CSE592 Convex Optimization: Home Work 1

Ashik Vetrivelu (111652994)

February 21, 2018

1. Gradient Descent

1. Prove that for any convex function with bounded Hessian, with a fixed step size $\eta_t = \frac{1}{M}$ after T iterations is,

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

Here since the function is bounded and since it's strongly convex and strongly smooth.
Since, $mI \leq \nabla^2 f(x) \leq MI$,

$$p^* = f(x^*) \geq f(x_t) + \nabla f(x_t)^T(x^* - x_t) + \frac{m}{2}||x_t - x^*||^2 \geq min_y f(x_t)^T(y - x_t) + \frac{m}{2}||y - x_t||^2$$

We know that,

$$\nabla f(x_t) + m(y - x_t) = 0$$
$$\implies y^* = x_t - \frac{1}{m}\nabla f(x_t$$

substituting this in the first equation,

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

Since f(x) is also upper-bounded, it's strongly smooth. Because of this,

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

The above result is obtained by substituting η with $\frac{1}{M}$. Combining this with the previous equation yields us,

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

$$\implies f(x_{t+1}) - p* \leq (f(x_t) - p*)(\frac{\kappa - 1}{\kappa}) \leq (\frac{\kappa - 1}{\kappa})^{t+1}(f(x_0) - p^*) \leq \epsilon$$
$$\implies (\frac{\kappa - 1}{\kappa})^{t+1} \leq (\frac{f(x_0) - p^*}{\epsilon})^{-1}$$
$$\implies (\frac{\kappa}{\kappa - 1})^{t+1} \leq (\frac{f(x_0) - p^*}{\epsilon})$$
$$\implies t + 1 \leq \frac{\log(\frac{f(x_0) - p^*}{\epsilon})}{\log(\frac{\kappa}{\kappa - 1})}$$
$$\implies T = \frac{1}{\log(\frac{\kappa}{\kappa - 1})} \log(\frac{f(x_0) - p^*}{\epsilon})$$

Hence, we obtained our number of iterations. For gradient descent, the number of gradient evaluations will be **'T'** in this case as we are only computing our gradient at one point per iteration. We aren't interested in the value of function in our gradient descent algorithm. So, the number of function evaluations is **0**.

2. Here, let's take a simple one dimensional quadratic function let's say ax^2. Since the function is bounded by assumption, the function f is strongly smooth. We know that,

$$\begin{aligned} \nabla f(x) &= 2ax \\ x^{t+1} &= x^t - 2a\eta x^t \\ \implies x^{t+1} - x^t &= 2a\eta x^t \\ \implies \eta &= \frac{x^{t+1} - x^t}{2ax^t} \end{aligned}$$

Here, taking t = 0 and x^0 = m,

$$\eta = \frac{x^1 - m}{2am}$$

We know that the optimum is at x^* = 0.

$$\begin{aligned} x^1 &= m - \eta 2am \\ x^1 &= m(1 - 2a\eta) \qquad \textit{(writing } 1 - 2a\eta \textit{ as } \gamma\textit{),} \\ x^1 &= m\gamma \\ x^2 &= m\gamma - 2a\eta m\gamma \\ x^2 &= m\gamma(1 - 2a\eta) \\ x^2 &= m\gamma^2 \\ \implies x^t &= m\gamma^t \end{aligned}$$

Since we know that at optimum is at t when x^t = 0, and $m\gamma^t$ can only be equal to 0 if γ = 0.
$i.e$,

$$\begin{aligned} 1 - 2a\eta &= 0 \\ \implies \eta &= \frac{1}{2a} \end{aligned}$$

Here, $2a$ is the hessian and η is dependent on the hessian to arrive at optimum. For other values of η, we won't even arrive at the optimum.

2. Newtons Method

Here, $y \in \mathbb{R}^m \to Ay + b$ and $g(y) = f(Ay + b)$. $Ay + b$ is the affine transformation of x.

