Ditch the Hessian Hassle with Riemannian Trust Regions

Nicolas Boumal, Inria & ENS Paris

Geometric Science of Information, GSI 2015
Oct. 30, 2015, Paris
The goal is to optimize a smooth function on a smooth manifold
The Trust Region method is like Newton’s with a safeguard.
On the tangent space, optimize the model in a trust region.
The trust-region way

Lift the cost to the tangent space,

Locally optimize an approximation,

And map back to the manifold.
Retractions map from the tangent space to the manifold

\[ \text{Retr}_x : T_x M \rightarrow M \]

Typical example on the sphere:

\[ \text{Retr}_x (\dot{x}) = \frac{x + \dot{x}}{\| x + \dot{x} \|} \]
The lifted cost is approximated

\[ \hat{f}_k = f \circ \text{Retr}_{x_k} : T_{x_k} M \to \mathbb{R} \]

First order Taylor

\[ \hat{m}_k(\dot{x}) = f(x_k) + \langle \text{grad} f(x_k), \dot{x} \rangle + \frac{1}{2} \langle \dot{x}, H_k[\dot{x}] \rangle \]

Expect better performance if \( H_k \) close to Hessian
Hessian approximated as identity

Convergence of RTR with various quadratic models

True Hessian

Gradient norm vs. Time [s]
Computing the Hessian often is a hassle
It is tempting to approximate the Hessian with finite differences

\[ \text{Hess } f(x)[\dot{x}] \approx \frac{\text{Tr}_x \left( \text{grad } f(\text{Retr}_x(t\dot{x})) \right) - \text{grad } f(x)}{t} \]

Gradients are transported to a common tangent space.
Hessian approximated as identity

Hessian approximated with finite differences

True Hessian
But the theory breaks down

\[ H[\dot{x}] = \frac{\text{Tr}_x \left( \text{grad} f(\text{Retr}_x(t \dot{x})) \right) - \text{grad} f(x)}{t} \]

This operator is **nonlinear**.

Quid of convergence?
Two key ingredients for global convergence to critical points

The model agrees with the cost to first order:

\[ (\hat{f}_k - \hat{m}_k)(\dot{x}) \leq c\|\dot{x}\|^2 \text{ for all } \dot{x} \]

At each step, obtain sufficient model decrease:

Optimize at least along steepest descent dir.
Make $H_k$ **radially** linear

$$\text{Tr}_x \left( \operatorname{grad} f \left( \operatorname{Retr}_x \left( \varepsilon \frac{\dot{x}}{\|\dot{x}\|} \right) \right) \right) - \operatorname{grad} f (x)$$

Then, $H[\alpha \dot{x}] = \alpha H[\dot{x}]$ for all $\alpha \geq 0$. 
The paper shows...

Finite difference ops are uniformly bounded,

And how to enforce Cauchy decrease with Steihaug-Toint tCG, even after losing linearity.

Leads to **global convergence to critical points**.

Caveat: no results regarding local rates.
That's Folks!
All the algorithms are in Manopt

Open source, documented and user friendly

Go to www.manopt.org