COM SCI 260B HW 1 Solution

Ashish Kumar Singh (UID:105479019)

April 12, 2022

Problem 1. Bound on gradient

Solution 1a. One example function is $f(x,y) = x^2 - y^2$, $\nabla f(x,y) = \begin{bmatrix} 2x \\ -2y \end{bmatrix}$
At $(x,y) = (0,0)$ we see that the gradient is zero, but it is not a local maximum or minimum as $f(0,0) = 0$ but $f(\delta, 0) = \delta^2$ and $f(0,\delta) = -\delta^2$
In the plot below we can see, $(x,y) = (0,0)$ is not a maxima or a minima

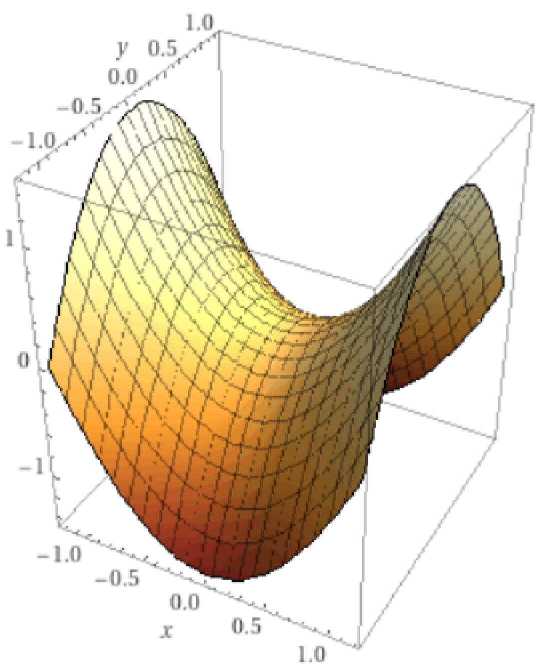

Figure 1: Plot of $f(x,y) = x^2 - y^2$

Solution 1b. From the monotonicity of GD, we have

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2 \tag{1}$$

Lets say GD converges to x_*, then

$$f(x_*) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2 \tag{2}$$

Now from the definition of $\beta - smoothness$, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|_2^2$$

Using $y = x_k$, $x = x_*$ and $\nabla f(x_*) = 0$, we get

$$f(x_k) \leq f(x_*) + \frac{\beta}{2}\|x_k - x_*\|_2^2 \tag{3}$$

Substituting eqn2 in eqn3 we get,

$$f(x_k) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2 + \frac{\beta}{2}\|x_k - x_*\|_2^2$$

$$\|\nabla f(x_k)\|_2^2 \leq \beta^2\|x_k - x_*\|_2^2$$

The right handside goes to 0 as we converge, which shows that with GD we can find $w = x_k$ which has arbitrarily small gradient norm.

Now , lets assume kth iteration of GD is the first iteration when the gradient norm is less then ϵ, i.e. $\|\nabla f(x_k)\|_2 \leq \epsilon$ and $\|\nabla f(x_i)\|_2 > \|\nabla f(x_k)\|_2$ for every $i < k$,
Writing eqn1 for different values of k we have,

$$f(x_1) \leq f(x_0) - \frac{1}{2\beta}\|\nabla f(x_0)\|_2^2$$

$$f(x_2) \leq f(x_1) - \frac{1}{2\beta}\|\nabla f(x_1)\|_2^2$$

$$\vdots$$

$$f(x_k) \leq f(x_{k-1}) - \frac{1}{2\beta}\|\nabla f(x_{k-1})\|_2^2$$

Adding the above inequalities, and using $\|\nabla f(x_i)\|_2 > \|\nabla f(x_k)\|_2$ for every $i < k$ we get

$$f(x_k) \leq f(x_0) - \frac{k}{2\beta}\|\nabla f(x_k)\|_2^2 \tag{4}$$

Substituting eqn4 in eqn2 we get,

$$f(x_*) \leq f(x_0) - \frac{k}{2\beta}\|\nabla f(x_k)\|_2^2 - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2$$

$$\|\nabla f(x_k)\|_2^2 \leq \frac{2\beta}{k+1}(f(x_0) - f(x_*))$$

For $\|\nabla f(x_k)\|_2 \leq \epsilon$, we will have

$$\frac{2\beta}{k+1}(f(x_0) - f(x_*)) \leq \epsilon^2$$

$$k \geq \frac{2\beta}{\epsilon^2}(f(x_0) - f(x_*)) - 1 \tag{5}$$

Thus, with GD we can find points with arbitrarily small gradient norm. The number of steps required to achieve it is bounded by eqn5

