[Some optional Exercises on Conjugate Grad and Trust-Region are up on blackboard]

3. Stepsize Selection and Descent

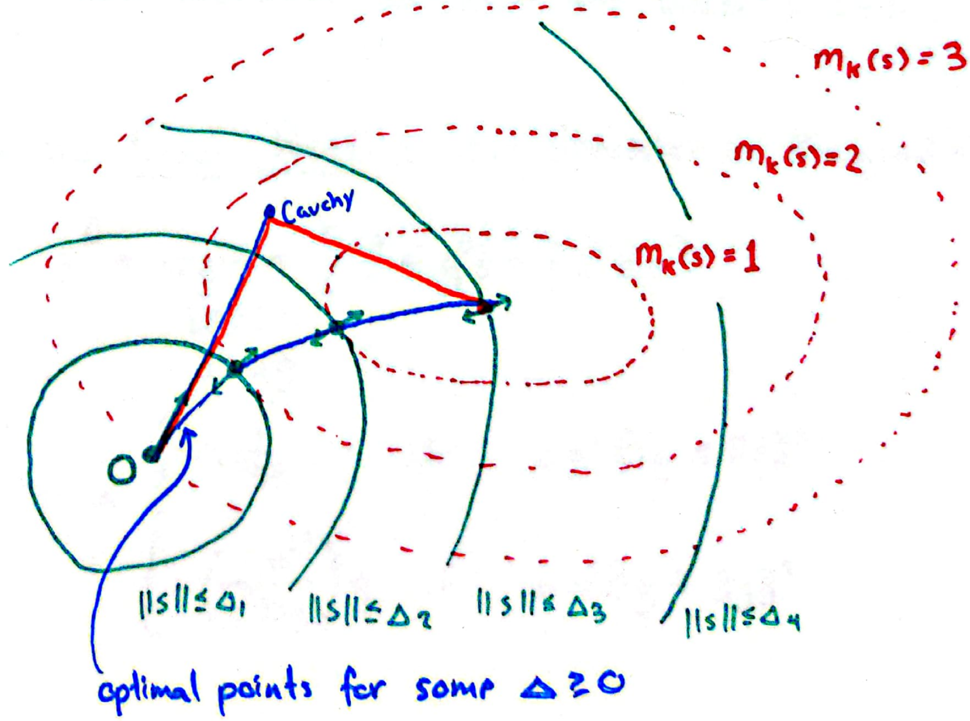
Recall Trust Region steps involve solving the nonconvex minimization (considered last time)

$$\boxed{s_k = \underset{\|s\|_2 \le \Delta_k}{\operatorname{argmin}} \left\{ \underbrace{f(x_k) + \nabla f(x_k)^T s + \tfrac{1}{2} s^T \overset{\text{indefinite}}{B_k} s}_{m_k(s)} \right\}}$$

