Lecture 4

Last week: algorithms for solving 1-d minimization problems e.g. Newton's Method

algorithms for mult-d minimization

Main example for today

Steepest descent: (*) x_{k+1} minimum point of f in the half-line $\{x_k - sd_k : s \geq 0,$ $d_k : \nabla f(x_k)\}$

We'll write $x_{k+1} \in A(x_k)$ when (*) holds

background

Goal: ① Use general theory to show "algorithm works"

② analyze its convergence – When does it work well / badly?

Theorem says: $f: E^n \to \mathbb{R}$ continuous

If A is a point-to-set mapping $E^n \to E^n$ and

① A is closed at x if x is not global minimizer

② If $y \in A(x)$ and X is not global minimum then $f(y) < f(x)$.

and if $x_{k+1} \in A(x_k)$ and (x_k) is bounded then any limit of a convergent subsequence is a minimizer

To use theorem, must check hypothesis ① and ②.

Let's check for method of steepest descent.

Assume f is C^1 and convex.

Check hypothesis ①

Have to check: If $x_k \to x$, $y_k \to A(x_k)$, $y_k \to y$ } then $y \in H(x)$

True because

$y_k = x_k - s_k d_k$, $d_k = \nabla f(x_k)$

s_k minimizes $g_k(s) = f(x_k - sd_k)$

Need to show. $y = x - s^* d$, $d = \nabla f(x)$

s^* minimizes $g(s) = f(x - sd)$

Claim 1: $d_k \to d$ easy because $d_k = \nabla f(x_k)$, $d = \nabla f(x)$, $x_k \to x$ and ∇f is continuous

Claim 2: $s_k \to s^*$ with $y = x - s^* d$

True because $s_k = \dfrac{|y_k - x_k|}{|d_k|} \to \dfrac{|y - x|}{d} = s^*$

since $|d| \neq 0$ (since $\nabla f(x) = 0$ only for global minimizers by convexity)

Claim 3: $y = x - s^* d$

True because $y_k = x_k - s_k d_k$ while $y_k \to y$, $x_k \to x$, $s_k \to s^*$, $d_k \to d$ everything converges

Claim 4: s^* minimizes $g(s) = f(x - sd)$

$g(s^*) = f(x - s^* d) = \lim_k f(x_k - s_k d_k) \le \lim_k f(x_k - s d_k)$ for every s

$= f(x - sd) = g(s)$

so s^* minimizes g.

So far we proved A closed.

Also need: $y \in A(x) \Rightarrow f(y) < f(x)$ (if x is not global min)

Note: in Claim 2 above, definition of s^* was $s^* = \frac{|y-x|}{|d|} = \lim s_k$

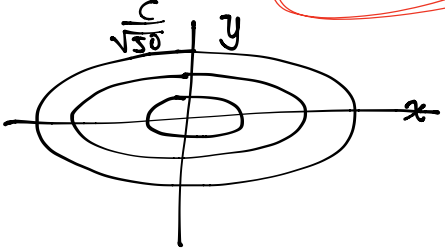
To show $f(y) < f(x)$, enough to check that $f(x - sd) < f(x)$ for some $s > 0$.

True by Taylor Thm

$f(x - sd) = f(x) - s d^T d + ds$ → negligible for small s.

Consider $f(x,y) = x^2/2 + 50y^2$ level curves: $f(x,y) = c^2$

i.e. $\frac{x^2}{2c^2} + \frac{y^2}{(c^2/50)} = 1$



very bad convergence

see below

Initial guess: $(x_0, y_0) = (100, 1)$

$\nabla f(x_0, y_0) = [x_0, 100 y_0] = [100, 100]$

$(x_1, y_1) = (x_0, y_0) - s(100, 100) = (100(1-s), (1-100s)$ where s minimizes: $g(s) = 100^2(1-s)^2/2 + 50(1-100s)^2$

Answer: $s = 2/101$

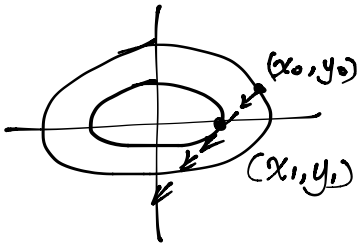
$(x_1, y_1) = (100 - 200/101, 1 - 200/101) = (98 \cdot 2/101, -99/101) = \frac{99}{101}(100, -1)$

Repeat to find: $(x_k, y_k) = (\frac{99}{101})^k (100, 1)$ ... k even

$(\frac{99}{101})^k (100, -1)$ ... k odd

e.g. $(x_{100}, y_{100}) \approx (13.5, .135)$ still far from 0.