Problem 2. Convergence rate of GD

Solution 2. Given that for $\alpha - convex$ function f,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2 \tag{6}$$

Also from monotonicity of GD for $\beta - smooth$ function f,

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2 \tag{7}$$

Using $y = x_*$ and $x = x_k$ in eqn6 and putting in eqn7,

$$f(x_{k+1}) \leq f(x_*) - \langle \nabla f(x_k), x_* - x_k \rangle - \frac{\alpha}{2}\|x_* - x_k\|_2^2 - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2$$

Using $\nabla f(x_k) = \beta(x_k - x_{k+1})$

$$f(x_{k+1}) - f(x_*) \leq \frac{1}{2\beta}\Big[2\beta\langle \beta(x_k - x_{k+1}), x_k - x_* \rangle - \alpha\beta\|x_* - x_k\|_2^2 - \beta^2\|x_k - x_{k+1}\|_2^2\Big]$$

$$f(x_{k+1}) - f(x_*) \leq \frac{\beta}{2}\Big[\|x_k - x_{k+1}\|_2^2 + \|x_* - x_k\|_2^2 - \|x_* - x_{k+1}\|_2^2 - \frac{\alpha}{\beta}\|x_* - x_k\|_2^2 - \|x_k - x_{k+1}\|_2^2\Big]$$

$$f(x_{k+1}) - f(x_*) \leq \frac{\beta}{2}\Big[(1 - \frac{\alpha}{\beta})\|x_k - x_*\|_2^2 - \|x_{k+1} - x_*\|_2^2\Big] \tag{8}$$

Note that the left hand side will always be positive because of the monotonicity of GD, so we can write,

$$0 \leq \frac{\beta}{2}\Big[(1 - \frac{\alpha}{\beta})\|x_k - x_*\|_2^2 - \|x_{k+1} - x_*\|_2^2\Big]$$

$$\|x_{k+1} - x_*\|_2^2 \leq (1 - \frac{\alpha}{\beta})\|x_k - x_*\|_2^2$$

For simplicity, lets assume $\lambda = (1 - \frac{\alpha}{\beta})$, note that by definition of convexity and smoothness, it is obvious that $\beta > \alpha$, hence $\lambda > 0$. Writing above eqn for different values of k, we get

$$\|x_1 - x_*\|_2^2 \leq \lambda\|x_0 - x_*\|_2^2$$
$$\|x_2 - x_*\|_2^2 \leq \lambda\|x_1 - x_*\|_2^2$$
$$\vdots$$
$$\|x_k - x_*\|_2^2 \leq \lambda\|x_{k-1} - x_*\|_2^2$$

On multiplying the above inequalities, we get

$$\|x_k - x_*\|_2^2 \leq \lambda^k\|x_0 - x_*\|_2^2$$

$$\|x_k - x_*\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^k\|x_0 - x_*\|_2^2$$

Problem 3. Differentiable convex function

Solution 3. Suppose for function $f : R^d \rightarrow R$ we have a local minimum at x_l, hence $\nabla f(x_l) = 0$
Now, lets assume that it is possible to find a global minimum at x_g, such that $f(x_g) < f(x_l)$.
Using the definition of convexity, for every u, v we have

$$f(v) \geq f(u) + \langle \nabla f(u), v - u \rangle$$

Using $v = x_g$ and $u = x_l$ and $\nabla f(x_l) = 0$,

$$f(x_g) \geq f(x_l) + \langle \nabla f(x_l), x_g - x_l \rangle$$

$$f(x_g) \geq f(x_l)$$

which contradicts our assumption that $f(x_g) < f(x_l)$, hence it is not possible to find another point whose function value is lower that $f(x_l)$, thus for convex differentiable function every local minimum is the global minimum.