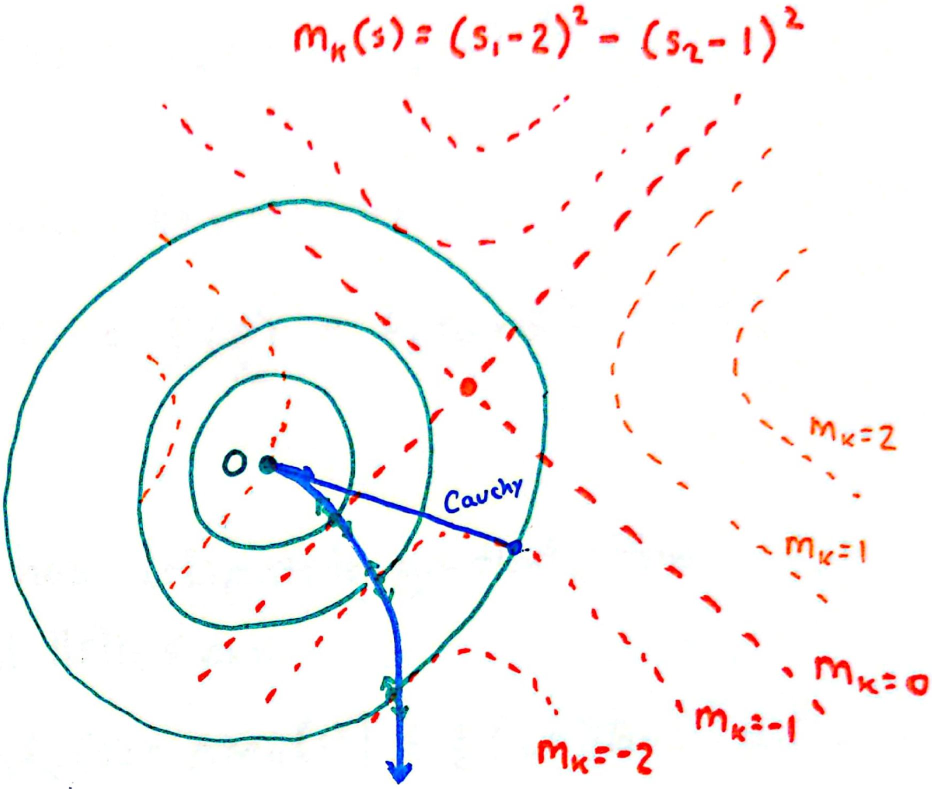
where B_k models the Hessian and Δ_k limits the nearby area to search.

What does the solution look like as we vary Δ_k?

If $m_k(s)$ is convex with $B_k \succ 0$, then



If $m_k(s)$ is nonconvex (say B_k indefinite), then



This motivates two heuristics for some (*) decently well...

Define the Cauchy Point as the minimize in the grad direction.

$$s^C = \underset{\substack{\|s\| \leq \Delta \\ s = \alpha g}}{\text{argmin}} \left\{ f + g^T s + \frac{1}{2} s^T B s \right\}$$

$$= \begin{cases} -\Delta \frac{g}{\|g\|} & \text{if} \quad \Delta g^T B g \leq \|g\|^2 \\ -\left(\frac{\|g\|^2}{g^T B g}\right) g & \text{if} \quad \Delta g^T B g \geq \|g\|^2 \end{cases}$$

Define the Dogleg Path as a mixture of gradient and Newton directions (assuming $B \succ 0$):

$$s^{DL}(\tau) = \begin{cases} \tau s^{GD} & \text{if } 0 \le \tau \le 1 \\ s^{GD} + (\tau - 1)(s^{N} - s^{GD}) & \text{if } 1 \le \tau \le 2 \end{cases}$$

where $s^{GD} = -\left(\frac{\|g\|^2}{g^T B g}\right) g$ and $s^N = -B^{-1} g$.

Pick s_k minimizing $m_k(s^{DL}(\tau))$.

One more heuristic. Solve in 2D subspace s^{GD}, s^{N},

$$s_k = \underset{\substack{\|s\| \le \Delta \\ s \in \operatorname{span}(-g, -B^{-1}g)}}{\operatorname{argmin}} \; f + g^T s + \frac{1}{2} s^T B s$$

Back to guaranteeing descent.

Not true that every Δ gives descent.

Small Δs work.

Define model objective decrease as

$$\Delta m_K(s) = m_K(0) - m_K(s) \ (>0)$$

and function value decrease as

$$\Delta f_K(s) = f(x_K) - f(x_K+s) \ (\gtrless 0)$$

<u>Lemma 1</u> If f has L-Lipschitz gradient, then for all $\|s\|_2 \le \Delta_K$

$$|\Delta f_K(s) - \Delta m_K(s)| \le \frac{1}{2}(L + \|B_K\|)\Delta_K^2 .$$

If f has Q-Lipschitz Hessian then

$$|\Delta f_K(s) - \Delta m_K(s)| \le \frac{Q}{6}\Delta_K^3 + \frac{\|B_K - \nabla^2 f(x_K)\|}{2}\Delta_K^2$$

