CSE 592: Convex Optimization
HW 1

Bhushan Sonawane
SBU ID: 111511679

February 21, 2018

1 Gradient Decent without Linesearch

1.1 How many gradient evaluations are performed to reach an ϵ-suboptimal solution? How many function evaluations

We are given that,

$$\nabla^2 f(x) \geq mI$$

$$\nabla^2 f(x) - mI \geq 0$$

We know that function is strongly convex, hence

$$p^* - f(x^*) \geq f(x_t) + \nabla f(x_t)^T (x^* - x_t) + m/2||x_* - x_t||^2$$

Right hand side of the equation is convex quadratic function and setting its gradient with respect to y to zero, we get

$$y = x - (1/m)f(x)$$

Hence,

$$p^* \geq f(x) - 1/2m||\nabla f(x)||^2$$

$$f(x) - p^* \leq 1/2m||\nabla f(x)||^2$$

We also know that,

$$\nabla^2 f(x) \leq MI$$

hence,

$$p^* - f(x^*) \leq f(x_t) + \nabla f(x_t)^T (x^* - x_t) + M/2||x_* - x_t||^2$$

We can take gradient of second term in RHS similar to what we did for $\nabla^2 f(x) \geq mI$ case. Hence,

$$f(x) - p^* \geq 1/2M||\nabla f(x)||^2$$

As f(x) is strongly convex, we can write,

$$f(x + p) \leq f(x) + \nabla f(x)^T p + M/2||p||^2$$

Hence,

$$\leq min_\eta f(x_t) + \nabla f(x_t)^T (-\eta \nabla f(x_t) + M/2||\eta \nabla f(x_t||^2$$

We know that, $\eta = 1/M$ Hence,

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

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

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

Substracting p^* from both sides,

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

Chaging $f(x_t)$ to $f(x_0)$ i.e. starting from first iteration instead of previous iteration.

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

Taking log on both side for finding iteration,
1 - m/M ¡ 1, which shows that $f(x^k)$ converges to p^* as k $- >$ ∞ In particular, we must have $f(x^k) - p^* \leq \epsilon$ after at most,

$$\frac{log((f(x^0 - p^0)/\epsilon)}{-log(1 - m/M)}$$

we can further simplify to, using $K = M/m$

$$\frac{1}{-log(1 - 1/K)} * \frac{log((f(x^0 - p^0)}{\epsilon}$$

$$\frac{1}{-log(\frac{K-1}{K})} * \frac{log((f(x^0 - p^0)}{\epsilon}$$

$$T = \frac{1}{log(\frac{K}{K-1})} * \frac{log((f(x^0 - p^0)}{\epsilon}$$

Hence, at T^{th} iteration, we have ϵ-suboptimal solution
Number of gradient evaluations= t (every iteration requires one gradient evaluation)
Number of function evaluations= 0 (η is given and finding η requires function evaluation)

1.2 For any , find a twice-continuously differentiable strongly convex function f (x) with bounded Hessian and an initial point x 0 such that gradient descent with fixed stepsize t = starting at x 0 yields a sequence of iterates that does not converge to the optimum

Let's prove this by an counter example, Consider an quadratic function $x^2 + 8$ and $\eta = 2$ assume initial point x = 200

Iteration 1
x = 200
Δx = -(2x) = -400
$\eta = 2$
x = 200 + 2(-400) = -600
Iteration 2
x = -600
Δx = -(2x) = 1200
$\eta = 2$
x = -600 + 2(1200) = 1800
Iteration 3
x = 1800
Δx = -(2x) = -3600
$\eta = 2$
x = 1800 + 2(-3600) = 5200
Iteration 4
x = 5200
Δx = -(2x) = -10400

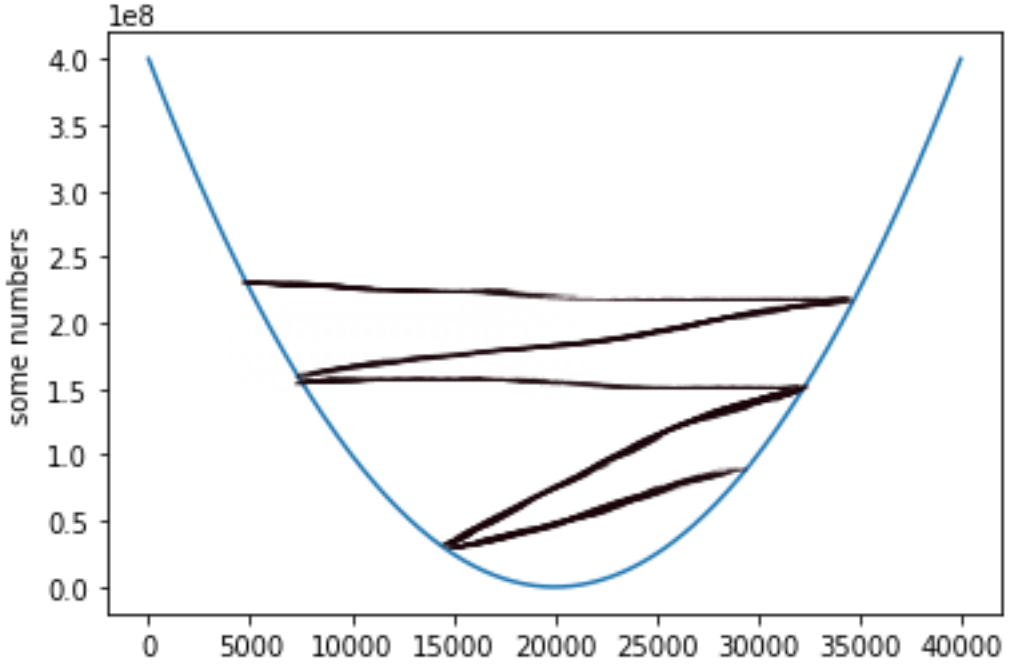
$\eta = 2$
x = 5200 + 2(-10400) = -15600
....
....
....

