Lecture 10:Gradient Descent: Downhill to a Minimum

Agenda

- Gradient Descent: Downhill to a Minimum

Source: Sections VI.4 in Linear Algebra and Learning from Data (2019) by **Gilbert Strang and Section 9.3 from Convex Optimization** by **Stephen Boyd and Lieven Vandenberghe**

Gradient Descent

Gradient descent is central algorithm of neural net deep learning, machine learning, and optimization in general. So we minimize a function. And if there are many, many variables, too many to take second derivatives, then we settle for first derivatives of the function

$x_{k+1}=x_k + s_k \nabla f(x_k)$

$s_k \rightarrow$ learning rate

We see gradient and Hessian and the role of convexity before we see the crucial example:

$F=\frac{1}{2} X^T S X = \frac{1}{2}(x^2+by^2)$ The function is a pure quadratic of two unknowns, x and y. Write symmetric matrix S

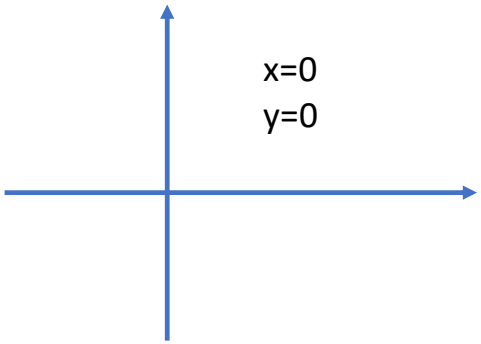
$$S=\begin{pmatrix} 1 & 0 \\ 0 & b \end{pmatrix}$$

Eigen values 1 and b on the diagonal. Condition number is the question of the speed of convergence. That is $\lambda_{max}/\lambda_{min}=1/b$.

In this case, the largest is 1 the smallest is b. So that's 1 over b. And when b is a very small number, then that's when we're in trouble

Gradient Descent

Steps of steepest descent



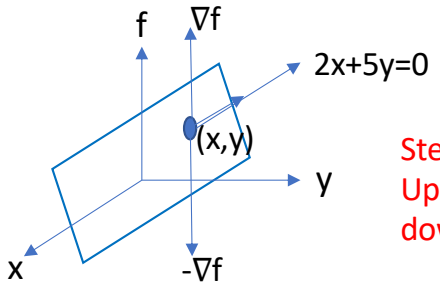
We just have a bowl sitting on the origin. Minimum point is x=0, y= 0. Question will be how quickly do we get to that one

Let us see the actual steps of steepest descent. How quickly they converge to x^*, the answer, the place where this thing is a minimum.

Gradient Hessian of F and Convexity of F

Take example after some general thoughts about gradients, Hessians.

Take an example of the gradient



Steepest Up and down

$f(x,y)=2x+5y$

$\nabla f = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$ and $H=\nabla^2 f=0=\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

Surface f=Constant
∇f=0 on surface
What about H= $\nabla^2 f$

H= $\nabla^2 f$=0 because the surface is flat. The surface was convex upwards from-- if it was a convex or a graph of F Hessian would be? What's the connection between Hessian and convexity of the function?

Gradient Descent

Convexity → H positive semi definite (includes pos def)

Strict Convexity → H positive semi definite

Semi-definite for convex. So, the linear function is convex, but not strictly convex. Strictly means it really bends upwards. The Hessian is positive definite. The curvatures are positive

Take an example: $f(x)=\frac{1}{2} X^T S X - a^T x - b$ ←We have linear terms $-a^T x$, and constant b

H=S. So this function is strictly convex when S is positive definite, because H is now S for that function, for that function H. Usually H is varying from point to point. The nice thing about a pure quadratic is its constant. It's the same S at all points

Differentiating all n vectors at once i.e., whole vector of first derivatives. Because here whole function with x for vector x. We take n =1

$\nabla f = S^T x - a$ And $S^T x - a = 0 \Rightarrow x = S^{-1} a$

$H = \nabla^2 f = S$

Minimum value of f(x) is f_{min} to compute at $x=S^{-1}a$ =argmin (x) point where f= f_{min}

argmin (x) is the argument that minimizes the function

argmin $f=x^*=(0,0)$

$x^*=(0,0)$

Gradient Descent

Gradient Descent $x_{k+1}=x_k - s_k \nabla f(x_k)$

Input some decision making is a step size, the learning rate. We can take it as constant. If we take too big a learning rate, the thing will oscillate all over the place and it's a disaster. If we take too small a learning rate, what's the matter with that

Optimize s_k i.e., estimate suitable s_k:

Exact line search: Choose s_k to make $f(x_{k+1})$ a min in the line search direction, which is given by gradient

An s_k-- estimate a suitable s_k, and then go with it for a while. And then look back to see if it was too big, they'll see oscillations. It'll be bouncing all over the place

