COMS 4771 Fall 2016 Homework 4

Solutions by Daniel Hsu

Problem 1

Solution.

(a) The Hessian of the objective function at any $\boldsymbol{w} \in \mathbb{R}^d$ is

$$\boldsymbol{H} \;:=\; \lambda \boldsymbol{I} + \frac{2}{|S|} \sum_{(\boldsymbol{x},y)\in S} \boldsymbol{x}\boldsymbol{x}^\top .$$

To show that $\boldsymbol{H}$ is positive definite, observe that for any non-zero vector $\boldsymbol{v}$,

$$\boldsymbol{v}^\top \boldsymbol{H} \boldsymbol{v} \;=\; \lambda \|\boldsymbol{v}\|_2^2 + \frac{2}{|S|} \sum_{(\boldsymbol{x},y)\in S} \langle \boldsymbol{v}, \boldsymbol{x}\rangle^2 \;\geq\; \lambda \|\boldsymbol{v}\|_2^2 \;>\; 0 ,$$

where the last inequality follows because $\boldsymbol{v} \neq \boldsymbol{0}$ and $\lambda > 0$. This implies that $\boldsymbol{H}$ is positive definite, which in turn implies that objective function is convex.

(b) Let the initial solution be $\boldsymbol{w}^{(1)} \in \mathbb{R}^d$, the step size in iteration t be η_t, and the number of iterations be T.

1: **for** $t = 1, 2, \dots, T$ **do**
2: $\quad \boldsymbol{\gamma}^{(t)} := \lambda \boldsymbol{w}^{(t)}$. # vector scaling
3: $\quad$ **for each** $(\boldsymbol{x}, y) \in S$ **do**
4: $\quad\quad s := 2 \times (\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}\rangle - y) / |S|$. # inner product, arithmetic
5: $\quad\quad \boldsymbol{\gamma}^{(t)} := \boldsymbol{\gamma}^{(t)} + s\boldsymbol{x}$. # vector scaling, vector addition
6: $\quad$ **end for**
7: $\quad \boldsymbol{w}^{(t+1)} := \boldsymbol{w}^{(t)} - \eta_t \boldsymbol{\gamma}^{(t)}$. # vector scaling, vector addition
8: **end for**

(It is fine to write Steps 2–6 using a summation.)

(c) **Yes**. In standard form, the i-th new constraint function is $f_i(\boldsymbol{w}) := w_i^2 - 1$. To see that f_i is convex, observe that its Hessian matrix $\boldsymbol{H}$ is all-zeros except for a 2 in the (i,i)-th position; this is positive semidefinite because $\boldsymbol{v}^\top \boldsymbol{H} \boldsymbol{v} = v_i^2 \geq 0$ for any vector $\boldsymbol{v}$.

(d) **Yes**. We can write the i-th new constraint as a pair of constraint functions $f_i(\boldsymbol{w}) := w_{2i-1} + w_{2i} - 1$ and $\tilde{f}_i(\boldsymbol{w}) := -w_{2i-1} - w_{2i} + 1 = -f_i(\boldsymbol{w})$. This is because $f_i(\boldsymbol{w}) \leq 0$ and $\tilde{f}_i(\boldsymbol{w}) \leq 0$ implies that $w_{2i-1} + w_{2i} - 1 = 0$. Each of f_i and $\tilde{f}_i$ is an affine function and hence is convex.

(e) **No**. The feasible region $\{\boldsymbol{w} \in \mathbb{R}^d : w_i^2 = 1 \,\forall i = 1, 2, \dots, d\}$ is not a convex set. To see this, observe that the vector $\boldsymbol{w} := (1, 1, \dots, 1)$ is in the set, and so is the vector $\tilde{\boldsymbol{w}} := (-1, -1, \dots, -1)$, but the vector $(\boldsymbol{w} + \tilde{\boldsymbol{w}})/2 = \boldsymbol{0}$ is not in the set, even though it is a convex combination of $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$.

□

Problem 2

Solution.

(a) Let the initial solution be $(\beta_0^{(1)}, \boldsymbol{\beta}^{(1)}) \in \mathbb{R} \times \mathbb{R}^d$, the step size in iteration t be η_t, and the number of iterations be T.

1: **for** $t = 1, 2, \dots, T$ **do**
2: $\quad g^{(t)} := 0$.
3: $\quad \boldsymbol{\gamma}^{(t)} := \boldsymbol{0}$.
4: $\quad$ **for** $i = 1, 2, \dots, n$ **do**
5: $\quad\quad s_i := \frac{1}{1+\exp(-\beta_0^{(t)} - \langle \boldsymbol{\beta}^{(t)}, \boldsymbol{x}_i\rangle)} - y_i$. # inner product, arithmetic, exp
6: $\quad\quad g^{(t)} := g^{(t)} + s_i/n$. # arithmetic
7: $\quad\quad \boldsymbol{\gamma}^{(t)} := \boldsymbol{\gamma}^{(t)} + (s_i/n)\boldsymbol{x}_i$. # vector scaling, vector addition
8: $\quad$ **end for**
9: $\quad \beta_0^{(t+1)} := \beta_0^{(t)} - \eta_t g^{(t)}$. # arithmetic
10: $\quad \boldsymbol{\beta}^{(t+1)} := \boldsymbol{\beta}^{(t)} - \eta_t \boldsymbol{\gamma}^{(t)}$. # vector scaling, vector addition
11: **end for**

(It is fine to write Steps 2–8 using a summation, and also to use the "lifting" trick to $d+1$ optimization variables.)

(b) It took 4659 iterations to achieve objective value at most 0.65064.

(c) The sample variance of the first and third features are about 400 times that of the second feature. We use the matrix

$$\boldsymbol{A} \;:=\; \begin{bmatrix} \frac{1}{20} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{20} \end{bmatrix}$$

to transform the data. It took 378 iterations to achieve objective value at most 0.65064.

(d)

	original data	transformed data
number of iterations	512	64
final objective value	0.65507	0.655698
final hold-out error rate	0.382927	0.389024

□