Lecture 9: Minimizing a Function Step by Step

Agenda

- Optimization

Source: Sections VI.3 and VI.4 in Linear Algebra and Learning from Data (2019) by **Gilbert Strang**

Optimization

Optimization-Basic facts of three terms of Taylor series

$F(x+\Delta x) \simeq F(x) + \Delta(x)\,\frac{dF}{dx} + \frac{1}{2}(\Delta(x))^2\,\frac{d^2F}{dx^2}$ ← F as a one function, and x as a one variable

$$F(x+\Delta x) \simeq F(x) + \Delta(x)^T \nabla F(x) + \frac{1}{2}\Delta(x)^T H(\Delta x)$$

$x=\{x_1,..,x_n)$

$$\nabla F(x) = \begin{pmatrix} \frac{\partial F}{\partial x_1} \\ \frac{\partial F}{\partial x_n} \end{pmatrix}$$

Hessian $H_{jk} = \frac{\partial^2 F}{\partial x_j \partial x_k} = H_{kj}$

Key fact is it's symmetric

← Change if F(x) is a function of n variables of x where $x=\{x_1,..,x_n)$. And the derivative of F w.r.t x_1, the derivative of F w.r.t x_2, and so on. These are vectors too.

So $\Delta(x)^T$ times dF/dx, this dF/dx replacing this by all the derivatives, and it's the gradient. So the gradient of F at x, $\nabla F(x)$ is partial derivatives of F w.r.t x_1 to x_n.

1/2 times $\Delta(x)$ is a vector and we have $\Delta(x)^T$ and 2nd derivatives function of n variables, called $H(\Delta x)$

It's an approximation. If n is very large, and we've a function of many variables.

Then, we had n derivatives to compute here, and about ½ times n^2 derivatives.

Computing the gradient is feasible if n is small or moderately large.

Optimization

$f=(f_1(x), f_2(x), \ldots, f_n(x))$,

$x=\{x_1,..,x_n)$

$f= \Delta(F)$ → nf_n s and n variables

$F(x+\Delta x)=f(x)+J\ \Delta x$

$J_{jk}=\frac{\partial f_j}{\partial x_k}$ ← Jacobain matrix, J_{jk} is the derivative of the J function w.r.t kth variable

That's the background what we discussed from first slide up to the above in this slide

Look at Optimization

Minimize F(x)

Solve f=0 means $f_1=0, \ldots, f_n=0$ (n eqns, n unknowns)

Start with Newton's method to solve these n equations and n unknowns

Application of gradients in Jacobians:: $F(x+\Delta x)=f(x)+J\ \Delta x$

Solve $f=0 \Rightarrow 0= f(x_k) + J(x_k)\ (x_{k+1}-x_k)$

$\mathbf{x_{k+1}=x_k - J(x_k)^{-1} f(x_k)}$ ←That's system of equations

Write down Newton's method for minimizing a function

Optimization

Suppose, Solve $f(x)=x^2-9=0 \Rightarrow \frac{\partial f}{\partial x_k} = 2x_k$

Suppose $f(x)=x^2-9=0 \Rightarrow x=\pm 3$, we take $x_k=3$

Putting in $x_{k+1}=x_k - J(x_k)^{-1} f(x_k) \Rightarrow x_{k+1}= x_k - \frac{1}{2x_k}(x_k^2-9)$

$$\Rightarrow x_{k+1}= \frac{1}{2} x_k + \frac{9}{2}\frac{1}{x_k}$$

$$\Rightarrow x_{k+1}= \frac{1}{2} * 3 + \frac{9}{2} * \frac{1}{3} \text{ as } x_k=3$$

$$= \frac{6}{2} = 3$$

So we've checked that the method is consistent which just means we kept the algebra straight

Important point about Newton's method is to discover how fast it converges. So now let us do

$$x_{k+1}-3= \frac{1}{2} x_k + \frac{9}{2}\frac{1}{x_k} -3 \quad \Rightarrow x_{k+1}-3= \frac{1}{2} x_k + \frac{9}{2}\frac{1}{x_k} -3 \quad \Rightarrow (x_{k+1}-3)= \frac{1}{x_k}\left[\frac{9}{2} + \frac{1}{2} x_k^2 -3x_k\right] = \frac{1}{2x_k}[9+ x_k^2-6x_k]$$

$$\Rightarrow (x_{k+1}-3)= \frac{1}{2x_k}[x_k-3]^2$$

A similar equation sort of centered at minus 3. They're all the starting points that approach 3, and that equation is with quadratic convergence the error being squared at every step. It zooms in on 3

Optimization

Convert all those equations over to Newton's method for optimization

Minimize F(x) (≈ solving $\nabla F=0$) At a minimum, finding a point where all the first derivatives are 0.
This ∇F in here is small f in Newton’s method

This is sort of the heart of our applications to deep learning-- we have very complicated loss functions to minimize, functions of thousands or hundreds of thousands of variables. So that means to use Newton's method, but often we can't. So we need to put down here two methods

