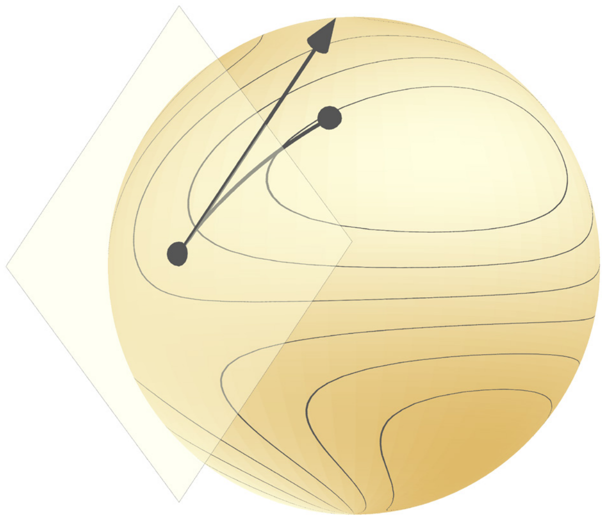
207

Newton's method



Spring 2023

Optimization on manifolds, MATH 512 @ EPFL

Instructor: Nicolas Boumal

Exploiting second-order information

We aim to minimize $f : \mathcal{M} \to \mathbf{R}$, smooth on a manifold.

Choose a retraction R, a Riemannian metric on $\mathcal{M}$, and $x_0 \in \mathcal{M}$.

Algorithms iterate $x_{k+1} = R_{x_k}(s_k)$ with some choice of s_k.

Gradient descent: $s_k = -\alpha_k \mathrm{grad} f(x_k)$. Fine, but slow…

Exploit $\mathrm{Hess} f(x_k)$ to choose a better s_k?

Recall second-order Taylor expansions, with $c(t) = R_x(ts)$:

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

$$+ \frac{1}{2}\langle \mathrm{grad} f(x), c''(0)\rangle_x + O(\|s\|_x^3)$$

$$f(R_x(s)) \simeq m_x(s) \triangleq f(x) + \langle \mathrm{grad} f(x), s\rangle_x + \frac{1}{2}\langle s, \mathrm{Hess} f(x)[s]\rangle_x$$

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

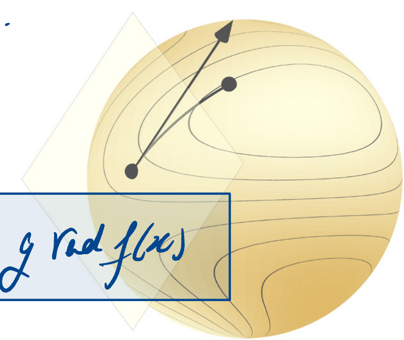
$$f(x_{k+1}) = f(R_{x_k}(s_k)) \simeq m_{x_k}(s_k).$$

Assume $\operatorname{Hess} f(x_k) \succ 0$; then the minimizer of m_{x_k} is attained at its critical point, because the model is convex.

$$D m_x(s)[\dot{s}] = \langle \operatorname{grad} f(x), \dot{s} \rangle_x + \cancel{\tfrac{1}{2}} \langle \dot{s}, \operatorname{Hess} f(x)[s] \rangle_x \cancel{+ \tfrac{1}{2} \langle s, \operatorname{Hess} f(x)[\dot{s}] \rangle_x}$$

$$= \langle \dot{s}, \underbrace{\operatorname{grad} f(x) + \operatorname{Hess} f(x)[s]}_{\doteq \operatorname{grad} m_x(s)} \rangle_x$$

$\Rightarrow$ The minimizer of m_x if $\operatorname{Hess} f(x) \succ 0$ is the vector $s \in T_x\mathcal{M}$ s.t. $\boxed{\operatorname{Hess} f(x)[s] = -\operatorname{grad} f(x)}$



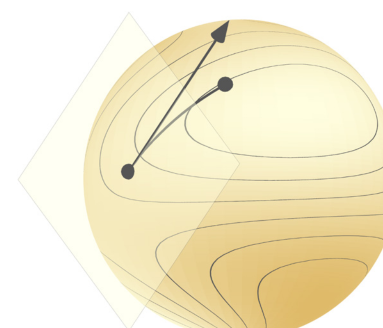
Newton's method

- Choose $x_0 \in \mathcal{M}$.
- For k in $0, 1, 2, 3, \ldots$

  in $T_{x_k}\mathcal{M}$.

  Solve the linear system $\operatorname{Hess} f(x_k)[s_k] = -\operatorname{grad} f(x_k)$.

  $x_{k+1} = R_{x_k}(s_k)$



Fast local convergence…

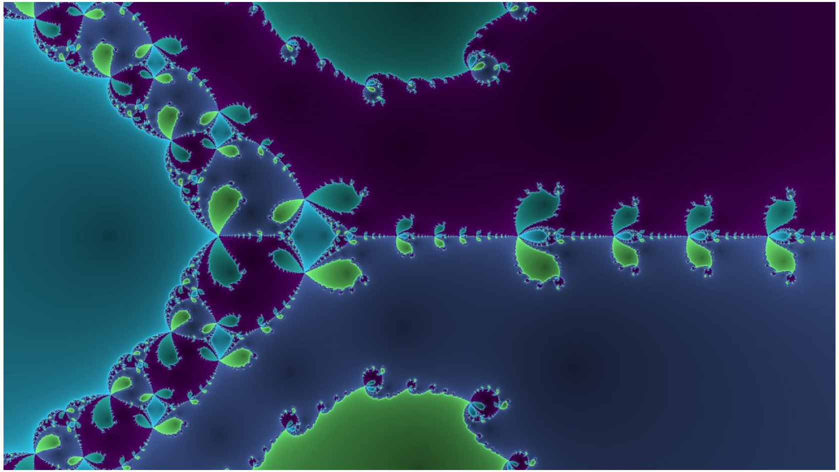
Theorem: Let $x_\star \in \mathcal{M}$ satisfy $\operatorname{grad} f(x_\star) = 0$ and $\operatorname{Hess} f(x_\star) \succ 0$. There exists a neighborhood $\mathcal{U}$ of $x_\star$ on $\mathcal{M}$ such that, for all $x_0 \in \mathcal{U}$, the sequence $x_0, x_1, x_2, \ldots$ generated by Newton's method converges to $x_\star$ at least quadratically.

§ 6.2

… and nothing else.

The global behavior of Newton's is horrendous.

See 3blue1brown video on Newton's fractals (picture)



There are several fixes. The classical "globalized" algorithms are:

Trust-region methods

Cubic regularization methods

These also aim to control the per-iteration computational cost.