Legi 16-352-137

Separating Hyperplane of Data Points

1. We need to prove (definition of subgradient)

$$g(y) \geq g(x) + \nabla g_{k(x)}^{\top}(y-x), \forall x, y \in \mathcal{R}^d.$$

But $g(x) = g_{k(x)}(y)$, $\; g(y) = \max_{i \in [m]} g_i(y) \geq g_{k(x)}(y)$ and from convexity and differentiability of $g_{k(x)}$ we have

$$g_{k(x)}(y) \geq g_{k(x)}(x) + \nabla g_{k(x)}^{\top}(y-x), \forall x, y \in \mathcal{R}^d,$$

and it's true only for one such value of $\nabla g_{k(x)}^{\top}(y-x)$ (differentiability), so

$$g(y) \geq g_{k(x)}(y) \geq g_{k(x)}(x) + \nabla g_{k(x)}^{\top}(y-x) = g(x) + \nabla g_{k(x)}^{\top}(y-x), \forall x, y \in \mathcal{R}^d d$$

and therefore $\nabla g_{k(x)}^{\top}(y-x)$ is subragient in $x \quad \square$.

Then, using this (because $g_i(x)$ now are $-y_i \mathbf{w}^{\top}\mathbf{x_i}$ — differentiable and convex (because linear)), we can claim that in any point x subgradient of $f(\mathbf{w})$ is $\nabla_{\mathbf{w}}(-y_i \mathbf{w}^{\top}\mathbf{x_i}) = -y_i \mathbf{x_i}$ where i is such that $-y_i \mathbf{w}^{\top}\mathbf{x_i}$ is maximal among all $i \in [m]$ (which indicates the worst classified point in the set (a we want it to be as less as possible, and in the end less than 0 for all i)).

2. From Exercise 24 we get that because $dom(f)$ is convex (the whole space is convex obviously) and there exist subgradient in every point (calculated in previous subtask), then f is convex. Also from the previous subtask we have that $\partial f(\mathbf{w}) = -y_i \mathbf{x_i}$ for some i. And $||-y_i \mathbf{x_i}|| = ||y_i|| \cdot ||\mathbf{x_i}|| = 1 \cdot ||\mathbf{x_i}|| \leq \max_{i \in [m]} ||\mathbf{x_i}|| = B$. From Lemma 4.4 (Exercise 25) we have that the last result is equivalent to B–Lipschitzness.

3. $y_i \mathbf{w}^{\top}\mathbf{x_i} \geq 1 \Rightarrow y_i \mathbf{w}^{\top}\mathbf{x_i} \leq -1$. Suppose $a := \min y_i \mathbf{w}^{\top}\mathbf{x_i} < -1$, then if we change $\mathbf{w} \to \frac{\mathbf{w}}{|a|}$, we will still have $y_i \mathbf{w}^{\top}\mathbf{x_i} \leq -1$ (all values shrinked no more than in $|a|$ times, and the smallest absolute value became 1), therefore still $-y_i \mathbf{w}^{\top}\mathbf{x_i} \geq 1$. But now the norm of $\mathbf{w}$ shrinked in $|a| > 1$ times, what is in contradiction with task and therefore proves the claim.

4. As the set is linearly separable there exist at least one $\mathbf{w}^* \neq \mathbf{0}$ such that $f(\mathbf{w}_*) < 0$.
We start from $\mathbf{w_0} = \mathbf{0}$ and do this at k–th step (k starts from 0):

$$\mathbf{w_{k+1}} = \mathbf{w_k} + \mathbf{g}(\mathbf{w_k}),\; \mathbf{g}(\mathbf{w_k}) \text{ — subgradient of } f(\mathbf{w}) \text{ as in 1. subtask}$$

until $\mathbf{w_{k+1}}^{\top}\mathbf{g}(\mathbf{w_{k+1}}) < 0$ (which means all points are classified correctly).
We know from subtask 1 that

$$\mathbf{g}(\mathbf{w_k}) = -y_i \mathbf{x_i},\; i = \arg\max_{j \in [m]} -y_j \mathbf{w_k} \mathbf{x_j}.$$

Let's prove that this algorithm converges to the solution.
After k steps we will have (obviously from recursion)

$$\mathbf{w_{k+1}} = \sum_{j=0}^{k} \mathbf{g}(\mathbf{w_j}) = \sum_{j=0}^{k} -y_i \mathbf{x_i}, \text{ all } i \in [m] \text{ and could be repeated in principle.} \tag{1}$$

Let's denote $A = \min_{j\in[m]} \frac{|\mathbf{w}_*^\top \mathbf{x_j}|}{||\mathbf{w}_*||} > 0$ (as it separates the points). Note that it doesn't on scale of $\mathbf{w}_*$.
Then multiplying (1) by $\mathbf{w}_*^\top$ from left in both sides we have

$$\mathbf{w}_*^\top \mathbf{w_{k+1}} = \sum_{j=0}^{k} \mathbf{w}_*^\top \mathbf{g}(\mathbf{w_j}) \geq (k+1)A||\mathbf{w}_*||.$$

Therefore from Cauchy-Schwarz inequality we have lower–bound

$$||\mathbf{w}_*||^2||\mathbf{w_{k+1}}||^2 \geq (k+1)^2A^2||\mathbf{w}_*||^2 \Rightarrow ||\mathbf{w_{k+1}}||^2 \geq (k+1)^2A^2 \tag{2}$$