Method I- steepest descent

$x_{k+1}=x_k - s_k \nabla F$ s --> Step size or learning rate

(Basic for large search job)

Steepest descent means that we move in the steepest direction which is the direction of the gradient of F. we move some distance, and decide what that distance should be. So this is a step size, s, or in deep learning, it's learning rate

Like to choose s_K so that minimize F. You take the point on this line, so this a line in R^n, a direction in R^n.
we exact line search would be, exact line search is the best s

Optimization

Method II- Newton’s method

$x_{k+1}=x_k - H^{-1}(\nabla F)$

(Parallel to Jacobian)

These two methods on the board parallel to each other, because we keep straight, are you solving equations (Newton’s method of solving equations), or are you minimizing functions (here in Newton’s method of optimization)

f= ∇F; Jacobian of the gradient is the Hessian because the first derivative of the first derivative is the second derivative

Do they converge? What rate do they converge?

1. Convergence rate for Newton's method will be quadratic. The error gets squared, and of course, that means super-fast convergence, if you start and close enough

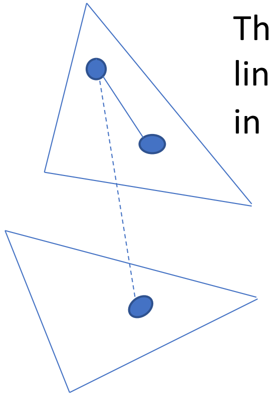
2. Rate of convergence for a steepest descent is not. You're not squaring errors here. So a linear rate of convergence would be right

Back-up Slides

Optimization

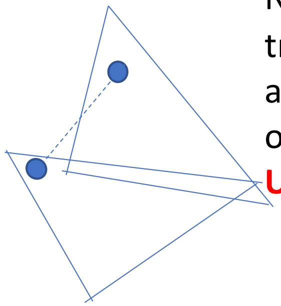
Convex Set K

Convex Set



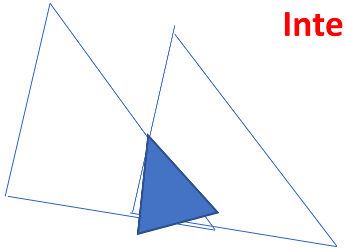
That is a convex set, because line between x_1 and x_2 stays in the set

that point and that point, of course, it goes outside, Not Convex



Lower triangle, overlaps upper triangle? Is that a convex set? No. Why how do we see that the union of those two triangles is not a convex set? Take one point from that corner and one from that corner, and the line between them went outside.

Union is usually not convex



This color part is convex.

Intersection is always convex

Proof that the intersection is convex:

x_1 in $K_1 \cap K_2$ and x_2 in $K_1 \cap K_2$

Line x_1 to x_2 in $K_1 \cap K_2$

So, **Intersection of sets is always convex**

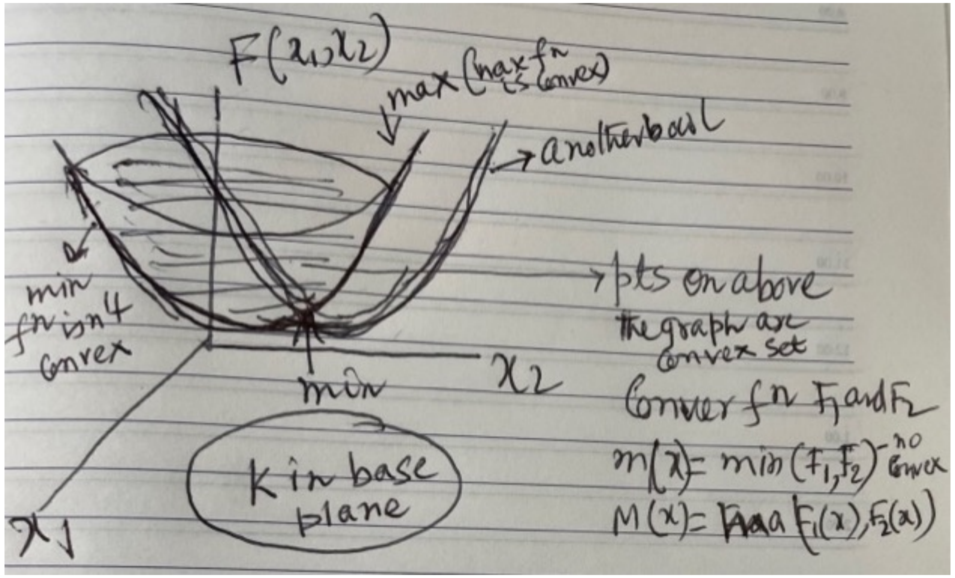
Optimization

Convexity

Convex Function F(x)
Convex Set K
Convex Minimization: F(x)
for x in K

Exp.: K: Ax=b
Subspace made in
called affine

Minimizing it over certain x’s, not all x's. K might be the set where Ax=b. K might be, a subspace or a shifted subspace



pts on above the graph are convex set

Convex fn F_1 and F_2

$m(x) = \min(F_1, F_2)$ - no convex

$M(x) = \max(F_1(x), F_2(x))$

The convex sets the constraint, so this is the constraint set. Constraint is that x must be in set K. Here was an example where it's flat plane.

