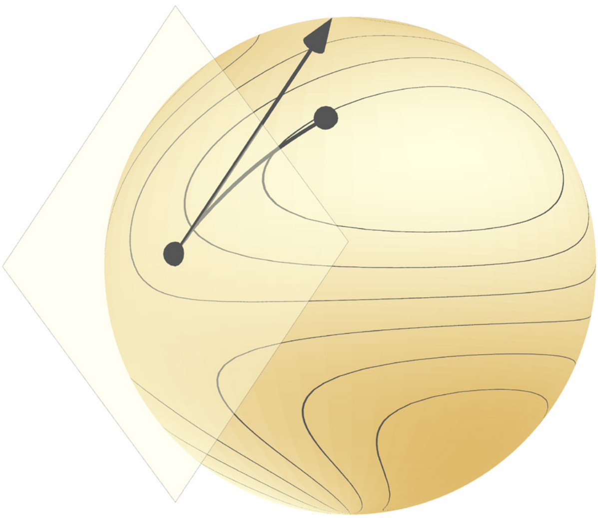
210

Trust-region methods



Spring 2023

Optimization on manifolds, MATH 512 @ EPFL

Instructor: Nicolas Boumal

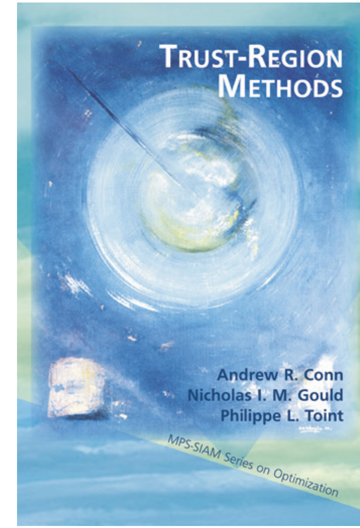
Aiming for the best of both worlds

Close to strict local minima, Newton's method converges fast.

Far from them, Newton is terrible, but gradient descent is fine.

Can we transition from GD-like to Newton-like behavior adaptively?

Historical notes on trust regions



Core ideas traced back to Levenberg 1944, for nonlinear least-squares with damped Gauss-Newton.

Bible of trust-regions on Euclidean spaces by Conn, Gould and Toint: a SIAM book from 2000.

Generalized to Riemannian manifolds by Absil, Baker and Gallivan in 2007, with retractions and an analysis.

Found. Comput. Math. 303–330 (2007)
© 2007 SFoCM
DOI: 10.1007/s10208-005-0179-9

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Trust-Region Methods on Riemannian Manifolds

P.-A. Absil,[1] C. G. Baker,[2] and K. A. Gallivan[2]

Example: Max-Cut Burer-Monteiro rank 2

Run on `/Manopt/examples/maxcut.m` with $\dim \mathcal{M} = 20$:

	iter	cost val	grad. norm	numinner	stopreason
	0	-3.288517e+01	4.768684e+00		
acc	1	-3.935870e+01	2.814923e+00	1	exceeded trust region
acc	2	-4.274683e+01	2.167945e+00	1	exceeded trust region
acc	3	-4.457110e+01	1.453372e+00	2	exceeded trust region
acc	4	-4.620138e+01	2.500653e+00	2	negative curvature
acc	5	-4.854677e+01	2.891663e+00	2	negative curvature
acc	6	-5.066439e+01	1.918719e+00	2	exceeded trust region
acc TR+	7	-5.233968e+01	1.180198e+00	3	exceeded trust region
acc	8	-5.280136e+01	6.871197e-01	7	reached target residual-kappa (linear)
acc	9	-5.297255e+01	9.966179e-02	5	reached target residual-kappa (linear)
acc	10	-5.297890e+01	1.352219e-02	6	reached target residual-theta (superlinear)
acc	11	-5.297897e+01	1.905915e-04	8	reached target residual-theta (superlinear)
acc	12	-5.297897e+01	3.996911e-08	13	reached target residual-theta (superlinear)

Newton method's shaky foundations

$m_x : T_x \mathcal{M} \to \mathbb{R}$

$$f\big(R_x(s)\big) \approx m_x(s) \stackrel{\text{def}}{=} f(x) + \langle \operatorname{grad} f(x), s \rangle_x + \frac{1}{2} \langle s, \operatorname{Hess} f(x)[s] \rangle_x$$

Newton's method blindly jumps to the critical point of $m_x \colon \mathrm{T}_x\mathcal{M} \to \mathbf{R}$.