Problem 4. GD vs NAGD

Solution 4a. Given,

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \sigma(\langle w, x_i\rangle))$$

$$L(w) = \frac{-1}{n}\sum_{i=1}^{n}(y_i log(\sigma(\langle w, x_i\rangle)) + (1-y_i)log(1-\sigma(\langle w, x_i\rangle)))$$

Differentiating wrt w_j and using $\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$

$$\frac{dL(w)}{dw_j} = \frac{-1}{n}\sum_{i=1}^{n} x_{ij}(y_i(1-\sigma(\langle w, x_i\rangle)) - (1-y_i)\sigma(\langle w, x_i\rangle))$$

$$\frac{dL(w)}{dw_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}(\sigma(\langle w, x_i\rangle) - y_i)$$

In matrix form,

$$\nabla L(w) = \frac{1}{n}X^T(\sigma(XW) - Y)$$

Solution 4b. The plots are below and the code is attached in following pages.

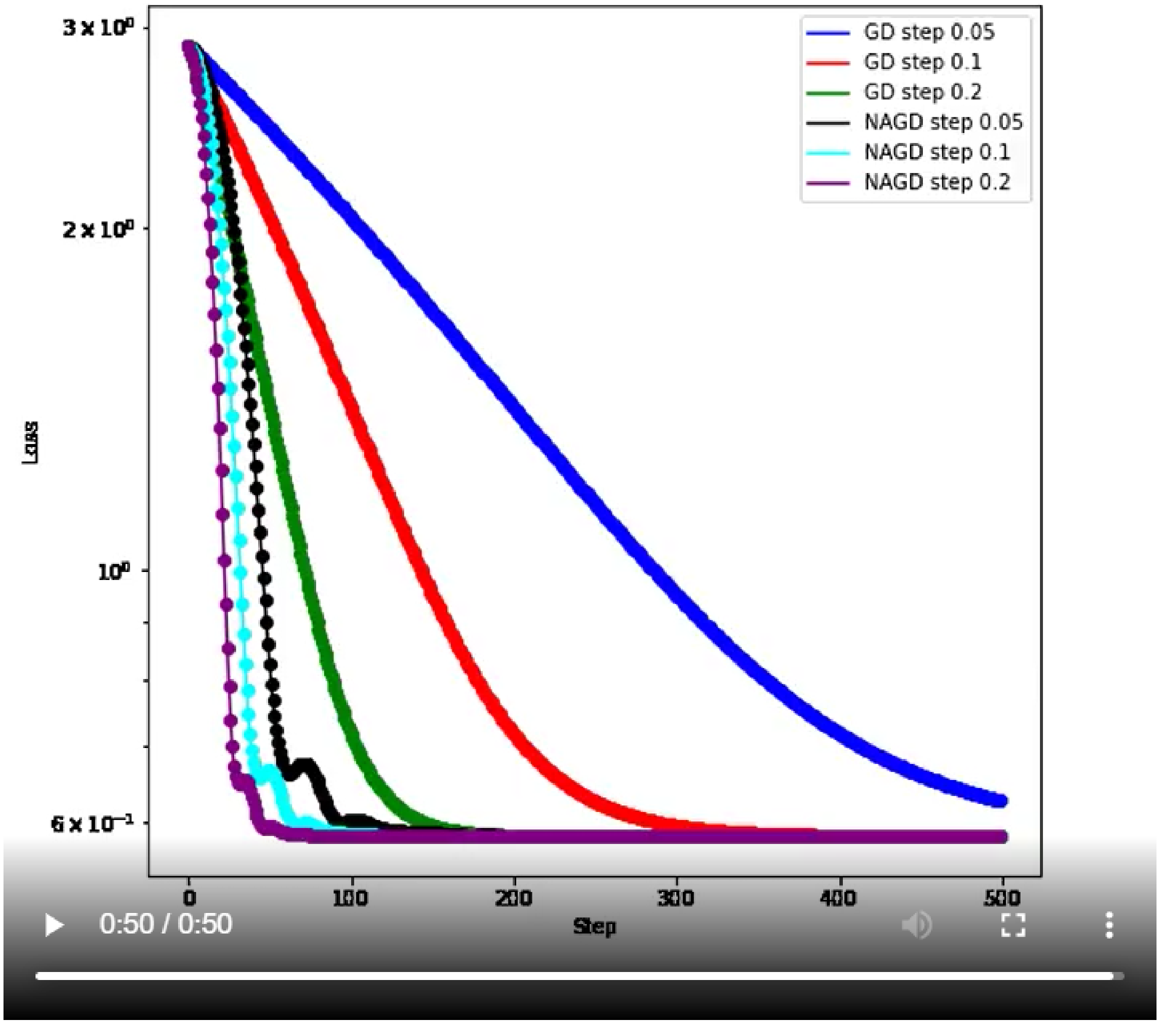

Figure 2: Loss function vs steps

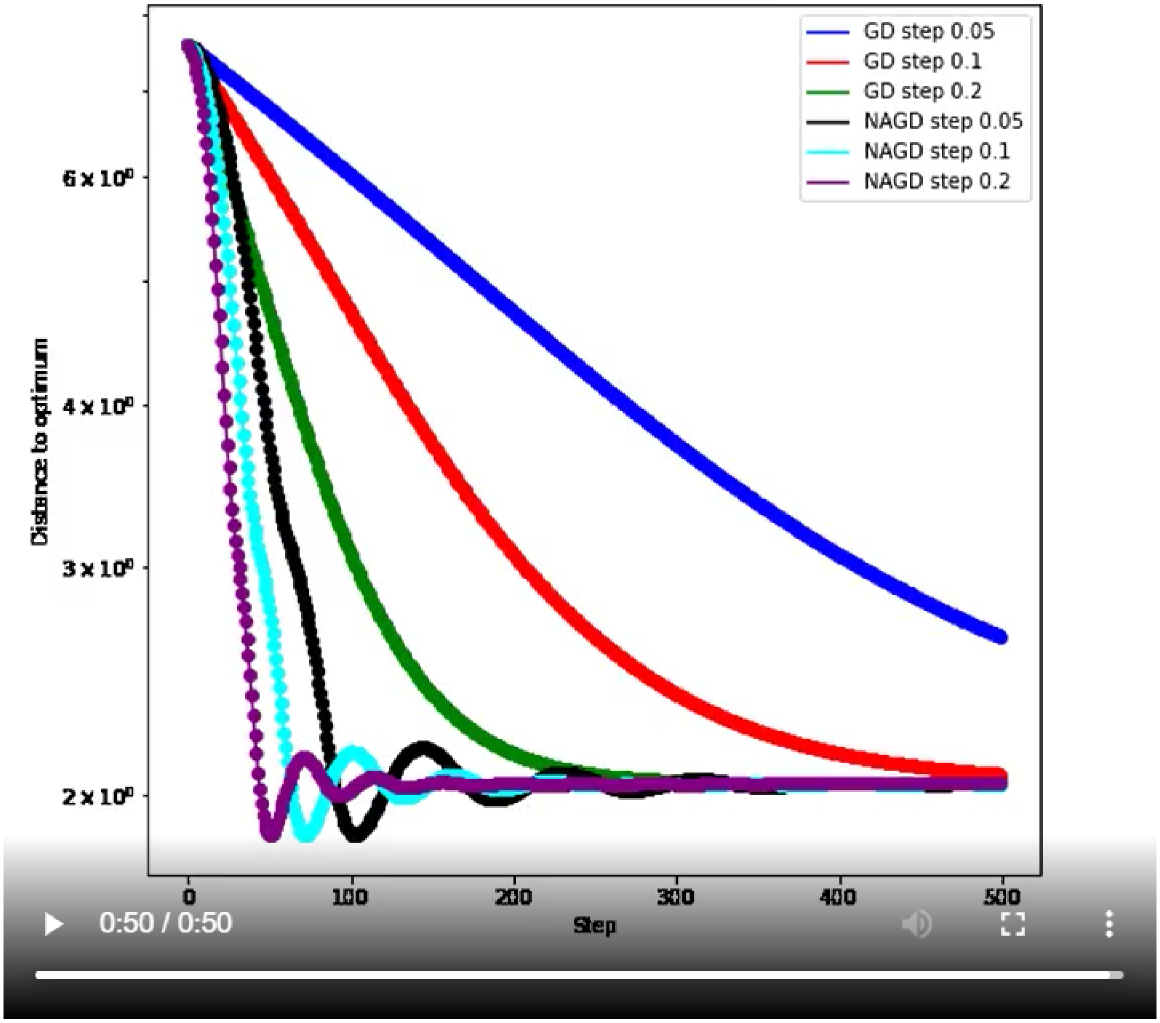

Figure 3: Distance to optimum w vs steps