Local descent - L2 - 29/09 - Optimization

$f(x) \to \min,\ x \in D \subseteq \mathbb{R}^n,\ n \geq 1$, f is smooth

goal is to iteratively find a sequence $x^{(1)}, x^{(2)}, \dots, x^*$ where x^* is a solution of the optimization problem (local or global minimum), realizing the descent

$f(x^{(1)}) > f(x^{(2)}) > \dots$ for all or most of the iterates

$\Leftrightarrow \nabla f(x^*) = 0$

General descent method

Given a starting point $x^{(1)} \in D$

Repeat:

1. Determine descent direction $p^{(k)}$ (often $\|p^{(k)}\| = 1$)
2. Determine step size/learning rate $\alpha^{(k)}$
3. Update $x^{(k+1)} = x^{(k)} + \alpha^{(k)} p^{(k)}$

Until stopping criteria is satisfied

Taylor formula:

$f(x+v) = f(x) + v^T \nabla f(x) = f(x) + \nabla_v f(x)$

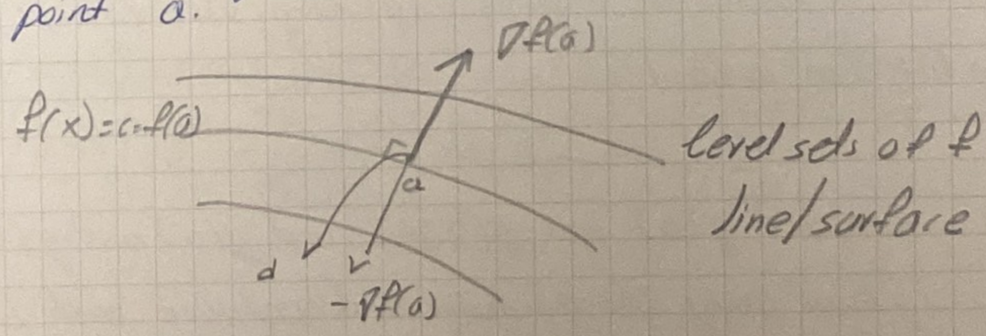
$\hookrightarrow$ directional derivative (want it to be negative)

Theorem:

$f: D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, $a \in D$, $d \in \mathbb{R}^n$ with $\|d\| = 1$, if θ is the angle between d and $\nabla f(a)$ then

$\nabla_d f(a) = d^T \nabla f(a) = \|\nabla f(a)\| \cos\theta$

The vector $-\nabla f(a)$ gives the maximum descent direction of f at the point a.



Stopping criteria/termination conditions:

→

Stopping criterio / termination conditions:

- maximum iterations - repeat until $K \leq K_{max}$
- absolute improvement - repeat until $f(x^{(K)}) - f(x^{(K+1)}) < t_a$
- relative improvement - repeat until $f(x^{(K)}) - f(x^{(K+1)}) < t_r |f(x^{(K)})|$
- gradient magnitude - repeat until $\|\nabla f(x^{(K+1)})\| < t_g$

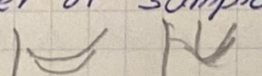
↳ one or more termination conditions can be used

↳ if there are several local minima, one can add random restart with $x^{(1),new}$ sampled randomly from ⊙

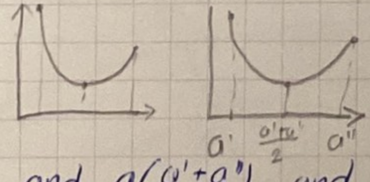
Step size / learning rate:

suppose $x = x^{(K)}$ and $p = p^{(K)}$ is given how to find $d = d^{(K)}$?

methods - Exact line search

- ▷ minimize $f(x + ap)$
  this is a univariate optimization problem for $g(a) = f(x+ap)$
- ▷ find a bracket for the optimal solution (an interval $[d', d'']$ containing d^*)
- ▷ Use univariate optimization methods to find an approximate of d^*:
  - dyadic search - supdivide internall in half at each step
  - fibonacci search - max reduction of interval size for given number of samples
  - quadratic fit search -  (check slides)
  - shubert-pixarski method - assuming g is Lipshitz

    $$|g(x) - g(y)| \leq l |x - y|, \; \forall x, y \in [d', d'']$$

  - bisection method 

    solve $g'(a) = 0$

compare $g'(a)$ and $g(\frac{a'+a''}{2})$ and $g(a'')$ and $g(\frac{a'+a''}{2})$ and throw away the maximum and continue with $[a', \frac{a'+a''}{2}]$

$|a'' - a'| = 1$: after n steps, the size of the interval will be $\frac{1}{2^n}$

→

How many evaluations of g need to be done in order to reduce the size by a factor of n?

$F_{n+2} = F_{n+1} + F_n$, $F_1 = F_2 = 1$ $(1, 1, 2, 3, 5, 8, 13, 21, \dots)$

n steps need to be done:

step 1 - take the interval and divide it proportionally to F_n & F_{n+1}

step 2 - divide proportionally to F_n, F_{n-1}

step 3 - divide proportionally to F_{n-1}, F_{n-2}

...

roughly the final interval will be proportionall to $\left[\frac{a'-a''}{F_n}\right]$

$F_n \sim \frac{1}{\sqrt{5}} \varphi^n \quad \left(\varphi = \frac{1+\sqrt{5}}{2}\right)$

for $\forall a, b, c$ $\exists!$ parabola $(y = Ax^2 + Bx + c)$ passing through

$(a, g(a)), (b, g(b)), (c, g(c))$

quadratic approximation has a high degree of tangency at the minimum

For example: $\ell = \max g'(x)$ & $x \in [a', a'']$

take into account all midpoints at all steps
cost: 2^n evaluations for g
or we can assume g is a unimodal function on $[a', a'']$
↳ there is a unique minimum