Proof. $$|\Delta f_K(s) - \Delta m_K(s)| = |f(x_K+s) - (f(x_K) + \nabla f(x_K)^T s + \frac{1}{2}s^T B_K s)|$$
$$\le |f(x_K+s) - (f(x_K) + \nabla f(x_K)^T s)| + \frac{1}{2}|s^T B_K s|$$
$$\le \frac{L}{2}\|s\|_2^2 + \frac{\|B_K\|}{2}\|s\|_2^2$$
$$\le \frac{1}{2}(L + \|B_K\|)\Delta_K^2 . \quad \checkmark$$

Adding and subtracting

$$|\Delta f_K(s) - \Delta m_K(s)| \le |f(x_K+s) - (2^{\text{nd}}\text{ order model})| + \frac{1}{2}|s^T(\nabla^2 f(x_K) - B_K)s|$$
$$\le \frac{Q}{6}\Delta_K^3 + \frac{1}{2}\|\nabla^2 f(x_K) - B_K\|\Delta_K^2 . \quad \square$$

Lemma 2 The Cauchy Point s^c has

$$\Delta m_k(s^c) \geq \frac{1}{2}\|\nabla f(x_k)\| \cdot \min\left[\frac{\|\nabla f(x_k)\|}{\|B_k\|}, \Delta_k\right].$$

Proof. If $\Delta_k g_k^T B_k g_k \leq \|g_k\|^2$, where $g_k = \nabla f(x_k)$

$$\Delta m_k(s^c) = \Delta\|g_k\| - \frac{1}{2}\Delta^2 \frac{g_k^T B_k g_k}{\|g_k\|^2}$$

$$\geq \Delta\|g_k\| - \frac{1}{2}\Delta\|g_k\|$$

$$= \frac{1}{2}\Delta\|g_k\|.$$

Otherwise $\Delta_k g_k^T B_k g_k > \|g_k\|^2$

$$\Delta m_k(s) = \frac{\|g_k\|^4}{g_k^T B_k g_k} - \frac{1}{2}\frac{\|g_k\|^4}{g_k^T B_k g_k}$$

$$= \frac{1}{2}\frac{\|g_k\|^4}{g_k^T B_k g_k}$$

$$\geq \frac{1}{2}\frac{\|g_k\|^2}{\|B_k\|} \quad \text{by } g_k^T B_k g_k \leq \|B_k\|\|g_k\|^2.$$

$\square$

Together these give a descent bound

$$|\Delta f_k(s) - \Delta m_k(s)| \leq \frac{1}{2}(L + \|B_k\|)\Delta_k^2 \quad \text{by Lemma 1}$$

$$\Rightarrow \Delta f_k(x_k) \geq \Delta m_k(s) - \frac{1}{2}(L + \|B_k\|)\Delta_k^2$$

$$\geq \frac{1}{2}\|\nabla f(x_k)\|\min\left\{\frac{\|\nabla f\|}{\|B_k\|}, \Delta_k\right\} - \frac{1}{2}(L + \|B_k\|)\Delta_k^2.$$

> 0 for small $\Delta > 0$.

4. A Full Trust Region Method

Let's measure how much we trust a step s

as $\rho_k(s) := \dfrac{\Delta f_k(s)}{\Delta m_k(s)}$ (Note $\to 1$ as $\underset{\substack{\Updownarrow \\ s \to 0}}{\Delta_k \to 0}$).

If $\rho_k(s)$ near one or greater, this step is great!

If $\rho_k(s)$ near zero or less, this step is bad!