How to understand?



Note: $d_0 = \nabla f(x_0, y_0)$ tangent to level curve of f at (x_1, y_1)

$d_1 = \nabla f(x_1, y_1)$ is orthogonal to the level curve of f at (x_0, y_0)

So for steepest descent in 2d, each direction is orthogonal to previous direction.

Facts about convergence

① Suppose f is quadratic: $f(x) = \frac{1}{2} x^T Q x - b^T x + c$

Then: $f(x_{k+1}) - f_{min} \leq \left(\frac{1-r}{1+r}\right)^2 (f(x_k) - f_{min})$

$Q = \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix}$ ↑

$r = \frac{A}{a}$ Condition number

a = smallest eigenvalue of Q

A = largest eigenvalue of Q

Note ~~$0 \leq r \leq 1$~~: $r \geq 1$ at least 1

i.e. ~~good convergence if $r=1$~~

~~bad convergence if $r=0$~~

in our example, $r = 100$, $\frac{1-r}{1+r} = \frac{99}{101}$

Similarly: for more general convex functions, things are similar.

We'll prove these facts later. First some applications

① Suppose I want to minimize convex function f.

2 possible procedures:

a. minimize w/ steepest descent or

b. minimize $F(x) = |\nabla f(x)|^2$ with steepest descent.

(since x^* minimizer $\Leftrightarrow \nabla f(x) = 0 \Leftrightarrow x^*$ minimizes f.)

Which is better?

For simplicity. assume $f(x) = \frac{1}{2} x^T Q x - b^T x$

$\nabla f(x) = x^T Q - b^T$

$F(x) = |\nabla f(x)|^2 = (x^T Q - b^T)(Qx - b) = x^T Q^2 x - 2 b^T Q x + b^T b$

Which is easier to minimize via steepest descent?

i.e. which one has bigger condition number

Q vs Q^2

for f for $|\nabla f|^2$

Recall: eigenvalues for Q^2 = (eigenvalues for Q)^2

so $r_{Q^2} = (r_Q)^2$ where r_S = cond # for S.

so $r_Q \geq r_{Q^2}$

Conclusion: better to stick with minimizing f.

Example: Suppose I want to solve:

minimize $f(x)$ with constraint $h(x) = 0$

One approach: minimize $f(x) + \frac{1}{2} \mu h(x)^2$

for μ very large, then should happen that h close to 0 at minimum point (so constraint nearly satisfied)

Can we do this well with steepest descent?

For simplicity, assume

$f(x) = \frac{1}{2} x^T Q x + b^T x$ (quardratic)

$h(x) = c^T x$ (linear)

$h(x)^2 = x^T c c^T x$

so $f(x) + \frac{1}{2} \mu h(x)^2 = \frac{1}{2} x^T (Q + \mu c c^T) x - b^T x$

μ huge what is condition number at $Q + \mu c c^T$?

we guess: condition number is tiny if μ huge.

Conclusion is: ~~steepest~~ descent probably will not work well.

Next convergence analysis
But first, why do we care?
i.e. to minimize $f(x)=\frac{1}{2}x^TQx-b^Tx$
why don't I just find minimizer?
That is: $\nabla f = x^TQ - b^T$
So $\nabla f(x^*)=0$ when $(x^*)^TQ=b^T$ i.e. $\boxed{x^*=Q^{-1}b}$

Why bother?
① to gain in sight
② Hard to compute Q^{-1}

Now: convergence analysis for steepest descent, when $f(x)=\frac{1}{2}x^TQx-b^Tx$
① Given x find formula for $y=A(x)$ = minimizer of f on half-line $\{x-sd : s>0$
$d=\nabla f(x)\}$

But first simplify: Since $x^*=Q^{-1}b$ i.e. $b=Qx^*$

$$f(x)=\frac{1}{2}x^TQx-(x^*)^TQx=\underbrace{\frac{1}{2}(x-x^*)^TQ(x-x^*)}_{\text{call this } E(x)}\underbrace{-\frac{1}{2}x^*Qx^*}_{\text{a number of Independent of } x}$$

So minimizing f = minimizing E
① Formula for $y=A(x)$ $d=(\nabla E)^T=Q(x-x^*)$
$y=x-s\cdot d$ for optimal choice of s.