$g\,[a', a'']$, g has a minimum, then g has to change its sign

$g'\left(\frac{a'+a''}{2}\right) = 0 \rightarrow \frac{a'+a''}{2}$ candidate for local minimum

$g'\left(\frac{a'+a''}{2}\right) > 0 \rightarrow$ update $[a', a'']$ to $\left[a', \frac{a'+a''}{2}\right]$

$g'\left(\frac{a'+a''}{2}\right) < 0 \rightarrow$ update $[a', a'']$ to $\left[\frac{a'+a''}{2}, a''\right]$

repeat

→

Example:

$f(x_1, x_2, x_3) = \sin(x_1 x_2) + \exp(x_2 + x_3) - x_3$ from $x = [1, 2, 3]$ in the direction of $d = [0, -1, -1]$

minimize $\sin((1+0a)(2-a)) + \exp((2-a)+(3-a)) - (3-a)$

minimize $\sin(2-a) + \exp(5-2a) + a - 3$

minimum is at $a \simeq 3.127$

approximate line search

find $a^{(k)}$ approximately and move on with the descent method

for simplicity $x_k = x^{(k)}$, $p_k = p^{(k)}$, $d_k = d^{(k)}$

conditions: $\varphi(a_k) = f(x_k + a_k p_k) = f(x_k) + c_1 a_k (\nabla f(x_k))^T p_k$, $c_1 \in (0,1)$

↳ sufficient decrease condition

- $\ell(a_k) = f(x_k) + c_1 a_k \nabla f^T(x_k) p_k$ is a linear function
- for small values of $a_k > 0$, we have $\varphi(a_k)^? \ell(a_k)$ this is because $a \in (0,1)$ and $\varphi'(0) = (\nabla f(x_k))^T p_k < a(\nabla f(x_k))^T p_k = \ell'(0) < 0$

$\varphi(a_k) = f(x_k + a_k p_k)$, $\varphi(0) = f(x_k)$

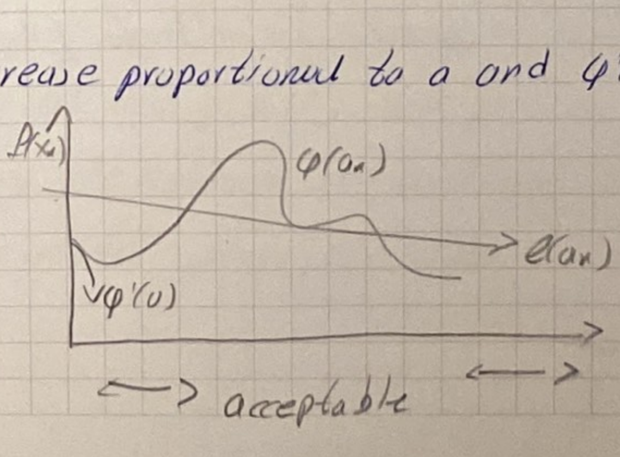
$\ell(a_k) = f(x_k) + a_k (\nabla f(x_k))^T p_k$

$\ell(0) = f(x_k) \Rightarrow \varphi(0) = \ell(0)$

$\varphi'(a_k) = (\nabla f(x_k + a_k p_k))^T p_k = (\nabla f(x_k))^T p_k$

Reminder: we always assume $(\nabla f(x_k))^T p_k < 0$ because we are doing descent

We ask for a decrease proportional to a and $\varphi'(0) = (\nabla f(x_k))^T p_k$
usually $c \approx 0.1$



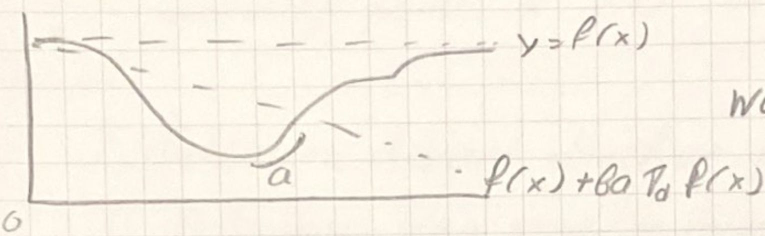
curvature condition:

since the previous is always satisfied for small values of α_k, we need to add further conditions for termination

$\rightarrow (\nabla f(x_k + \alpha_k p_k))^T p_k \geq c_2 (\nabla f(x_k))^T p_k, \quad c_2 \in (0,1)$

if $\varphi'(\alpha_k)$ is not negative enough, we terminate the k step

Known as the Wolfe conditions



Wolfe conditions satisfied

Lemma: Suppose $f: D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be a C^1 function. Let p_k a descent direction at the point $x_k \in D$ and assume f/L_{p_k} is bounded below where $L_{p_k} = \{x \in \mathbb{R}^n \mid x = \; ... \}$

Convergence:

Definition of the process - the election of p_k and α_k we need to study if the process converges somewhere. Let p_k be a descent direction, and let θ_k be the angle of p_k and

$\nabla f(x^*)$, $\cos(\theta_k) = -\frac{1}{\|\nabla f(x_k)\| \|p_k\|} (\nabla f(x_k))^T p_k$

Theorem: assume notation above with p_k a descent direction and α_k satisfying Wolfe's conditions suppose f is C^2 and bounded below in $\mathbb{R}^n$. Then $\sum_{k=0}^{\infty} \cos^2(\theta_k) \|\nabla f(x_k)\| < \infty$