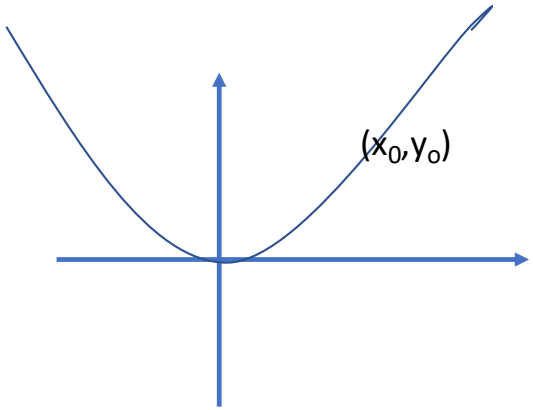
In this example $F=\frac{1}{2} X^T S X = \frac{1}{2}(x^2+by^2)$, where we will do exact line searches that for a small value of b, it's extremely slow, that the condition number controls the speed. So an exact line search would be that. So backtracking line search– backtracking would be taken a fixed s=1.

Backtracking: $s_0, \frac{1}{2}s_0, \frac{1}{4}s_0$,.... so on until you're satisfied with that step

Fundamental question. On an exact line search, how much does that reduce the function

Reduction involves m/M

Gradient Descent



Take a point, x_0,y_0 on the surface and iterate. New point:

$$\begin{pmatrix} x \\ y \end{pmatrix}_{K+1} = \begin{pmatrix} x \\ y \end{pmatrix}_{K} - s_k \begin{pmatrix} 2x \\ 2by \end{pmatrix}_{K}$$

Choose best s_k

Start $(x_0,y_0) = (b,1)$

$x_k=b\left(\frac{b-1}{b+1}\right)^k \quad y_k=b\left(\frac{1-b}{1+b}\right)^k$

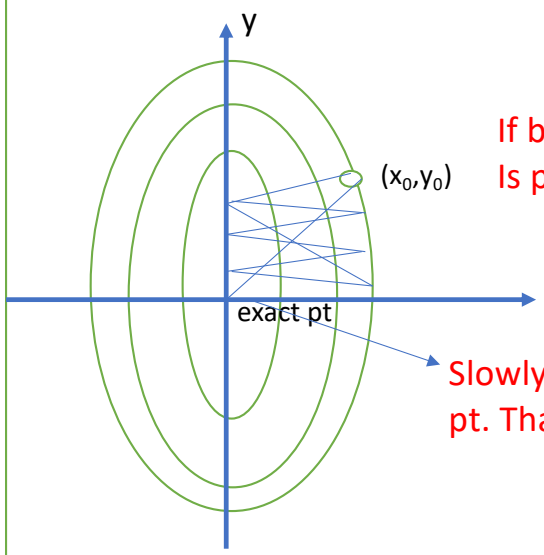
$f_k=\left(\frac{1-b}{1+b}\right)^{2k} f_0$

Taken from the book by Steven Boyd and Vandenberghe – *Convex Optimization* (See Appendix 1)

$\frac{1-b}{1+b}$ is crucial, b near to 1 converges quickly and b is near to 0, i.e., b is small then it is a hard case

Source: Page 469, Convex optimization- Boyd and Vandenberghe

Gradient Descent



If b is small, search
Is perpendicular

Slowly we reach exact pt. That depends on s_k

$f=x^2+by^2=\text{constant}=c$
If b is small

What kind of a curve is that? That's an ellipse. Because there is no xy term, that ellipse is like, well lined up with the axes.

First search direction? What direction do we move from x_0, y_0? Do we move along the ellipse? Because along the ellipse f is constant. The gradient direction is perpendicular to the ellipse

See for the curve, the track of the Zigzag

We would like to get here, but we're not aimed here at all. So we zig, zig, zig zag, and very slowly approach that point. And how slowly? With that multiplier, 1-b over 1+b

So that will be the critical quantity. And we see it in this ratio 1 minus b over 1 plus b. So if b is 100, this is 0.99 over 1.01. It's virtually 1

Appendix 1

A quadratic problem in $\mathbf{R}^2$

Our first example is very simple. We consider the quadratic objective function on $\mathbf{R}^2$

