \in \mathbb{R}^N) \qquad T(x^{[k]}, d) = f(x^{[k]}) + \nabla f(x^{[k]})^T d + \frac{1}{2} d^T B^{[k]} d$$

where B_k is the Hessian $H_f(x^{[k]})$ or an approximation to it.

- From the Taylor theorem, we know that the model will be good in a neighbourhood of $x^{[k]}$.
- We look for a step $d^{[k]}$ that minimizes $T(x^{[k]}, d)$ under the constraint that $x^{[k]} + d$ lays in a neighbourhood $\mathcal{B}_{\Delta_k}(x^{[k]})$ of $x^{[k]}$, called **trust region**.
- Most often a ball is used as trust region, but other choices are possible depending on the problem, for example elliptic or rectangular trust regions (used for example when box constraints are present).

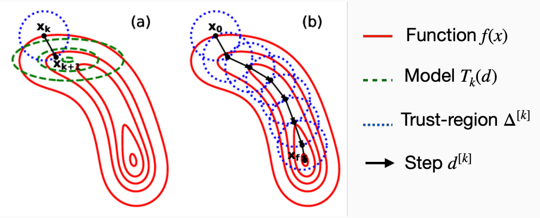
The trust-region subproblem

The trust-region subproblem

We look for a step $d^{[k]}$ solution of the problem

$$\min_{d \in \mathbb{R}^N : \|d\| \leq \Delta_k} T(x^{[k]}, d), \tag{TR}$$

where $\Delta_k > 0$ is called trust-region radius.



An important quantity

$$\rho_k := \frac{\text{ared}}{\text{pred}} = \frac{\overbrace{f(x^{[k]}) - f(x^{[k]} + d^{[k]})}^{\text{actual reduction}}}{\underbrace{T(x^{[k]}, 0) - T(x^{[k]}, d^{[k]})}_{\text{predicted reduction}}}$$

- Since $d = 0$ belongs to the trust region, it holds $T(x^{[k]}, d^{[k]}) \leq T(x^{[k]}, 0)$, but it may not hold $f(x^{[k]} + d^{[k]}) < f(x^{[k]})$.
- If this happens it means that the model is not a good approximation of $f(x)$ in the current trust region, the trust region is too large: we then chose $\Delta_{k+1} < \Delta_k$ and we solve again problem (TR).
- Otherwise we accept the step, i.e. we set $x^{[k+1]} = x^{[k]} + d^{[k]}$. It is possible to show that after a finite number of reductions $f(x^{[k]} + d^{[k]}) < f(x^{[k]})$.

Trust-region algorithm

Given $x^{[0]}$, $\Delta^{[0]}$, toll, $\eta \in [0, 1/4)$, set $k = 0$.

While $\|\nabla f(x^{[k]})\| > \mathrm{toll}$

1. Obtain $d^{[k]}$ by solving (TR).
2. Evaluate $\rho_k = \frac{f(x^{[k]}) - f(x^{[k]} + d^{[k]})}{T(x^{[k]}, 0) - T(x^{[k]}, d^{[k]})}$
3. If $\rho_k < \eta$ set $\Delta_{k+1} = \frac{1}{4}\Delta_k$ and $x^{[k+1]} = x^{[k]}$ (unsuccessful step)
4. If $\rho_k \geq \eta$ set $x^{[k+1]} = x^{[k]} + d^{[k]}$ and $\Delta_{k+1} = 2\Delta_k$ (successful step)
5. Set $k = k + 1$

Trust-region subproblem solution

Lemma
$d^{[k]}$ is a solution of the TR subproblem

$$\min_{d} T(x^{[k]}, d^{[k]}) \text{ subject to } \|d\| \leq \Delta_k$$

for some $\Delta_k > 0$ if and only if $d^{[k]}$ is feasible and there is a scalar $\lambda_k \geq 0$ such that

$$\begin{aligned}
&(B^{[k]} + \lambda_k I)d^{[k]} = -\nabla f(x^{[k]}), \\
&\lambda_k(\Delta_k - \|d^{[k]}\|) = 0 \\
&(B^{[k]} + \lambda_k I) \text{ is positive semidefinite}
\end{aligned}$$

Interior and boundary Solutions

Two situations may occur.

