Problem: Convex analysis

This problem has three sub-parts:

1. Consider the following set

$$\mathcal{F} := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \leq 0 \text{ and } x_2 \leq 0\} \cup \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \geq 0 \text{ and } x_2 \geq 0\},$$

   where $\cup$ denotes the union operator. Is $\mathcal{F}$ a convex set? Justify your answer.

2. Compute the subdifferential of a function $g(x) := \max\{2x - 6, -3x + 9\}$ at the point $x = 3$. That is, compute $\partial g(3)$. Hint [1]

3. Is the function $h(x) = -2x^2$ convex? Justify your answer.

Problem: Convex analysis

This problem has three sub-parts:

1. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$

$$f(x) := x^\top A x,$$

   with matrix A given as

$$A := \begin{bmatrix} \texttt{sd}_1 - 3 & 0 \\ 5 & \texttt{sd}_2 + 1 \end{bmatrix},$$

   where you need to substitute $\texttt{sd}_1$ and $\texttt{sd}_2$ with the digits in your student ID[2]. Is f a convex function? Justify your answer.

2. Compute the subdifferential of a function $g(x) := |3x - 6|$ at the point $x = 2$. That is, compute $\partial g(2)$.

3. Is the function $h(x) = x^4$ strongly convex? Is it strictly convex? Justify your answer.

Note: Convex analysis

For additional practice problems on Convex Analysis, please look at chapter 2 and chapter 3 of Stephen Boyd's book on "Convex Optimization". Check out exercise problems at the end of the chapters.

[1] To elaborate more on the max of two functions, at a particular x the value that the function $g(x) = \max\{2x - 6, -3x + 9\}$ takes is the maximum of the two values: $f_1(x)$ and $f_2(x)$, where $f_1(x) = 2x - 6$ and $f_2(x) = -3x + 9$. As an example, the absolute value function $g(x) = |x|$ is the max of two functions $g_1(x) = x$ and $g_2(x) = -x$. It might be helpful to plot the function g and understand how it behaves.

[2] The variable $\texttt{sd}_1$ is the second last digit in your student ID. The variable $\texttt{sd}_2$ is the last digit in your student ID. For example, if your student ID is: s1234567, then $\texttt{sd}_1 = 6$ and $\texttt{sd}_2 = 7$.

Problem: Neural networks

For a neural network you are given the following information. The input vector u is of dimension 2, that is, $u = (u_1, u_2) \in \mathbb{R}^2$ and the weight vector w is of dimension 20. The output is one dimensional and the formula for the output given input and weights is given as

$$v(u, w) = \sigma\left(\sum_{k=1}^{2} w_k^{(3)} h\left(\sum_{j=1}^{3} w_{kj}^{(2)} h\left(\sum_{i=1}^{2} w_{ji}^{(1)} u_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) + w_0^{(3)}\right)$$

Answer the following:

1. How many hidden layers are there in the network? Justify your answer.
2. Draw the network showing input nodes, hidden layer nodes, output nodes, and connections. Justify how the expression of v given above lead you to this network.

Problem: Nesterov accelerated gradient method

Consider the following function $f : \mathbb{R}^3 \to \mathbb{R}$ given by

$$f(x) = \|x\|^4,$$

where $\|x\|$ is the Euclidean norm of the vector [3]. We will implement the Nesterov accelerated gradient method and the gradient descent to minimize this function. Note first that the minimizer of the function is $x = (0, 0, 0)^\top$ and the optimal cost is zero. Do the following

1. Compute the analytical expression of the gradient of f at any point x, that is, compute $\nabla f(x)$. See hint [4].
2. Compute the bound L on the norm of the gradient over the set $\mathcal{X} = \{x : \|x\| \leq 1\}$, that is, find an $L > 0$ such that
$$\|\nabla f(x)\| \leq L, \text{ for all } x \in \mathcal{X}.$$
3. Implement the Nesterov's accelerated gradient method using the above calculated L (in case you could not find L in the above part, then set $L = 6$) and using the initial x to be a random vector belonging to the set $\mathcal{X}$. Note that the function f is not strongly convex, so you need to use the second Nesterov's algorithm presented in the lecture where you need to compute the stepsizes $\{\gamma_k\}$ using the sequence $\{\lambda_k\}$. Select 100 iterations. Plot the cost function and the $\|x\|$ at each iteration of the algorithm.
4. Implement the gradient descent method with the initial condition equal to the initial condition that you picked for the Nesterov's accelerated gradient method in part (3) above. Use constant step size for each iteration. Vary your step size value and pick one that gives you convergence. Plot the cost function and the $\|x\|$ at each iteration of the algorithm.
5. Compare the two methods (Nesterov's accelerated gradient method and the gradient descent) by plotting, in one figure, the costs and $\|x\|$ for both methods. Discuss which method is better.