Makes sense if $\operatorname{Hess} f(x) \succ 0$ and if s is small.

What if not?

Retain the core idea, with nuance

$$f\big(R_x(s)\big) \approx m_x(s) \stackrel{\text{def}}{=} f(x) + \langle \mathrm{grad} f(x), s\rangle_x + \frac{1}{2}\langle s, \mathrm{Hess} f(x)[s]\rangle_x$$

Keep the idea of a local model for the pullback $f \circ R_x \colon \mathrm{T}_x\mathcal{M} \to \mathbf{R}$.

But remember: it's local, so only trust it in a small region of $\mathrm{T}_x\mathcal{M}$;

we mean to minimize f, so aim to minimize m_x;

minimizing m_x is a means to an end, don't overdo it.

Consider a subproblem at each iteration

$$f\big(R_x(s)\big) \approx m_x(s) \stackrel{\text{def}}{=} f(x) + \langle \operatorname{grad} f(x), s \rangle_x + \frac{1}{2}\langle s, \operatorname{Hess} f(x)[s] \rangle_x$$

what is good enough?

Select $s \in \mathrm{T}_x\mathcal{M}$ as an approximate solution of:

$$\min_{v \in \mathrm{T}_x\mathcal{M}} m_x(v) \quad \text{subject to} \quad \|v\|_x \leq \Delta$$

figure it out adaptively

Then, maybe go to $R_x(s)$.

what if we don't?

$\rho' = 1/10$ (typical)

Parameters: $\rho' \in (0, 1/4)$, $\bar{\Delta} > 0$

Initialize: $x_0 \in \mathcal{M}$; $\Delta_0 \in (0, \bar{\Delta}]$

For k in 0,1,2, ...

$H_k = \mathrm{Hess} f(x_k)$

$$m_k(v) = \frac{1}{2}\langle v, H_k v\rangle_{x_k} + \langle \mathrm{grad} f(x_k), v\rangle_{x_k} + f(x_k)$$

- Compute s_k $\in T_{x_k}\mathcal{M}$ as approx. solution of $\min_{v \in \mathrm{T}_{x_k}\mathcal{M}} m_k(v)$ s.t. $\|v\|_{x_k} \leq \Delta_k$

  Trust Region Subproblem (TRS)

- Tentative next iterate: $x_k^+ = R_{x_k}(s_k)$
- Assess its quality: $\rho_k = \dfrac{f(x_k) - f(x_k^+)}{m_k(0) - m_k(s_k)}$ $\quad m_k(0) - m_k(s_k) > 0$
- Accept or reject: $x_{k+1} = \begin{cases} x_k^+ & \text{if } \rho_k > \rho' \quad \text{: successful step} \\ x_k & \text{otherwise} \quad \text{: unsuccessful step.} \end{cases}$
- Update the radius: $\Delta_{k+1} = \begin{cases} \frac{1}{4}\Delta_k & \text{if } \rho_k < 1/4 \\ \min(2\Delta_k, \bar{\Delta}) & \text{if } \rho_k > 3/4 \text{ and } \|s_k\|_{x_k} = \Delta_k \\ \Delta_k & \text{otherwise} \end{cases}$

Minimal effort for the subproblem

The trust-region subproblem (TRS) takes the form

$$\min_{v \in \mathrm{T}_x\mathcal{M}} m_x(v) \quad \text{subject to} \quad \|v\|_x \leq \Delta$$

with $m_x(v) \stackrel{\text{def}}{=} f(x) + \langle \operatorname{grad} f(x), v \rangle_x + \frac{1}{2}\langle v, \operatorname{Hess} f(x)[v] \rangle_x$.

Cauchy step: let $s^C = -t \cdot \operatorname{grad} f(x)$ with optimal $t \geq 0$.