1. From the question, $x = Ay + b$.
For Newton's method we know that,

$$\begin{aligned}\Delta x &= -[\nabla^2 f(x)]^{-1} \nabla f(x) \\ \Delta y &= -[\nabla^2 g(y)]^{-1} \nabla g(y)\end{aligned}$$

Since $g(y)$ can be transformed as $f(Ay + b)$, we can write the gradients and hessians of $g(y)$ as follows,

$$\begin{aligned}\nabla g(y) &= A^T \nabla f(Ay + b) \\ &= A^T \nabla f(x) \\ \nabla^2 g(y) &= A^T \nabla^2 f(Ay + b) A \\ &= A^T \nabla^2 f(x) A\end{aligned}$$

Substituting these equations in Δy, we get,

$$\begin{aligned}\Delta y &= -[A^T \nabla^2 f(x).A]^{-1}.A^T.\nabla f(x) \\ &= \frac{1}{A}\Big(- [\nabla^2 f(x)]^{-1} \nabla f(x)\Big) \\ &= \frac{1}{A} \Delta x \\ \implies \Delta x &= A \Delta y\end{aligned}$$

2. Here we can rewrite the equation,

$$\begin{aligned}g(y + \eta\Delta y) &> g(y) + \alpha\eta[\nabla g(y)]^T \Delta y \\ f(A(y + \eta\Delta y) + b) &> f(Ay + b) + \alpha\eta A.[\nabla f(Ay + b)]^T.A^{-1}.\Delta x \\ f(Ay + A\eta\Delta y + b) &> f(x) + \alpha\eta[\nabla f(x)]^T.\Delta x \qquad \textit{(Simplifying the equation to transform } Ay + b \textit{ as } x\textit{)} \\ f(x + A\eta\Delta y) &> f(x) + \alpha\eta[\nabla f(x)]^T.\Delta x \\ f(x + \eta\Delta x) &> f(x) + \alpha\eta[\nabla f(x)]^T.\Delta x \qquad \textit{(As we know } x = A\Delta y\textit{)}\end{aligned}$$

Thus we got the B.T.L.S stop condition of x from g(y). This proves that the condition solely depends on y.

3. Here we know that the equation holds when $t = 0$. That is, $x^{(0)} = Ay^{(0)} + b$ and $f(x^{(0)}) = g(y^{(0)})$.
We can prove this by mathematical induction assuming that these equations are true for $t = k$.
That is,
$x^{(k)} = Ay^{(k)} + b$
We know that, in Newton's Method,

$$\begin{aligned}x^{(k+1)} &= x^{(k)} + \eta\Delta x_k \\ y^{(k+1)} &= y^{(k)} + \eta\Delta y_k\end{aligned}$$

Since we already proved that $\Delta x = A\Delta y$, applying that here,

$$\begin{aligned} x^{(k+1)} &= (Ay^{(k)} + b) + \eta(A\Delta y_k) \\ &= A(y^{(k)} + \eta\Delta y_k) + b \\ &= A(y^{(k+1)}) + b \qquad \textit{(by substituting } y^{(k)} + \eta\Delta y_k \textit{ with } y^{(k+1)}\textit{)} \end{aligned}$$

Hence we know this holds for $x^{(k+1)}$ also and by induction, this holds for all the values of t.
Also we know that,

$$\begin{aligned} & g(y^{(k)}) = f(Ay^{(k)} + b) \qquad \textit{(by definition)} \\ \implies & g(y^{(k)}) = f(x^{(k)}) \qquad \textit{(Since we already proved } x^{(k)} = Ay^{(k)} + b\textit{)} \end{aligned}$$

Hence, we have proved that the conditions satisfy for any iterate in Newtons.

4. We can simply write Newton's Decrement for x and y as,

$$\begin{aligned} \lambda^2(x) = (\nabla f(x))^T(\nabla^2 f(x))^{-1}\nabla f(x) \\ \lambda^2(y) = (\nabla g(y))^T(\nabla^2 g(y))^{-1}\nabla g(y) \end{aligned}$$

From the first solution we already know that the Gradient and Hessian is transformed as

$$\begin{aligned} \nabla g(y) = A^T.\nabla f(x) \\ \nabla^2 g(y) = A^T.\nabla^2 f(x).A \end{aligned}$$

We can substitute the Gradient and Hessian of g with these in the the Newton's decrement for y.

$$\begin{aligned} \lambda^2(y) &= A.(\nabla f(x))^T[A^T.\nabla^2 f(x).A]^{-1}.\nabla f(x).A^T \\ &= (\nabla f(x))^T(\nabla^2 f(x))^{-1}\nabla f(x) \qquad \textit{(Canceling out all the A)} \\ &= \lambda^2(x) \end{aligned}$$

Hence the Newton's decrement is equal for both $f(\cdot)$ at x and $g(\cdot)$ at y. The stopping condition is nothing but, $\lambda^2(\cdot) \leq \epsilon$ so this is identical for both x and y. $f(x - \eta A^T\nabla f(x))$