[3] Hint: The formula is $\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$ where the vector $x = (x_1, x_2, x_3)$

[4] Hint: Note that the gradient in this case means the vector

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \frac{\partial f(x)}{\partial x_3} \end{bmatrix}$$

Problem: Sample average approximation

Consider the following stochastic optimization problem

$$\begin{aligned} \min \quad & f(x) && \text{(1a)} \\ \text{subject to} \quad & x \geq 0, && \text{(1b)} \end{aligned}$$

where the function $f : \mathbb{R}^2 \to \mathbb{R}$ is given as

$$f(x) = x^\top \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix} x - \mathbb{E}[u] \begin{bmatrix} 8 \\ 16 \end{bmatrix}^\top x,$$

and u is a real-valued (scalar) random variable uniformly distributed over the interval $[0, 1]$. Here, note that x is two-dimensional, that is, $x \in \mathbb{R}^2$. The inequality $x \geq 0$ implies that both components of the vector x are non-negative. Do the following:

1. Find the optimizer x^* and the optimal value $f(x^*)$ of the optimization problem (1). Use the CVX toolbox if necessary.

2. Find the sample average approximate solution of the problem taking $N = 15$ i.i.d samples of the uncertainty u (you can use the rand function in MATLAB to do this). Repeat this approximation 100 times. For each of the sample average approximation solution, calculate

$$\|\widehat{x}_N - x^*\|,$$

   where $\widehat{x}_N$ is the solution of the sample average approximation scheme and x^* is the solution of the stochastic optimization problem (1). Plot the cumulative distribution function of $\|\widehat{x}_N - x^*\|$. Discuss your findings.[5]

3. Repeat the above part with $N = 50$ i.i.d samples for the uncertainty (keeping the number of runs to be 100). What difference do you notice in the cumulative distribution function as compared to the case of $N = 15$ samples? Discuss your findings.

Problem: True/false

Identify the following statements to be true or false. Justify your answer.

1. Every run of a stochastic gradient descent algorithm with the same initial condition and the same step sizes might create different iterate values.

2. Larger the value of the margin of a learned hyperplane in a support vector machine based classification example, worse is the separation between the classes.

3. For any given online optimization algorithm, bigger is the class of functions from which costs are selected by the "environment", bigger is the regret incurred by the algorithm.

[5] Note that for a vector $v \in \mathbb{R}^2$, the notation $\|v\|$ means the Euclidean 2-norm. That is, $\|v\| = \sqrt{v_1^2 + v_2^2}$.

Problem: True/false

Identify the following statements to be true or false. Justify your answer.

1. The following set is convex:

$$\mathcal{F} := \{x \in \mathbb{R}^2 \mid x_1 + x_2 = 1, x_1 \geq 0, \text{ and } x_2 \geq 0\}.$$

2. Consider the following two chance-constrained optimization problems (as studied in lecture 7)

$$\begin{array}{lcl} & \min & c(x) \\ \text{(Problem 1)} & \text{subject to} & \mathbb{P}(g(x,\xi) \leq 0) \geq 1 - \alpha_1 \\ & & x \in \mathcal{X}, \end{array}$$

and

$$\begin{array}{lcl} & \min & c(x) \\ \text{(Problem 2)} & \text{subject to} & \mathbb{P}(g(x,\xi) \leq 0) \geq 1 - \alpha_2 \\ & & x \in \mathcal{X}. \end{array}$$

The difference between the two problems is the probability with which the uncertain constraint $g(x,\xi) \leq 0$ is required to hold. In Problem 1, this value is $1 - \alpha_1$ and in Problem 2, it is $1 - \alpha_2$. We have the following statement: if $0 < \alpha_1 \leq \alpha_2 < 1$, then the optimal value of Problem 1 is not higher than the optimal value of Problem 2.

3. In a classification problem over input $u = (u_1, u_2)$ of two dimensions, you ended up with the learned separating hyperplane as $u_1 = 0$ (that is, the vertical axis in the (u_1, u_2) plane. All points $u = (u_1, u_2)$ with $u_1 > 0$ are assigned class "1" and points with $u_1 < 0$ as "-1". Suppose the point closest to the separating hyperplane in the set of points in class "1" is $(2, 5)$ and the point closest to the separating hyperplane in the set of points in class "-1" is $(-1.5, 4)$. Then, the margin value is 1.5.