Interior solution

If $B^{[k]} \succ 0$ and $\|(B^{[k]})^{-1}\nabla f(x^{[k]})\| < \Delta_k$, then

$$\lambda_k = 0 \text{ and } d^{[k]} = -(B^{[k]})^{-1}\nabla f(x^{[k]}),$$

which is simply the Newton step.

Boundary solution

Otherwise, $\lambda_k > 0$, and $\|d^{[k]}\| = \Delta_k$.
The solution satisfies $(B^{[k]} + \lambda_k I)d^{[k]} = -\nabla f(x^{[k]})$.

The Secular Equation

Define

$$d(\lambda) = -(B^{[k]} + \lambda I)^{-1} \nabla f(x^{[k]}).$$

The unknown multiplier λ is chosen so that

$$\boxed{\|d(\lambda)\| = \Delta_k.}$$

This nonlinear equation is called the **secular equation**. Each iteration for its solution requires to solve a linear system of the form $(B^{[k]} + \lambda I)d(\lambda) = -\nabla f(x^{[k]})$ for a different λ.

Trust-region for large scale problems

The exact solution of the subproblem is often too expensive in large-scale optimization, motivating approximate solvers such as

- the Cauchy point,
- the Dogleg method,
- the Truncated Conjugate Gradient method.

See "Numerical optimization", Nocedal and Wright.

Special case: 1st order trust-region

$$\boxed{\min_{d \in \mathbb{R}^N : \|d\| \leq \Delta_k} f(x^{[k]}) + \nabla f(x^{[k]})^T d}$$

- Optimal solution **analytically known**
- Lagrangian: $L(d, \lambda) = \nabla f(x^{[k]})^T d + \lambda(\|d\| - \Delta_k)$.
- KKT:

$$\nabla_d L(d, \lambda) = \nabla f(x^{[k]}) + \lambda d / \|d\| = 0$$
$$\lambda(\|d\| - \Delta_k) = 0$$
$$\|d\| \leq \Delta_k$$

Solution of the KKT

- It must hold $\lambda \neq 0$, otherwise $\nabla f(x^{[k]}) = 0$.
- Then $\|d\| = \Delta_k$. But

$$d = -\frac{\nabla f(x^{[k]})\Delta_k}{\lambda},$$

and

$$\|d\| = \frac{\|\nabla f(x^{[k]})\|\Delta_k}{\lambda} = \Delta_k,$$

so $\lambda = \|\nabla f(x^{[k]})\|$ and

$$\boxed{d = -\frac{\nabla f(x^{[k]})\Delta_k}{\|\nabla f(x^{[k]})\|.}}$$

Outline

Adaptive regularization framework

- Quadratic model $T(x^{[k]}, d)$ that approximates $f(x)$ in a neighbourhood of the current position $x^{[k]}$:

$$T(x^{[k]}, d) = f(x^{[k]}) + \nabla f(x^{[k]})^T d + \frac{1}{2} d^T B^{[k]} d, \qquad d \in \mathbb{R}^N,$$

  where $B^{[k]} = H_f(x^{[k]})$ or an approximation to it.
- Instead of using a trust region we add a regularization term to the model to control the step length:

$$m(x^{[k]}, d) = f(x^{[k]}) + \nabla f(x^{[k]})^T d + \frac{1}{2} d^T B^{[k]} d + \frac{\lambda_k}{2} \|d\|^2, \qquad p \in \mathbb{R}^N,$$

- λ_k plays the role of $1/\Delta_k$.
- Much easier to solve the subproblem: λ_k is explicitly given, just solve a linear system $(B^{[k]} + \lambda_k I)d = -\nabla f(x^{[k]})$

Link between Δ_k and λ_k: 1st order

- TR: $d = -\Delta_k \frac{\nabla f(x^{[k]})}{\|\nabla f(x^{[k]})\|}$ and $\|d\| = \Delta_k$
- AR: $d = -\frac{\nabla f(x^{[k]})}{\lambda_k}$ and $\|d\| = \frac{\|\nabla f(x^{[k]})\|}{\lambda_k}$

Then

$$\Delta_k \approx \frac{\|\nabla f(x^{[k]})\|}{\lambda_k}$$

so

$$\boxed{\lambda_k \propto \frac{1}{\Delta_k}}$$

Link between Δ_k and λ_k: 2nd order

- TR: $(B^{[k]} + \lambda I)d = -\nabla f(x^{[k]})$ and $\|d\| = \Delta_k$
- AR: $(B^{[k]} + \lambda_k I)d = -\nabla f(x^{[k]})$

$$
\begin{aligned}
d(\lambda) &= (B^{[k]} + \lambda I)^{-1} \nabla f(x^{[k]}) \\
B^{[k]} &= U \Sigma V^T \\
\|d(\lambda)\|^2 &= \sum_{i=1}^{N} \frac{r_i^2}{(\sigma_i + \lambda)^2}, \; r_i = U^T \nabla f(x^{[k]})
\end{aligned}
$$

