§9 Derivative free optimization ($\min_{x \in \mathbb{R}^n} f(x)$)

• If no derivatives are available one can always use finite difference approx:

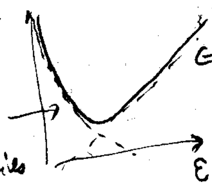
$$\nabla_\varepsilon f(x) = \left[\frac{f(x+\varepsilon e_i) - f(x)}{\varepsilon} \right]_i \qquad \text{FWD diff}$$

or

$$= \left[\frac{f(x+\varepsilon e_i) - f(x-\varepsilon e_i)}{\varepsilon} \right]_i \qquad \text{CENTRAL diff}$$

However this sensitive to "noise" in f.

Recall



(for fwd differences but similar for central diff)

Model based methods

Idea: min quadratic model of function:

$$m_k(x_k+p) = c + g^T p + \frac{1}{2} p^T H p$$

Newton's method: $c = f_k$, $g = \nabla f_k$, $H = \nabla^2 f_k$

Since we don't have derivatives we use a quadratic interpolant on last q points; requiring that

$$m_k(y_\ell) = f(y_\ell), \quad \ell = 1, 2, \dots, q \qquad (*)$$

How many points do we need to (in principle) uniquely determine c, g, H?

$$\underbrace{c}_{1},\ \underbrace{g}_{n},\ \underbrace{H}_{\frac{n(n+1)}{2}} \qquad \rightsquigarrow q = \frac{(n+2)(n+1)}{2}$$

($(*)$ is nothing else than a linear system w/ $\frac{(n+2)(n+1)}{2}$ unknowns)

Once we have m_k, we can compute a step p by solving a trust region subproblem:

$$\min_{p} \; m_k(x_k + p)$$
$$\text{s.t. } \|p\|_2^2 \leq \Delta$$

Model Based Derivative free method

choose $Y = \{y_1, \ldots, y_q\}$ s.t. quadratic interpolant can be found uniquely

choose x_0 s.t. $f(x_0) \leq f(y)$, $y \in Y$.

Δ_0, $\eta \in (0,1)$

for $k = 1 \ldots$

form $m_k(x_k + p)$ quadratic s.t $m_k(y_i) = f(y_i)$

Solve TR subproblem $\rightarrow p_k$

$$\rho = \frac{f(x_k + p_k) - f(x_k)}{m_k(x_k + p_k) - m_k(x_k)} = \frac{\text{ared } k}{\text{pred } k}$$

if $\rho \geq \eta$

replace a node in Y by $x_k + p_k$

choose $\Delta_{k+1} \geq \Delta_k$

$x_{k+1} = x_k + p_k$

break;

else if set Y is good

choose $\Delta_{k+1} < \Delta_k$

$x_{k+1} = x_k + p_k$

break;

Change Y by at least changing a point so that condition number of interpol. eq is better.

$\Delta_{k+1} = \Delta_k$

Let $y_* = \arg\min_{y \in Y} f$. $\rho = \frac{\text{ared } k}{\text{pred } k}$

| if $\rho \geq \eta$
| $x_{k+1} = y_*$
| else
| $x_k = x_k$

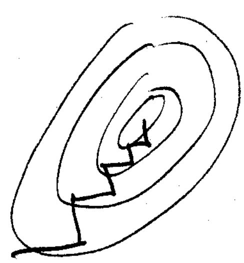
end for

becomes quickly expensive: $O(n^2)$ to start. Even doing smart updates of quadratic model complexity of one iteration is $O(n^4)$.
$\rightsquigarrow$ Can do hybrid method where model is linear (only c & g $= n+1$ params) so each step is $O(n^3)$ and when enough points have been explored switch to quadratic model

Pattern search methods

simplest one (but not very efficient):

- Coordinate search method



do line search on $e_1, e_2, e_3 \dots e_n, e_1 \dots$
other patterns possible $e_1, \dots e_n, e_{n-1}, \dots, e_1 \dots$
problem: is not guaranteed to converge to a stationary point ($\|\nabla f\| = 0$)

- pattern search methods are more general in that search directions D_k are chosen in order to guarantee convergence to a stationary point.

Pattern search algorithm

γ_{tol} = convergence tolerance

θ_{max} = contraction param. < 1

$\rho : [0,\infty) \to \mathbb{R}$, $\rho(t)$ increasing and $\frac{\rho(t)}{t} \to 0$ as $t \to 0$.

x_0, $\gamma_0 > \gamma_{tol}$, $\mathcal{D}_0$ = initial set of dir.
steplength

for $k = 1, 2, \dots$ ~ SDC

if $\gamma_k \leq \gamma_{tol}$ STOP

if $f(x_{k+1} + \gamma_k p_k) < f(x_k) - \rho(\gamma_k)$ for some $p_k \in \mathcal{D}_k$

$x_{k+1} = x_k + \gamma_k p_k$ for that p_k

$\gamma_{k+1} = \phi_k \gamma_k$ $(\phi_k \geq 1)$ (increase steplength)

else

set $x_{k+1} = x_k$

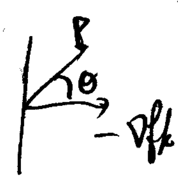
$\gamma_{k+1} = \theta_k \gamma_k$, where $0 < \theta_k \leq \theta_{max} < 1$ (decrease steplength)

What condition is used?

Recall Zoutendijk condition for convergence for a method that gives direction p_k and unit stepsize α_k satisfying Wolfe conditions:

convergence $\Rightarrow$ $\displaystyle\sum_{k=0}^{\infty} \cos^2\theta_k \, \|\nabla f_k\|^2 < \infty$

$$\cos\theta_k = \frac{-p \cdot \nabla f_k}{\|p\| \, \|\nabla f_k\|}$$



We require that:

$$K(\mathcal{D}_k) = \min_{v \in \mathbb{R}^n} \max_{p \in \mathcal{D}_k} \frac{v^T p}{\|v\| \|p\|} \geq \delta$$

$\to$ at least a direction $p \in \mathcal{D}_k$ gives $\cos\theta \geq \delta$

For convergence it is also assumed that δ ~ size of step that is:

$$\forall p \in D_k : \beta_{min} \leq \|p\| \leq \beta_{max}$$

Then:

$$-\nabla f_k^T p \geq K(D_k) \|\nabla f_k\| \|p\| \geq \delta \beta_{min} \|\nabla f_k\|$$

Sample D_k:

Coordinate:

$$\{e_1, \ldots e_n, -e_1, \ldots, -e_n\}$$ $\rightarrow$ Here $K(D_k) = 0$ so method may not converge to stationary point.

Simplex:

$$p_i = \frac{1}{2n} e - e_i, \quad i = 1 \ldots n$$

$$p_{n+1} = \frac{1}{2n} e.$$

Of course other directions can be added to D_k, coming from e.g. heuristics