CSE 592: Convex Optimization, Assignment 1

111447198: Nishant Borude

20th February 2018

1 Gradient Decent without Linesearch

1.1 Number of iterations:

Given $f(x)$ is a strongly convex twice-continuously differentiable function. The bounds for the Hessian are $mI \leq \nabla^2 f(x) \leq MI$. Consider the lower bound first. We know that the second order Taylor series expansion of $f(x)$ is:

$$f(y) \geq f(x_t) + \nabla f(x_t)^T (y - x_t) + \frac{m}{2}\|y - x_t\|^2$$

Our objective for the descent is to find the optimal value of $f(x_t)$. Therefore minimizing over the right hand side of the equation for finding y we get:

$$min_y f(y) + \nabla f(y)^T (x_t - y) + \frac{m}{2}\|x_t - y\|^2$$

Take the gradient w.r.t y and equate to 0.
Therefore,

$$\nabla f(x_t) + m(y - x_t) = 0$$

$$y^* = x_t - \frac{1}{m}\nabla f(x_t)$$

For optimal conditions, we can replace $f(y)$ by p^* and substituting the values, we get

$$p^* \geq f(x_t) + \nabla f(x_t)^T \Big(- \frac{1}{m}\nabla f(x_t)\Big) + \frac{m}{2}\Big\|\frac{1}{m}\nabla f(x_t)\Big\|^2$$

$$p^* \geq f(x_t) + \|\nabla f(x_t)\|^2 \Big(- \frac{1}{m} + \frac{1}{2m}\Big)$$

$$p^* \geq f(x_t) - \frac{\|\nabla f(x_t)\|^2}{2m}$$

$$f(x_t) - p^* \leq \frac{1}{2m}\|\nabla f(x_t)\|^2 \leq \epsilon \tag{1}$$

Similarly for M strongly smooth function we can write:

$$f(x + p) \leq f(x) + \nabla f(x)^T p + \frac{M}{2}\|p\|^2$$

To solve this equation, we find the minimization of right hand side w.r.t to η. But in this case as $\eta = \frac{1}{M}$ is fixed, we can solve by substituting $p = -\eta \nabla f(x)$ as the gradient descent step.
Therefore, the next $f(x_{t+1})$ is $f(x_t - \eta \nabla f(x_t))$

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t)^T (-\eta \nabla f(x_t)) + \frac{M}{2}\|\eta \nabla f(x_t)\|^2$$

$$f(x_{t+1}) \leq f(x_t) + \|\nabla f(x_t)\|^2 \Big(- \frac{1}{M} + \frac{M}{2}\frac{1}{M^2}\Big)$$

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2M}\|\nabla f(x_t)\|^2$$

From equation (1) we can substitute the value for $\|\nabla f(x_t)\|^2$ and $f(x_t)$ becomes p^*.

$$f(x_{t+1}) \leq f(x_t) - \frac{2m}{2M}(f(x_t) - p^*)$$

$$f(x_{t+1}) - p^* \leq (f(x_t) - p^*)\left(1 - \frac{m}{M}\right)$$

By condition of strong convexity and simplifying further, we get

$$f(x_{t+1}) - p^* \leq (f(x_0) - p^*)\left(1 - \frac{m}{M}\right)^{t+1} \leq \epsilon$$

Replace t + 1 by T and M/m by K

$$Tlog(1 - \frac{1}{K}) + log(f(x_0) - p^*) \leq log\epsilon$$

$$Tlog(1 - \frac{1}{K}) \leq log\left(\frac{\epsilon}{log(f(x_0) - p^*)}\right)$$

$$Tlog\left(\frac{K-1}{K}\right) \leq log\left(\frac{\epsilon}{log(f(x_0) - p^*)}\right)$$

$$T \leq \frac{1}{log\left(\frac{K-1}{K}\right)} log\left(\frac{\epsilon}{log(f(x_0) - p^*)}\right)$$

$$T \leq \frac{1}{log\left(\frac{K}{K-1}\right)} log\left(\frac{log(f(x_0) - p^*)}{\epsilon}\right)$$

Therefore we can approximate it to:

$$T = \frac{1}{log\left(\frac{K}{K-1}\right)} log\left(\frac{log(f(x_0) - p^*)}{\epsilon}\right)$$

1.1.1 Gradient Evaluations

According to the gradient descent algorithm, we evaluate the gradient every time we update the value the parameters i.e. once per iteration.
Therefore, the total number of gradient evaluations for T iterations = T

1.1.2 Function Evaluations

In this case, we have the η parameter fixed. If it was otherwise, we would evaluate function every time to update η. Therefore, in this case, the number of function evaluations are = 0.