Decreasing function of λ

Adaptive regularization strategies

λ_k is updated with the same philosophy as the trust region radius, but the update goes in the opposite direction:

- If $\rho_k < \eta$ set $\lambda_{k+1} = 2\lambda_k$, else set $\lambda_{k+1} = \frac{1}{4}\lambda_k$

ARq: generalization to higher order models

ARq: adaptive regularization methods of order $q \geq 1$

- Build a model : $T_q(x^{[k]}, d) = f(x) + \sum_{i=1}^{q} \frac{1}{i!} \nabla_x^i f(x^{[k]}) \, [d]^i \, .$
- Compute a step $d^{[k]}$:

$$\min_d m_q(x^{[k]}, d) = T_q(x^{[k]}, d) + \frac{\lambda_k}{q+1} \|d\|^{q+1}, \qquad \lambda_k > 0$$

Notable example: ARC (adaptive cubic regularization $q = 2$)

$$m_2(x^{[k]}, d) = f(x^{[k]}) + d^T \nabla f(x^{[k]}) + \frac{1}{2} d^T \nabla^2 f(x^{[k]}) d + \frac{\lambda_k}{3} \|d\|^3$$

[Cartis, Gould, Toint, **Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives**, 2022]

ARq: a better wort-case complexity

- TR/AR: $\|\nabla f(x^{[k]})\| \leq \epsilon$ requires $k = O(\epsilon^{-2})$
- ARq: $\|\nabla f(x^{[k]})\| \leq \epsilon$ requires $k = O(\epsilon^{-(q+1)/q})$
- AR2: $\|\nabla f(x^{[k]})\| \leq \epsilon$ requires $k = O(\epsilon^{-3/2})$

Bottleneck

Subproblem solution: $\min_{d} m_q(x^{[k]}, d)$

AR2

$$T_2(x^{[k]}, d) = f(x^{[k]}) + d^T \nabla f(x^{[k]}) + \frac{1}{2} d^T H_f(x^{[k]}) d,$$
$$m_2(x^{[k]}, d) = T_2(x^{[k]}, d) + \frac{\lambda_k}{3} \|d\|^3$$

$$\rho_k := \frac{\text{ared}}{\text{pred}} = \frac{f(x^{[k]}) - f(x^{[k]} + d^{[k]})}{T_2(x^{[k]}, 0) - T_2(x^{[k]}, d^{[k]})}$$

Approximately minimize m_2 until

$$m_2(x^{[k]}, d) \leq m_2(x^{[k]}, 0) \qquad \|\nabla m_2(x^{[k]}, d)\| \leq \theta \|d\|^2 \tag{1}$$

AR2 algorithm

Given $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_2 \leq \gamma_1 < 1 < \gamma_3$, $\lambda_{\min} > 0$.
Input: $x^{[0]} \in \mathbb{R}^N$, $\lambda_0 > \lambda_{\min}$, $\epsilon > 0$.
$k = 0$
while $\|\nabla f(x^{[k]})\| > \epsilon$ **do**
  • **Model minimization:** Compute the step $d^{[k]}$ by minimizing $m_{k,2}$ until (1) holds.
  • **Acceptance of the trial point:** Compute $\rho_k = \dfrac{f(x^{[k]}) - f(x^{[k]} + d^{[k]})}{T_2(x^{[k]}, 0) - T_2(x^{[k]}, d^{[k]})}$.
  if $\rho_k \geq \eta_1$ (successful step) **then**
    $x^{[k+1]} = x^{[k]} + d^{[k]}$
  else
    (unsuccessful step) $x^{[k+1]} = x^{[k]}$.
  end if
    • **Regularization parameter update:**
    if $\rho_k \geq \eta_1$ **then**

$$\lambda_{k+1} = \begin{cases} \max\{\lambda_{\min}, \gamma_2 \lambda_k\}, & \text{if } \rho_k \geq \eta_2 \text{(very successful step)} \\ \max\{\lambda_{\min}, \gamma_1 \lambda_k\}, & \text{if } \rho_k < \eta_2 \text{(successful step)} \end{cases}$$

    else
      $\lambda_{k+1} = \gamma_3 \lambda_k$.
    end if
    $k = k + 1$
  end while