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2),$$

where $\gamma > 0$. Clearly, the optimal point is $x^\star = 0$, and the optimal value is 0. The Hessian of f is constant, and has eigenvalues 1 and γ, so the condition numbers of the sublevel sets of f are all exactly

$$\frac{\max\{1,\gamma\}}{\min\{1,\gamma\}} = \max\{\gamma, 1/\gamma\}.$$

The tightest choices for the strong convexity constants m and M are

$$m = \min\{1,\gamma\}, \qquad M = \max\{1,\gamma\}.$$

We apply the gradient descent method with exact line search, starting at the point $x^{(0)} = (\gamma, 1)$. In this case we can derive the following closed-form expressions for the iterates $x^{(k)}$ and their function values (exercise 9.6):

$$x_1^{(k)} = \gamma\left(\frac{\gamma-1}{\gamma+1}\right)^k, \qquad x_2^{(k)} = \left(-\frac{\gamma-1}{\gamma+1}\right)^k,$$

and

$$f(x^{(k)}) = \frac{\gamma(\gamma+1)}{2}\left(\frac{\gamma-1}{\gamma+1}\right)^{2k} = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f(x^{(0)}).$$

Appendix 1

Quadratic problem in $\mathbf{R}^2$. Verify the expressions for the iterates $x^{(k)}$ in the first example of §9.3.2.

Solution. For $k = 0$, we get the starting point $x^{(0)} = (\gamma, 1)$.

The gradient at $x^{(k)}$ is $(x_1^{(k)}, \gamma x_2^{(k)})$, so we get

$$x^{(k)} - t\nabla f(x^{(k)}) = \begin{bmatrix} (1-t)x_1^{(k)} \\ (1-\gamma t)x_2^{(k)} \end{bmatrix} = \left(\frac{\gamma-1}{\gamma+1}\right)^k \begin{bmatrix} (1-t)\gamma \\ (1-\gamma t)(-1)^k \end{bmatrix}$$

and

$$f(x^{(k)} - t\nabla f(x^{(k)})) = (\gamma^2(1-t)^2 + \gamma(1-\gamma t)^2)\left(\frac{\gamma-1}{\gamma+1}\right)^{2k}.$$

This is minimized by $t = 2/(1+\gamma)$, so we have

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - t\nabla f(x^{(k)}) \\ &= \begin{bmatrix} (1-t)x_1^{(k)} \\ (1-\gamma t)\gamma x_2^{(k)} \end{bmatrix} \\ &= \left(\frac{\gamma-1}{\gamma+1}\right)\begin{bmatrix} x_1^{(k)} \\ -x_2^{(k)} \end{bmatrix} \\ &= \left(\frac{\gamma-1}{\gamma+1}\right)^{k+1}\begin{bmatrix} \gamma \\ (-1)^k \end{bmatrix}. \end{aligned}$$

Back-up Slides

Gradient Descent

Fundamental minimization question, to minimize a quadratic

$f(x)=\frac{1}{2}x^TSx-a^Tx$

$\nabla f=Sx-a$

$Sx-a=0 \Rightarrow x=S^{-1}a$

Looking at the function f(x). Let us to remove b, because that just shifts the function by b

To plug in $x=S^{-1}a$ in $\frac{1}{2}x^TSx - a^Tx \Rightarrow \frac{1}{2}S^{-1}a^T SS^{-1}a - a^T S^{-1}a \Rightarrow \frac{1}{2}a^TSa - a^T S^{-1}a \Rightarrow -\frac{1}{2}\mathbf{a^T S^{-1}a}$

That's what a quadratic bowl, a perfect quadratic problem minimizes to that's its lowest level.

Remarkable Convex Function

$f(x) = - \log (\det X)$

Take the determinant of the matrix. That's clearly a function of all the n^2 variables.
Then you take the log of the determinant and put in a minus sign because we want convex.
That turns out to be a convex function

Function of a matrix, n^2 variables x_{ij}

∇f= entries of X^{-1}

Where did that come from? It might be minus the entries. So we've got n^2 function-- typical entry in x inverse?

$$X_{ij}^{-1} = \frac{\text{derivatives of determinant}}{\det X}$$

Because when you take derivatives of a log, that will put determinant of x in the denominator. And then the numerator will be the derivatives of the determinant of x

Gradient Descent

$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nn} \end{pmatrix}$

That one smaller determinant that we are throwing away the first row and first column. It's called a Minor

Find $X_{11}^{-1} = \frac{\partial detX}{\partial x_{11}}$ = what? We need derivatives of the determinant ∂x_{11}. what's the change in the determinant. We have formula for determinants like

Cofactor expansion det X= x_{11} ( )+x_{12} ( ) +..+x_{1n} ( )

Something

And what is something ? x_{11} multiply when you compute determinants? x_{11} will not multiply any other elements in its row, because you're never multiplying two x's in the same row or the same column. What x_{11} is multiplying all these elements. And in fact, it turns out to be is the determinant

Minor as this smaller matrix, and the co-factor, which is the determinant of the minor.

Minor cofactor (c)=$\pm$det (minor)

So it's the co- factor. Let us call it c_{11}.

Cofactor expansion det X= x_{11} (c_{11})+x_{12} (c_{12}) +....

$$X_{11}^{-1} = \frac{c_{11}}{\det X}$$