Exercise: Find how to compute s^C and check:

$$m_x(0) - m_x(s^C) \geq \frac{1}{2} \min\left(\Delta, \frac{\|\operatorname{grad} f(x)\|_x}{\|\operatorname{Hess} f(x)\|_x}\right) \|\operatorname{grad} f(x)\|_x$$

Globally not worse than gradient descent

A0 $f(x) \geq f_{\text{low}}$ for all $x \in \mathcal{M}$

A1 R is a second-order retraction

A2 $\left|f\left(R_x(v)\right) - f(x) - \langle \operatorname{grad} f(x), v \rangle_x\right| \leq \frac{L}{2} \|v\|_x^2$ for all $(x, v) \in \mathrm{T}\mathcal{M}$

A3 Subproblem solver ensures $m_k(0) - m_k(s_k) \geq$ Cauchy decrease.

Theorem: The algorithm finds x_k with $\|\operatorname{grad} f(x_k)\|_{x_k} \leq \varepsilon$ for some

$$k \leq \frac{48L(f(x_0) - f_{\text{low}})}{\rho'} \frac{1}{\varepsilon^2} + \frac{1}{2}\log_2\left(\frac{16L\Delta_0}{\varepsilon}\right)$$

given any $\varepsilon \leq 16L\Delta_0$.

Why does this work?

- The trust region radius cannot become arbitrarily small.

- Successful steps yield good decrease when the gradient is large.

- Most steps are successful.

Parameters: $\rho' \in (0,1/4)$ and $\bar{\Delta} > 0$

Initialize: $x_0 \in \mathcal{M}$ and $\Delta_0 \in (0, \bar{\Delta}]$

For k in $0,1,2,\ldots$

- Compute s_k as approx. solution of $\min_{v \in \mathrm{T}_{x_k}\mathcal{M}} m_k(v)$ s.t. $\|v\|_{x_k} \leq \Delta_k$

$$m_k(v) = \frac{1}{2}\langle v, H_k v\rangle_{x_k} + \langle \mathrm{grad} f(x_k), v\rangle_{x_k} + f(x_k)$$

- Tentative next iterate: $x_k^+ = R_{x_k}(s_k)$
- Assess its quality: $\rho_k = \frac{f(x_k) - f(x_k^+)}{m_k(0) - m_k(s_k)}$
- Accept or reject: $x_{k+1} = \begin{cases} x_k^+ & \text{if } \rho_k > \rho' \\ x_k & \text{otherwise} \end{cases}$

$$f(x_k) - f(x_{k+1}) = f(x_k) - f(x_k^+) = \rho_k \left(m_k(0) - m_k(s_k) \right) \geq \rho' \left(\text{Cauchy decrease} \right).$$

- Update the radius: $\Delta_{k+1} = \begin{cases} \frac{1}{4}\Delta_k & \text{if } \rho_k < \frac{1}{4} \\ \min(2\Delta_k, \bar{\Delta}) & \text{if } \rho_k > \frac{3}{4} \text{ and } \|s_k\|_{x_k} = \Delta_k \\ \Delta_k & \text{otherwise} \end{cases}$

Numerical notes

- Riemannian trust-regions (RTR) is available in Manopt.
- Default parameters: $\rho' = 0.1$ and $\bar{\Delta} = \operatorname{diam}\mathcal{M}$ or $\bar{\Delta} = \sqrt{\dim \mathcal{M}}$.
- Default initialization: $\Delta_0 = \bar{\Delta}/8$ and x_0 random on $\mathcal{M}$.
- Default subproblem solver: truncated conjugate gradients (tCG)
- Don't check "$\|s_k\|_{x_k} = \Delta_k$" in floating point arithmetic:
  have the subproblem solver return s_k + a boolean "limitedbyTR".
- Computing ρ_k in floating point arithmetic is tricky: regularize.

See §6.4.6 for details.

Relaxed assumptions and finer guarantees

- $m_k(v) \stackrel{\text{def}}{=} f(x_k) + \langle \mathrm{grad} f(x_k), v \rangle_{x_k} + \frac{1}{2} \langle v, H_k[v] \rangle_{x_k}$

  We don't have to set $H_k = \mathrm{Hess} f(x_k)$. E.g., finite differences.

- $\|\mathrm{grad} f(x_k)\|_{x_k} \to 0$ under mild assumptions (§6.4.5)

- RTR can find a point with small gradient *and* nearly positive semidefinite Hessian, i.e., approximately second-order critical.