1.2

1.2.1 Dependence on Value of function

To calculate the constant value η we minimize the following:

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{M}{2}\|p\|^2$$

i.e. for p $= -\eta \nabla f(x_t)$, we can solve for

$$f(x_{t+1}) = min_\eta f(x_t - \eta \nabla f(x_t))$$

$$f(x_{t+1}) = min_\eta f(x_t) + \nabla f(x_t)^T(-\eta \nabla f(x_t)) + \frac{M}{2}\|\eta \nabla f(x_t)\|^2$$

$$f(x_{t+1}) = min_\eta f(x_t) + \|\nabla f(x_t)\|^2(-\eta + \frac{M}{2}\eta^2)$$

Differentiating w.r.t η, we get

$$\|\nabla f(x_t)\|^2(1 - M\eta) = 0$$

$$\eta = \frac{1}{M}$$

Thus, we derived that the value for constant parameter for step size depends on the magnitude of the Hessian as M is the max eigen value which is the upper bound.

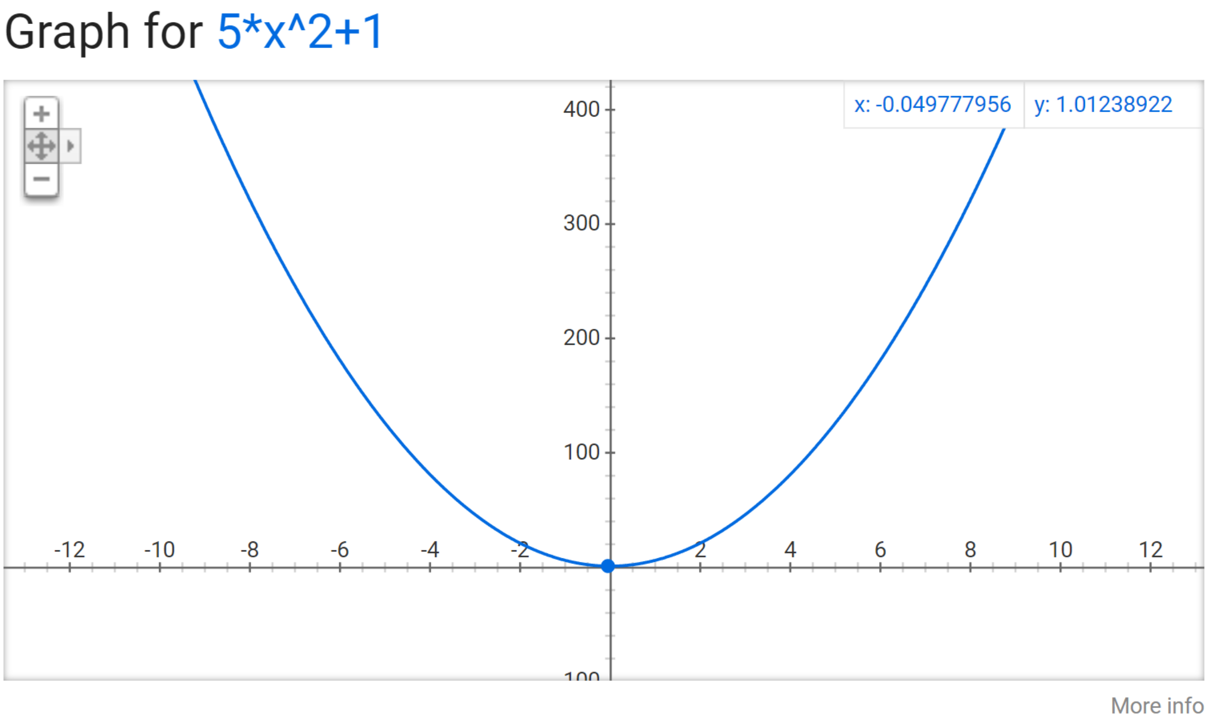
1.2.2 Divergence example

Consider the following minimization objective where the function is a 1D scalar function.