Picking Thresholds $0 < \underset{\approx 0.1}{\eta_s} \leq \underset{\approx 0.9}{\eta_{vs}} \leq 1$, x_0, Δ_0,

we iterate with

for $k = 0, 1, 2, \dots$

1. Build $m_k(s)$
2. Find s_k minimizing $m_k(s)$ s.t. $\|s\| \leq \Delta_k$ (at least as well as Cauchy.)
3. Compute $\rho_k(s_k)$
4. If $\rho_k(s_k) \geq \eta_{vs}$ $\boxed{\text{Very Successful!}}$

   $x_{k+1} = x_k + s_k$
   $\Delta_{k+1} = 2\Delta_k \leftarrow \gamma_{sv} \cdot \Delta_k$

   Else if $\rho_k(s_k) \geq \eta_s$ $\boxed{\text{Success!}}$

   $x_{k+1} = x_k + s_k$
   $\Delta_{k+1} = \Delta_k$

   Else

   $x_{k+1} = x_k$
   $\Delta_{k+1} = \Delta_k / 2$. $\boxed{\text{Unsuccessful :c}}$

5. Convergence Guarantees

We won't have many unsuccessful steps

$$\text{since } \rho_k(s_k) = \frac{\Delta f_k(s_k)}{\Delta m_k(s_k)}$$

$$\geq \frac{\Delta m_k(s_k) - \frac{1}{2}(L+\|B_k\|)\Delta_k^2}{\Delta m_k(s_k)}$$

$$= 1 - \frac{(L+\|B_k\|)\Delta_k^2}{\|\nabla f(x_k)\| \min\left\{\frac{\|\nabla f\|}{\|B_k\|}, \Delta_k\right\}}$$

$\Rightarrow$ Halving Δ_k makes ρ_k converge linearly to 1.

Claim: $\Delta_k \to 0$ only if $\|\nabla f(x_k)\| \to 0$.

Proof. Suppose all $\|\nabla f(x_k)\| > \varepsilon > 0$. Unsuccessful steps have

$$\eta_s \geq \rho_k = 1 - \frac{(L+\|B_k\|)\Delta_k^2}{\|\nabla f\| \min\left\{\frac{\|\nabla f\|}{\|B_k\|}, \Delta_k\right\}}$$

$$\Leftrightarrow \quad 1-\eta_s \leq \max\left\{\frac{\beta(L+\beta)\Delta_k^2}{\varepsilon^2}, \frac{(L+\beta)\Delta_k}{\varepsilon}\right\}$$

where $\beta \geq \|B_k\|$.

$$\Leftrightarrow \quad \Delta_k \geq \min\left\{\sqrt{\frac{1-\eta_s}{\beta(L+\beta)}}\,\varepsilon, \frac{1-\eta_s}{L+\beta}\,\varepsilon\right\}.$$

$\Rightarrow$ At any iteration $\Delta_k \geq \frac{1}{2}\left(\qquad \right)$. $\Rightarrow \Delta_k \not\to 0$. $\square$

Theorem (Global Convergence, 2018, Curtis, Lubberts, Robinson)

AMS

"Concise Complexity Analyses for Trust Region Methods)

$O(1/\varepsilon^2)$ rate, $(O(1/\varepsilon^{3/2})$ rate with improvements$)$

Theorem If $B_k = \nabla^2 f(x_k)$ (or converging to it), superlinear convergence

(Nocedal + Wright, Thm 4.9).

Semester Recap

We have built the machinery and theory for solving

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

for a huge variety of functions f.

Optimality Conditions - What can we locally guarantee and when is this globally meaningful.

First-Order Optimization - Methods that scale in dimension

- Smooth OPT with optimal acceleration,
- Nonsmooth OPT with subgradients and prox,
- Stochastic/Coordinate Methods with even cheaper per iteration costs,
- Conjugate Gradients and Least Squares.

Second-Order Methods - Methods that scale in accuracy

- Newton's Method with Quadratic Convergence,
- Quasi Newton (BFGS) with Superlinear Convergence,
- Trust-Regions for Indefinite Local Improvement.

Thank you all for your attention