Eventually a step is accepted $q = 2$

$$
\begin{aligned}
|1-\rho_k| &= \left|1 - \frac{f(x^{[k]}) - f(x^{[k]} + d^{[k]})}{T_2(x^{[k]}, 0) - T_2(x^{[k]}, d^{[k]})}\right| \\
&= \left|\frac{T_2(x^{[k]}, 0) - T_2(x^{[k]}, d^{[k]}) - f(x^{[k]}) + f(x^{[k]} + d^{[k]})}{T_2(x^{[k]}, 0) - T_2(x^{[k]}, d^{[k]})}\right| \\
&\leq \left|\frac{f(x^{[k]} + d^{[k]}) - T_{k,2}(d^{[k]})}{T_2(x^{[k]}, 0) - T_2(x^{[k]}, d^{[k]})}\right| \\
&\leq \frac{L/2\|d\|^3}{\lambda_k/3\|d\|^3} = \frac{3L}{2\lambda_k}
\end{aligned}
$$

If the step is not accepted λ_k is augmented, the ratio becomes smaller, until $|1-\rho_k| < 1-\eta_1$, i.e., $\rho > \eta_1$. So $\lambda_k \leq \lambda_{\max} := \frac{3L}{2(1-\eta_1)}$.

Bound on the gradient

$$
\begin{aligned}
\|\nabla f(x^{[k+1]})\| &\leq \|\nabla f(x^{[k]} + d^{[k]}) - T_2(x^{[k]}, d^{[k]})\| + \|T_2(x^{[k]}, d^{[k]}) + \lambda_k d^{[k]}\|d^{[k]}\|\| \\
&\quad + \lambda_k \|d^{[k]}\|^2 \\
&\leq L\|d^{[k]}\|^2 + \theta\|d^{[k]}\|^2 + \lambda_k\|d^{[k]}\|^2 \leq (L + \theta + \lambda_{\max})\|d^{[k]}\|^2
\end{aligned}
$$

Global convergence

k_0 first index of success, f bounded from below

$$
\begin{aligned}
f(x^{[k_0]}) - \liminf f &\geq \sum_{ksucc} f(x^{[k]}) - f(x^{[k+1]}) \\
&\geq \eta_1 \sum_{ksucc} T_2(x^{[k]}, 0) - T_2(x^{[k]}, d^{[k]}) \geq \frac{\eta_1 \lambda_k}{3} \sum_{ksucc} \|d^{[k]}\|^3 \\
&\geq \frac{\eta_1 \lambda_{\min}}{3(L+\theta+\lambda_{\max})^{3/2}} \sum_{ksucc} \|\nabla f(x^{[k+1]})\|^{3/2}
\end{aligned}
$$

The series is bounded, then $\lim_{k\to\infty} \|\nabla f(x^{[k+1]})\| = 0$.

Complexity

How many successful iterations are needed to drive $\|\nabla f(x^{[k]})\| < \epsilon$?

k_0 first index of success, f bounded from below, k_f the first index for which $\|\nabla f(x_{k_f})\| \geq \epsilon$ and $\|\nabla f(x_{k_f+1})\| < \epsilon$ and K_s set of successful iterations before $k_f - 1$.

$$
\begin{aligned}
f_{k_0} - \liminf f \geq f_{k_0} - f(x_{k_f}) &\geq \sum_{K_s} f(x^{[k]}) - f(x^{[k+1]}) \\
&\geq \eta_1 \sum_{K_s} T_2(x^{[k]}, 0) - T(x^{[k]}, d) \geq \frac{\eta_1 \lambda_k}{3} \sum_{K_s} \|d^{[k]}\|^3 \\
&\geq \frac{\eta_1 \lambda_{\min}}{3(L + \theta + \lambda_{\max})^{3/2}} \sum_{K_s} \|\nabla f(x^{[k+1]})\|^{3/2} \\
&\geq \frac{\eta_1 \lambda_{\min}}{3(L + \theta + \lambda_{\max})^{3/2}} |K_s| \epsilon^{3/2}
\end{aligned}
$$

$$
\boxed{K_s = \mathcal{O}(\epsilon^{3/2})}
$$