$$min_x 5x^2 + 1$$

The gradient is calculated as:

$$f'(x) = 10x$$



The graph of this equation is a parabola passing through the origin and is a convex function. Consider $\eta = 1$ as the constant learning parameter. For starting point as $x_0 = 3$ we can use the update as

$$x_{t+1} = x_t - \eta f'(x)$$

$$x_1 = 3 - 1 * 30$$

$$x_1 = -27$$

The next update is

$$x_2 = -27 + 1 * 270$$

$$x_2 = 243$$

The next update is

$$x_3 = 243 - 1 * 2430$$

$$x_3 = 2187$$

Thus we observe that the value of our next parameter is overshooting and instead of converging to 0, it keeps on increasing.

2 Newton's Method

2.1 Newton Step

In this example, we have $x = Ay + b$ and $f(x) = g(y)$ i.e. $f(Ay + b) = g(y)$.
For the independent variable y, we have the Newton step as follows:

$$\Delta y = -(\nabla^2 g(y))^{-1} \nabla g(y)$$

Similarly for x,

$$\Delta x = -(\nabla^2 f(x))^{-1} \nabla f(x)$$

we know,

$$g(y) = f(x)$$

Therefore,

$$g(y) = f(Ay + b)$$

$$\nabla g(y) = A^T \nabla f(Ay + b)$$

and

$$\nabla^2 g(y) = A^T \nabla^2 f(Ay + b) A$$

Therefore,

$$\Delta y = -(\nabla^2 g(y))^{-1} \nabla g(y)$$

$$\Delta y = -(A^T \nabla^2 f(Ay + b) A)^{-1} A^T \nabla f(Ay + b)$$

$$\Delta y = -A^{-1} \nabla^2 f(Ay + b)^{-1} {A^T}^{-1} A^T \nabla f(Ay + b)$$

$$\Delta y = -A^{-1} (\nabla^2 f(x))^{-1} \nabla f(x)$$

$$\Delta y = A^{-1} \Delta x$$

$$\Delta x = A \Delta y$$

2.2 Backtracking Condition

For some α such that $0 \leq \alpha \leq 0.5$, the exit condition for backtracking line search for $g(y)$ is

$$g(y + \eta\Delta y) > g(y) + \alpha \nabla g(y)^T \Delta y$$

$$g(y + \eta\Delta y) > g(y) - \alpha\eta \nabla g(y)^T (\nabla^2 g(y))^{-1} \nabla g(y) \tag{2}$$

Similarly for f(x) we can write:

$$f(x + \eta\Delta x) > f(x) - \alpha\eta \nabla f(x)^T (\nabla^2 f(x))^{-1} \nabla f(x)$$

$$f(A(y + \eta\Delta y) + b) > f(Ay + b) - \alpha\eta \nabla f(Ay + b)^T (\nabla^2 f(Ay + b))^{-1} \nabla f(Ay + b)$$

$$f(A(y + \eta\Delta y) + b) > f(Ay + b) - \alpha\eta (A^T \nabla f(Ay + b))^T (A^T \nabla^2 f(Ay + b) A)^{-1} A^T \nabla f(Ay + b)$$

From question 2.1,

we can substitute the values of:

$$f(Ay + b) = g(y)$$

$$A^T \nabla f(Ay + b) = \nabla g(y)$$

$$A^T \nabla^2 f(Ay + b) A = \nabla^2 g(y)$$

Therefore the equation becomes:

$$f(A(y + \eta\Delta y) + b) > g(y) - \alpha\eta \nabla g(y)^T (\nabla^2 g(y))^{-1} \nabla g(y)$$

Note that the rhs of the equation is similar to that of equation (2). Thus we can say that the termination condition for the backtracking algorithm for f(x) in the direction of Δx depends on the stopping condition for g(y) in the direction of Δy. y being the independent variable, we can prove by contradiction that if g(y) doesn't satisfy the stopping criteria for backtracking then f(x) will also fail to satisfy the condition.