The function is convex, and draw a convex function, graph, say a bowl. So that's a graph of F of x, and then here are the x’s. Let us put x_1 and x_2 in the base and the graph of $F(x_1,x_2)$ up here

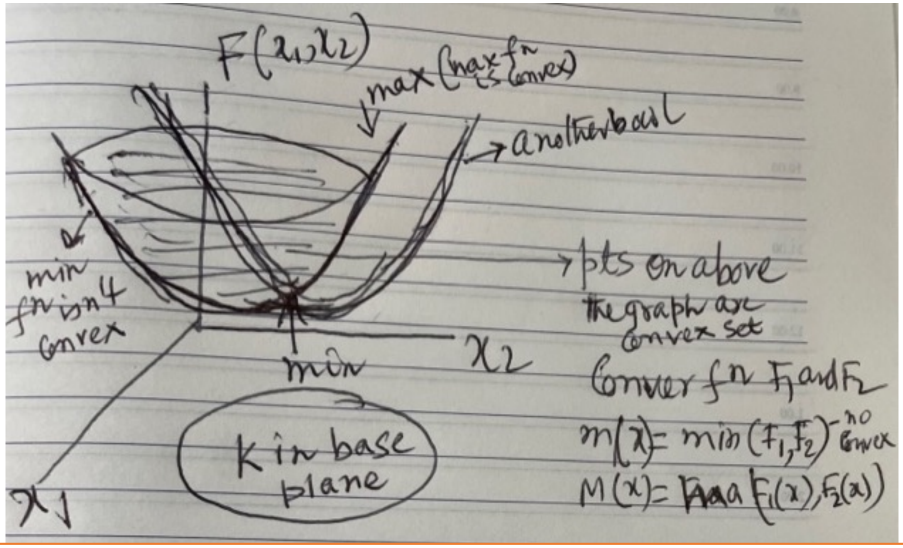
Optimization

Our Prototype

Convex Function F(x)= points above the graph are convex set

Convex Set K

Convex Minimization: F(x) for x in K



Two convex functions, F_1 and F_2. Create the min or the max. m(x)=min (F_1, F_2) and M(x)= max (F_1(x), F_2(x)) at x

We've a bowl and we've another bowl, and suppose they're both convex. We consider the minimum of those two functions and also the maximum of those two functions. One of these will be convex, and the other won't.

The minimum is this point until they meet somehow on some surface and then this min (see figure). Is that convex? Absolutely no. It's got this bad kink in it. So the maximum is the one that is above, all the points or things that are above or on (see figure)

There is the maximum function. That was the minimum function. It had a kink. The maximum function is like that, and it is convex, so maximum yes, minimum no.

Optimization

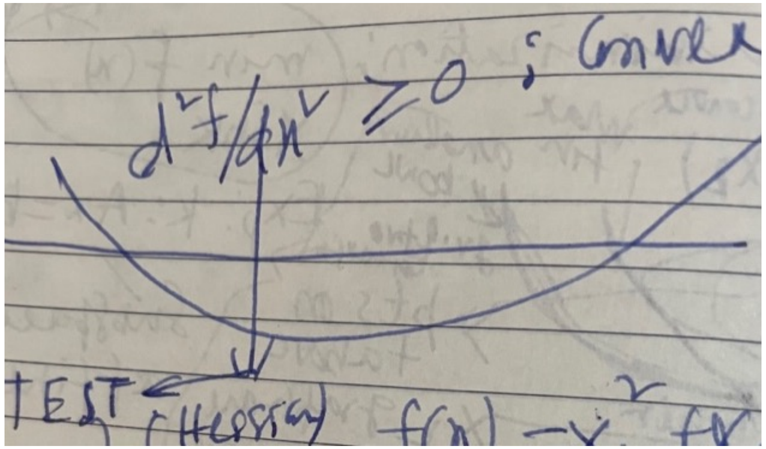
Definition of smooth convex function

Function *f(x)* is convex if $\frac{d^2f}{dx^2} \geq 0$. Reason:

The slope $\frac{d^2f}{dx^2}$ is increasing. *The curve bends upward*

(like the parabola $f = x^2$ with second derivative $\frac{d^2f}{dx^2} = 2$).

The extension to *n* variables involves the *n* by *n* matrix *H* (x) of second derivatives. If *F* (*x)* is a smooth function then there is an almost perfect test for convexity



F(x$_1$*, ... , x*$_n$*)* is convex if and only if its second derivative matrix *H(x)* is positive semidefinite at all *x.*

A quadratic $F = \frac{1}{2}x^T Sx$ has gradient *Sx* . Its symmetric second derivative matrix is *S.*

Above its graph is a bowl, when S is positive definite. This function *F* is strictly convex.