$$E(x-s\cdot d)=\frac{1}{2}(x-x^*-s\cdot d)^TQ(x-x^*-s\cdot d)=\frac{s^2}{2}(\quad)+s(\quad)$$
$$=\frac{s^2}{2}(d^TQd)-s(x-x^*)Qd+(\text{independent of } s)$$
$$=g(s)$$
$$g'(s)=s\cdot d^TQd-(x-x^*)^TQd$$
$$g'(s)=0 \text{ if } s=\frac{(x-x^*)^TQd}{d^TQd}=\boxed{\frac{d^Td}{d^TQd}}$$

We have used: $\nabla E(x)=\nabla f(x)=x^TQ-b^T=(x-x^*)^TQ$
So $d=(\nabla E)^T=Q(x-x^*)$
Thus $y=x-s^*d=x-\left(\frac{d^Td}{d^TQd}\right)d$ ← This is $A(x)$

$d=\nabla E(x)^T$
$=\nabla f(x)^T$
$=ax-b$

② Formula for $E(y)$ compared to $E(x)$
$y=A(x)$
Idea: just plug into definition of E & compute:
This leads to:

$$E(y)=\underbrace{\left(1-\frac{(d^Td)^2}{(d^TQd)(d^TQ^{-1}d)}\right)}E(x)$$

happy if this is small $\ll 1$

③ I claim that $1-\frac{(d^Td)^2}{(d^TQd)(d^TQ^{-1}d)}\le\left(\frac{1-r}{1+r}\right)^2=\left(\frac{A-a}{A+a}\right)^2=1-\frac{4aA}{(A+a)^2}$

$r=\frac{a}{A}$ = condition number

I need to show that

$$\frac{(d^Td)^2}{(d^TQd)(d^TQ^{-1}d)} \geq \frac{4aA}{(A+a)^2}$$

a) divide numerator and denominator by $(d^Td)^2 = |d|^4$

Let $y = \frac{d}{|d|}$ = unit vector

Then left-hand side is $\frac{1}{(y^TQy)(y^TQ^{-1}y)}$

Let $w_1, \dots, w_n$ orthonormal basis of eigenvectors for Q.

Any unit vector y can be written $y = a_1w_1 + \cdots + a_nw_n$ where $1 = |y|^2 = y^Ty = a_1^2 + \cdots + a_n^2$

$= \theta_1 + \cdots + \theta_n$

$\theta_i = a_i^2$

Then substitute & expand $Qw = \lambda w$

$$\frac{1}{(y^TQy)(y^TQ^{-1}y)} = \frac{1}{(\sum_{i=1}^n \theta_i\lambda_i)(\sum_{i=1}^n \theta_i/\lambda_i)} = \frac{\psi(\sum \theta_i\lambda_i)}{\sum \theta\, \psi(\lambda_i)} \quad \psi(s) = \frac{1}{s}$$

The point is: geometric reasoning tells us that smallest value of $\frac{1}{(y^TQy)(y^TQ^{-1}y)}$ for $|y|=1$ occurs when

$\theta_1, \dots, \theta_n$ nonzero

$\theta_2, \dots, \theta_{n-1} = 0$ i.e. only smallest + largest eigenvalues "archive"

if $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = A$

i.e. minimum of $\frac{(d^Td)^2}{(d^TQd)(d^TQ^{-1}d)}$ occurs for θ minimizing

$$\frac{1}{(\theta a + (1-\theta)A)(\frac{\theta}{a} + \frac{1-\theta}{A})} \qquad 0 \leq \theta \leq 1$$

In fact, this is minimized if $\theta = \frac{1}{2}$ This finally leads to our formula.

$$E(y) \leq \left(\frac{1-r}{1+r}\right)^2 E(x) \quad \text{if } y = A(x)$$

One other fact: Suppose f is not quadratic, but $\boxed{aI \leq \nabla^2 f(x) \leq AI \text{ for all } x}$

Where $Q \leq R$ means $V^TQV \leq V^TRV$ for all V

So $Q \leq AI$ means $V^TQV \leq V^T(AI)V = A(V)^2$

(Then $\frac{a}{A}$ is analog of condition number)

When this holds, $[f(x_{k+1}) - f_{min}] \leq (1 - \frac{a}{A})[f(x_k) - f_{min}] = (1-r)[f(x_k) - f_{min}]$ for method of steepest descent.