We will further establish also upper–bound

$$\mathbf{w_{k+1}} = \mathbf{w_k} + \mathbf{g_k},$$

taking square Euclidean norm we have

$$||\mathbf{w_{k+1}}||^2 = ||\mathbf{w_k}||^2 + 2\mathbf{w_k}^\top\mathbf{g_k} + ||\mathbf{g_k}||^2$$

Assuming that the algorithm isn't finished yet (it's the main assumption we have and which we will have to end at some point as we will see in (4)) we have $\mathbf{w_k}^\top\mathbf{g}(\mathbf{w_k}) \geq 0$, therefore

$$||\mathbf{w_{k+1}}||^2 \leq ||\mathbf{w_k}||^2 + ||\mathbf{g_k}||^2$$

and therefore

$$||\mathbf{w_{j+1}}||^2 - ||\mathbf{w_j}||^2 \leq ||\mathbf{g_j}||^2, \ \forall j \in 0, \dots, k$$

Then summing telescopically for k from 0 to final k we have (using $\mathbf{w_0} = \mathbf{0}$)

$$||\mathbf{w_{k+1}}||^2 \leq \sum_{i=0}^{k} ||\mathbf{g_i}||^2$$

As the set of points is finite, then exist $B = \max_{i\in[m]} ||\mathbf{x_i}|| \geq \max_{i\in[m]} ||\mathbf{g}(\mathbf{w})||\forall\mathbf{w}$.
Then we have upper–bound (we could have also obtained it straightly from the subtask 2 using B–Lipschitzness)

$$||\mathbf{w_{k+1}}||^2 \leq (k+1)B \tag{3}$$

Combining (2) and (3) we have

$$(k+1)^2A^2 \leq ||\mathbf{w_{k+1}}||^2 \leq (k+1)B \tag{4}$$

which means that after

$$K = \lceil \frac{B}{A^2} - 1 \rceil \tag{5}$$

our assumption that algorithm hasn't finished should be violated (lower bound higher than upper) and it will stop obtaining needed $\mathbf{w}$.

5. As we have seen in subtask 6 after no more than K iterations from (5) we will finish. At each iteration k we need

   to obtain $\arg\max_{j\in[m]} -y_j \mathbf{w_k x_j}$ (which is $O(nm)$, as inner product in $\mathcal{R}^n$ costs $O(n)$ and we need to repeat it for at most $O(m)$ points),

   check that $\max_{j\in[m]} -y_j \mathbf{w_k x_j} \geq 0$ ($O(1)$),

   therefore overall we need no more that $O(\frac{nmB}{A^2})$ elementary operations. Constants A and B are determined as

$$A = \min_{j\in[m]} \frac{|\mathbf{w}_*^\top \mathbf{x_j}|}{||\mathbf{w}_*||},\ B = \max_{i\in[m]} ||\mathbf{x_i}||.$$

Accelerated Gradient Descent for Strongly Convex Functions

6. From the Theorem 2.8 from lectures (its conditions are satisfied from the task description) we have (because $z_0 \coloneqq x_0$)

$$\frac{2L||z_0 - x^*||^2}{T(T+1)} = \frac{2LR^2}{T(T+1)} \geq f(y_T) - f(x^*). \tag{6}$$

   From strong convexity

$$f(y) \geq f(x) + \nabla f(x)(y-x) + \frac{\mu}{2}||y-x||^2.$$

   for $y = y_T$ and $x = x^*$ we have (as $\nabla f(x^*) = 0$ in global minimum of differentiable convex function determined in the whole $\mathcal{R}^n$):

$$f(y_T) - f(x^*) \geq \frac{\mu}{2}||y_T - x^*||^2 \tag{7}$$

   which together with (6) gives

$$\frac{2LR^2}{T(T+1)} \geq f(y_T) - f(x^*) \geq \frac{\mu}{2}||y_T - x^*||^2. \tag{8}$$

   After regrouping it leads to

$$\frac{||y_T - x^*||^2}{R^2} \leq \frac{4L}{\mu T(T+1)}. \tag{9}$$

   It means that after T such that $\frac{4L}{\mu T(T+1)} \leq 0.5$ we can choose y_T as x and have needed condition of twice decreasing of the square of distance to the solution. Therefore, we need $T(T+1) \geq \frac{8L}{\mu}$, but $T(T+1) \geq T^2$ as T is non-negative number, then it's sufficient to choose

$$T = \lceil \sqrt{\frac{8L}{\mu}} \rceil$$

which is obviously $O(\sqrt{\frac{L}{\mu}})$ as needed.

7. From smoothness definition

$$f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}||y-x||^2$$

using $x = x^*$ (which means $\nabla f(x^*) = 0$ as in previous) and $y = x_T$ we have

$$f(x_T) - f(x^*) \leq \frac{L}{2}||x^* - x_T||^2$$

We need $f(x_T) - f(x^*) \leq \varepsilon$, therefore it's enough to make

$$\frac{L}{2}||x^* - x_T||^2 \leq \varepsilon$$

i.e.

$$||x^* - x_T||^2 \leq \frac{2\varepsilon}{L}.$$

From Assignment 6 we know that we can make $||x^* - x_T||^2$ decrease twice in $O(\sqrt{\frac{L}{\mu}})$ iterations. If we now repeat the process from the previous subtask starting it anew several times, we need the square of distance R^2 to decrease at least $\frac{R^2 L}{2\varepsilon}$ times by decreasing it at least twice per step, which will be done in at most

$$\lceil \sqrt{\frac{8L}{\mu}} \rceil \log_2 \frac{R^2 L}{2\varepsilon}$$

which is

$$O(\sqrt{\frac{L}{\mu}} \ln(\frac{R^2 L}{2\varepsilon}))$$

as needed.