2.3 Sequence of Iterates

2.3.1

Given $x^{(0)} = Ay^{(0)} + b$

Therefore, consider calculating the term $x^{(1)}$

$$x^{(1)} = x^{(0)} + \eta\Delta x$$

where Δx is calculated from the above formula. We can substitute the value of $x^{(0)}$.

$$x^{(1)} = Ay^{(0)} + b + \eta\Delta x$$

From question 2.1 we can substitute for Δx

$$x^{(1)} = Ay^{(0)} + b + \eta A\Delta y$$

$$x^{(1)} = A(y^{(0)} + \eta\Delta y) + b$$

But $y^{(1)} = y^{(0)} + \eta\Delta y$

$$x^{(1)} = Ay^{(1)} + b$$

Similarly, we can say by Induction, for $(k-1)^{th}$ term,

$$x^{(k-1)} = Ay^{(k-1)} + b$$

Proving for case k,

$$x^{(k)} = x^{(k-1)} + \eta\Delta x$$

$$x^{(k)} = Ay^{(k-1)} + b + \eta\Delta x$$

$$x^{(k)} = Ay^{(k-1)} + b + A\eta\Delta y$$

$$x^{(k)} = A(y^{(k-1)} + \eta\Delta y) + b$$

$$x^{(k)} = Ay^{(k)} + b$$

2.3.2

We know that $g(y) = f(Ay + b)$ i.e. $g(y) = f(x)$
for the first iterate, we have,

$$f(x^{(1)}) = f(x^{(0)} + \eta\Delta x)$$

the first order Taylor series expansion is:

$$f(x^{(1)}) = f(x^{(0)}) + \nabla f(x^{(0)})^T(\eta\Delta x)$$

$$f(x^{(1)}) = f(x^{(0)}) - \eta\nabla f(x^{(0)})^T(\nabla^2 f(x^{(0)}))^{-1}\nabla f(x^{(0)})$$

$$f(x^{(1)}) = f(Ay^{(0)} + b) - \eta\nabla f(Ay^{(0)} + b)^T(\nabla^2 f(Ay^{(0)} + b))^{-1}\nabla f(Ay^{(0)} + b)$$

$$f(x^{(1)}) = f(Ay^{(0)} + b) - \eta(A\nabla f(Ay^{(0)} + b))^T(A^T\nabla^2 f(Ay^{(0)} + b)A)^{-1}A^T\nabla f(Ay^{(0)} + b)$$

$$f(x^{(1)}) = g(y^{(0)}) - \eta(\nabla g(y^{(0)})^T(\nabla^2 g(y^{(0)}))^{-1}\nabla g(y^{(0)})$$

$$f(x^{(1)}) = g(y^{(1)})$$

Similarly, we can prove for the rest that $f(x^{(k)}) = g(y^{(k)})$

2.4 Newton decrement

We can write the Newton decrement for f(x) as follows:

$$\lambda(x) = (\nabla f(x)^T\nabla^2 f(x)^{-1}\nabla f(x))^{\frac{1}{2}}$$

From above problems we can substitute the values:

$$\lambda(x) = (\nabla f(Ay + b)^T\nabla^2 f(Ay + b)^{-1}\nabla f(Ay + b))^{\frac{1}{2}}$$

$$\lambda(x) = (\nabla f(Ay + b)^T AA^{-1}\nabla^2 f(Ay + b)^{-1}A^{T^{-1}}A^T\nabla f(Ay + b))^{\frac{1}{2}}$$

$$\lambda(x) = (\nabla f(Ay + b)^T\nabla^2 f(Ay + b)^{-1}\nabla f(Ay + b))^{\frac{1}{2}}$$

$$\lambda(x) = (\nabla g(y)^T\nabla^2 g(y)^{-1}\nabla g(y))^{\frac{1}{2}}$$

But the equation for Newton decrement for g(y) is:

$$\lambda(y) = (\nabla g(y)^T\nabla^2 g(y)^{-1}\nabla g(y))^{\frac{1}{2}}$$

$$\lambda(x) = \lambda(y)$$

The stopping criteria for f(x) is:

$$\lambda^2(x) \leq \epsilon$$

and the stopping criteria for g(y) is

$$\lambda^2(y) \leq \epsilon$$

As:

$$\lambda(x) = \lambda(y)$$

$$\lambda^2(x) = \lambda^2(y)$$

Therefore the stopping criteria for both of them are same.