Name:

Convex optimization

Homework set 10

Date due: **December 18 2019, 17:21**

Explain your reasoning in all the problems.

Problem	Pts max	Pts
1	2	
2	2	
3	2	
4	2	
5	2	
Σ	10	

Problem 1. Let $n \in \mathbb{N}$ and let U be a random matrix whose all entries are independent random variables that are 1 with probability $1/2$ and -1 also with probability $1/2$. Show that the expected value of $U^T U$ is nE.

Problem 2. Let $n \in \mathbb{N}$ and let U be a random matrix whose all entries are pairwise independent random variables with mean 0 and a finite variance (different entries of U can come from different distributions). Show that then the expected value of $U^T U$ is a diagonal matrix.

Problem 3. Prove that the following problem (which in fact turns out to be rather hard) is *not* quasiconvex: Let A be a $n \times m$ matrix, $\mathbf{b}$ a vector of dimension n and $\epsilon > 0$ a number. Given these (fixed) objects, the problem is:

$$\begin{aligned} &\text{minimize the number of nonzero entries in } \mathbf{x} \\ &\text{subject to } \|A\mathbf{x} - \mathbf{b}\|_2 \leq \epsilon. \end{aligned}$$

Problem 4 (How to train your SVM)**.**

1. Download the Banknote Authentication dataset from

   `http://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt`

   This dataset (in the CSV format) contains on each line the optical properties $\mathbf{x} \in \mathbb{R}^4$ of a banknote followed by a label – either 0 for authentic banknotes or 1 for forgeries (the optical properties are derived from the wavelet transform of a photo of the banknote). Treat the label 0 as $y = -1$ and label 1 as $y = 1$.

2. We will need training and testing data. Divide the dataset into files `train.csv` and `test.csv` where the file `test.csv` contains every 10th line of the original file `data_banknote_authentication.txt` and `train.csv` all the other lines (feel free to use Python for this separation). Let N be the number of training data points (your N should be around 1240).

3. Use CVXOPT/CVXPY to find $\mathbf{a} \in \mathbb{R}^4, b \in \mathbb{R}$ that minimize

   $$\frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - y_i(\mathbf{a}^T \mathbf{x} - b)) + \epsilon \|\mathbf{a}\|_2^2,$$

   where $\mathbf{x}_i, y_i$ are the data from `train.csv` (remember that label 0 gives $y_i = -1$). Choose ϵ equal to 0.1, 1, 2, and 5.

   Send your four $\mathbf{a}, b$ together with your code to Jiří.

4. Test your four linear classifiers (given by the $\mathbf{a}$ and b) on the data from `test.csv`: For a given $\mathbf{a}, b$ and $\mathbf{x}$ the prediction of your classifier is $y = 1$ (banknote is authentic) if $\mathbf{a}^T \mathbf{x} \geq b$ and $y = -1$ (counterfeit) if $\mathbf{a}^T \mathbf{x} < b$. Calculate the predictions and compare them to the labels from the dataset. Report to Jiří the success rates of your classifiers.

Problem 5. Goblins want to compete with dwarves in gold mining. We will assume that the amount of gold mined in a given day is roughly the sum of the productivity of the goblins who went to work that day.

Fortunately, goblins happen to have an electronic record of who went to work when and how much gold was mined for the last 100 days. Unfortunately, the records for about 10 % of the days were rewritten by a self-serving goblin. These fake records should show up as outliers when you try to estimate the goblin's productivity using a norm approximation.

1. Download the file `goblins.csv` from the course website with the mining records. As with dwarves, each line lists one goblin's attendance and the final line lists the total amounts of gold mined on each day. Use least squares to estimate the goblin's productivities. List the resulting estimates here. Since least squares should be easy by now, you do *not* have to explain anything or send Jiří your code.

2. Use the idea that the days with fake records should act as outliers to identify approximately 10 days with suspicious records. This part does not have a unique solution – propose something that sounds sane. Write down your method here together with a list of suspicious days.

3. Remove the suspicious days from the records and re-run the least squares estimate for the "clean" records. Report here on how does this changes the vector of estimated productivities and which goblins were the most affected.

You can consult with your friends when solving the homework, but you have to **write** your solutions (including Python code) **on your own** and **do not show your fininished solutions** to your peers before the due date.