Optimization →

Given $f: \mathbb{R}^n \to \mathbb{R}$. find x s.t. $f(x) \leq f(\hat{x})$ $\forall \hat{x}$ near x.

review from last class

necessary conditions for x to be a minimizer

→ first order condition

$$\nabla f(x) = 0$$

→ second order condition

$\nabla^2 f(x)$ is positive semi-definite

↑

Hessian

a sufficient condition (guarantee x is a local min)

$$\nabla f(x) = 0$$

$\nabla^2 f(x)$ positive definite.

in fact, this $\Rightarrow$ $f(x) < f(\hat{x})$ $\forall \hat{x}$ near x

Algorithms for finding minimizers:

- generate $x_1, x_2, \ldots$ (starting w/ x_0)
- given x_{k-1} find d_k (descent direction)

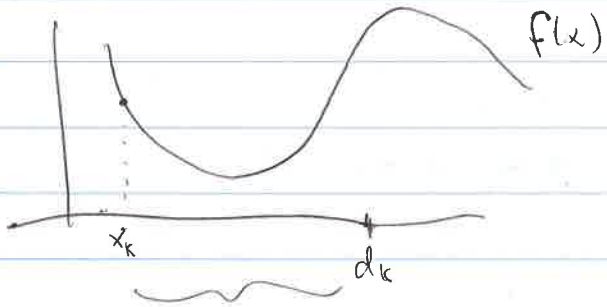
s.t. $(\nabla f)(x_k)^T d_k < 0$

- update $x_{k+1} = x_k + \alpha_k d_k$ for some $\alpha_k > 0$

So, ∃ 2 issues we need to look on:

1. choice of direction d_k
2. choice of scalar α_k.

Let's consider α_k (interpreted as steplength) for a moment: 1-d example.



want to find direction to move that will give us a smaller f° value.

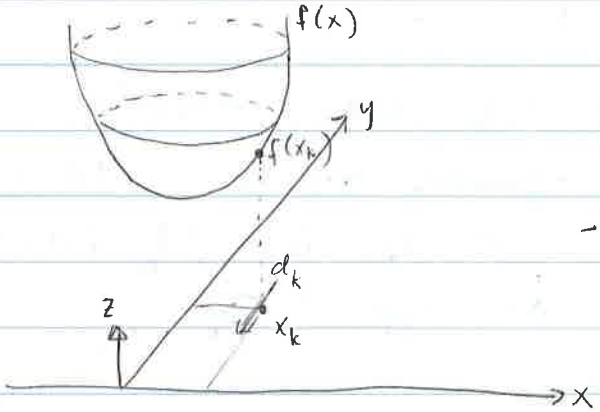
Suppose $|d_k| = 1$. also for x_k as shown,

$$\nabla f(x_k) = f'(x_k) < 0$$

for $d_k = 1$, $f'(x_k)^T d_k < 0$

$f'(x_k)$ $\quad$ 1

2D example

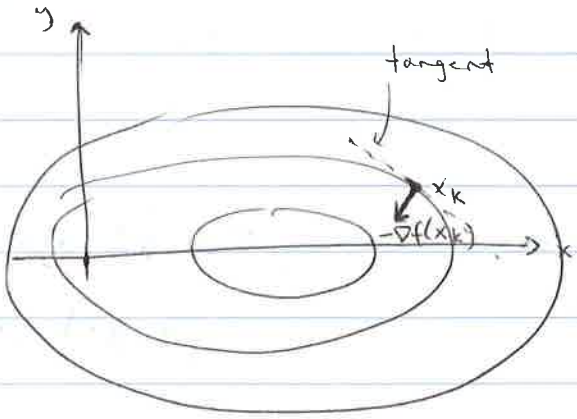


know

$$\nabla f(x_k)^T \cdot (-\nabla f(x_k)) < 0$$

$-\nabla f(x_k)$ is a descent direction.

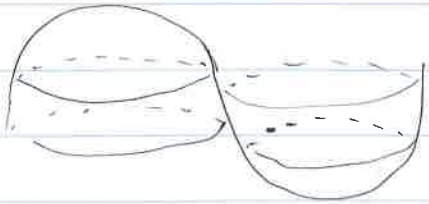
Contours



Contour Lines
$\Rightarrow f(\vec{x}) = \text{constant}$.

on the contour plot: $\nabla f(x_k) \perp$ tangent line to the contour through x_k.

but, suppose it later turned out the bowl in the graph turned over its surface, eg.



then the steplength α_k could make a big difference.