We can see that as iteration increase, gradient is over-shooting i.e. going far from the optimum. Thus, We can derive an quadratic function with constant η that does not reach the optimum.

Proof that step size should depend on the function or at least on the magnitude of the Hessian:

We have seen that fix step size might lead to overshoot and solution will start going away from the optimum. Clearly, this can be avoided if we update step size every time according to current state of the function.

Hence, updating step size depending on the function or on the magnitude of the Hessian will lead to the optimum solution.



2 Newton's Method

2.1 For x = Ay + b, let Δx and Δy be the Newton steps for f (x) and g(y) respectively. Prove that Δx = AΔy

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

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

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

we know that,

$$x = Ay + b$$

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

but, g(y) = f(Ax+b)

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

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

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

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

But, we know that,

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

Hence,

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

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

Hence prooved.

2.2 Prove that for any $\eta > 0$, the exit condition for backtracking line search on f(x) in direction Δx will hold if and only if the exit condition holds for g(y) in direction Δy.

Backtracking line search exit condition is for g(y) is:

$$f(y + \eta \Delta f(y)^T \Delta y) < f(Ay+b) - \alpha \eta \nabla f(Ay+b)^T \Delta y$$

substituting Δy from 2.1

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

Hence,

$$f(y + \eta \Delta f(y)^T \Delta y) < f(Ay+b) - \alpha \eta \nabla f(Ay+b)^T \nabla^2 f(Ay+b) \nabla f(Ay+b))$$

$$f(x + \eta \Delta f(x)^T \Delta x) < f(x) + \alpha \eta \nabla f(x)^T \Delta x$$

Consider RHS for now,

$$f(x) + \alpha \eta \nabla f(x)^T \Delta x$$

putting x = Ay+b and

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

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

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

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

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

$$f(Ay+b) - \alpha\eta\nabla f(Ay+b)^T \nabla^2 f(Ay+b)\nabla f(Ay+b))$$

Hence,

$$f(x+\eta\Delta f(x)^T \Delta x) < f(Ay+b) - \alpha\eta\nabla f(Ay+b)^T \nabla^2 f(Ay+b)\nabla f(Ay+b))$$

We also know that,

$$f(y+\eta\Delta f(y)^T \Delta y) < f(Ay+b) - \alpha\eta\nabla f(Ay+b)^T \nabla^2 f(Ay+b)\nabla f(Ay+b))$$

Exit condition for both f(x) and g(y) is same and hence, exit condition for f(x) will hold in direction Δx will hold only if the exit condition hold for g(y) in direction Δy and vice versa.

2.3 Consider running Newton's method on g(·) starting at some y(0) and on f(·) starting at x(0) = Ay(0) +b. Use the above to prove that the sequences of iterates obeys x(k) = Ay(k) +b and f(x(k)) = g(y(k))

We update x as follows,

$$x(1) = x(0) + \eta\Delta(x)$$

Let's consider the base case, i.e. x(0) and y(0)

Put, x = Ay + b and Δx = AΔy in above equation,

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

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

but, y(1) = y(0) + η Δ (y)

$$x(1) = Ay(1) + b$$

Hence, we can say that

$$f(x(1)) = f(Ay(1) + b)$$

$$g(y(1)) = f(Ay(1) + b)$$

by Induction, assuming above is true for $(i-1)^{th}$ iteration, Let's prove for i-th iteration

$$x(i) = Ay(i-1) + b + \eta A\Delta(y)$$

$$x(i) = A(y(i-1) + \eta\Delta(y)) + b$$

but, y(i) = y(i-1) + η Δ (y)

$$x(i) = Ay(i) + b$$

Hence, we can say that

$$f(x(i)) = f(Ay(i) + b)$$

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

Hence, Proved by induction.

2.4 Prove that Newton's decrement for f(·) at x is equal to Newton's decrement for g(·) at y, and so the stopping conditions are also identical

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

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

substituting value of x = Ay+b

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

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

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

but, we know that, f(Ay+b) = g(y)

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

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

Stopping condition is $()^(2) < \epsilon$ Given that,

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

Hence,
Stopping condition of f(.) is same as g(.)
Hence proved.

3 Implementation

Output of executation -

Optimum of the weird function found by bisection 1.4340255537854398

Optimum of the weird function found by GD with exact line search [[1.43402555]]

Optimum of the weird function found by Newton with exact line search [[1.43402555]]

Optimum of the weird function found by GD with backtracking line search [[1.43375809]]

Optimum of the weird function found by Newton with backtracking line search [[1.43916293]]

Optimum of the quadratic function found by GD with exact line search [[0.00765583 -0.00011342]]

Optimum of the quadratic function found by Newton with exact line search [[0. 0.]]

Optimum of the quadratic function found by GD with backtracking line search [[0.00708124 -0.00023187]]

Optimum of the quadratic function found by Newton with backtracking line search [[0. 0.]]

Optimum of the Boyd's function found by GD with exact line search [[-3.47097476e-01 2.47005594e-05]]

Optimum of the Boyd's function found by Newton with exact line search [[-0.34126199 0.000805]]

Optimum of the Boyd's function found by GD with backtracking line search [[-3.47261101e-01 1.96441251e-04]]

Optimum of the Boyd's function found by Newton with backtracking line search [[-0.34513821 0.00071872]]