So far, our only tool approach is to choose the negative gradient to go from there. but it's not the best way.

re, Concerning Item #1, one possible descent direction is $d_k = -\nabla f(x_k)$

For another approach, consider the Taylor series for f.

$$f(x_k + d) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k)\, d + O(\|d\|^3) \quad \text{for small } d.$$

call this region of interest $\Phi(d)$ s.t. $\Phi : \mathbb{R}^n \to \mathbb{R}$.

$$\nabla \Phi = (\nabla f)(x_k) + \underbrace{\left[\nabla^2 f(x_k)\right]}_{H_k} d \to \text{set} = 0$$

want to relate this back to the first order condition that $\nabla f(x) = 0$.

$$\Rightarrow d = -H_k^{-1} (\nabla f) x_k$$

(but we never multiply by inverse - instead we solve $H_k d = -\nabla f(x_k)$ for d)

$\nabla^2_d \Phi = H_k$ (b/c $\nabla\Phi$ ~~dependent~~ on is linear f^n of d)

We know that d minimizes Φ if H_k is positive definite.

This strategy is known as Newton's method.

So - do we choose Newton's or steepest descent method?

Newton	Steepest descent
expensive	Cheap
Could lead you astray if H_k is not positive definite	slow
much faster	safe (guaranteed to work)

Strategy: start w/ something safe & then switch. in a specific region of interest.

Convergence note: x_k converges to x with rate r if

$$\lim_{k\to\infty} \frac{\|x - x_{k+1}\|}{\|x - x_k\|^r} = C$$

Constant, independent of k.

ie, Concerning Item #1, one possible descent direction is $d_k = -\nabla f(x_k)$

For another approach, consider the Taylor series for f.

$$f(x_k + d) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k)\, d + O(\|d\|^3) \text{ for small } d.$$

call this region of interest $\Phi(d)$ s.t. $\Phi: \mathbb{R}^n \to \mathbb{R}$.

$$\nabla \Phi = (\nabla f)(x_k) + \underbrace{\left[\nabla^2 f(x_k)\right]}_{H_k} d \to \text{set} = 0$$

want to relate this back to the first order condition that $\nabla f(x) = 0$.

$$\Rightarrow d = -H_k^{-1} (\nabla f) x_k$$

(but we never multiply by inverse - instead we solve $H_k d = -\nabla f(x_k)$ for d)

$\nabla_d^2 \Phi = H_k$ (b/c $\nabla\Phi$ ~~d dependent~~ on is linear f° of d)

We know that d minimizes Φ if H_k is positive definite.

This strategy is known as Newton's method.

So - do we choose newton's or steepest descent method?

newton	steepest descent
expensive	Cheap
Could lead you astray if H_k is <u>not</u> positive definite	slow
much faster	safe (guaranteed to work)

strategy: start w/ something safe & then switch. in a specific region of interest.

(convergence note: x_k converges to x with rate r if $\lim_{k\to\infty} \frac{\|x - x_{k+1}\|}{\|x - x_k\|^r} = C$ constant, independent of k.

$\exists$ 3 types of rates we are concerned about:

$r = 1$ linear convergence, in this case we need $c < 1$ (* true for steepest descent).

$r = 2$ quadratic rate of convergence. don't § need $c < 1$, but we need denominator to not affect c. (* this is true for newton's method)

Suppose $c = 1/2$ for steepest descent.
Suppose $c = 2$ for newton.

$c = 2$ / Newton	← Compare w/ steepest descent much slower convergence
0	$1/4$
1	$2 \cdot 1/16 = 1/8$
2	$2 \cdot 1/64 = 1/32$
3	$2 \cdot 1/32^2 = 2 \cdot 1/1024 = 1/512$
4	$2 \cdot 1/2^{10} = 1/2^9$

One more thing: It's possible to have $1 < r < 2$ - this is superlinear convergence. obtained by combining newton & other methods - "quasi" newton methods. saves some overhead.