Convexity & Machine Learning
Assignment 1

Please send me

- the **original script** detailing your computations.
  - The script must be **documented**, i.e. the code corresponding to each answer must be delimited and your loops/variables briefly explained.
  - The script must be **executable**: by just running your script, all results should appear **automatically**.
  - Do not use external functions, everything must be coded **by yourself** using elementary linear algebra functions.
- A **document** (in pdf format) which will contain your answer and your analysis. Provide illustrations and graphs but do not put code in the pdf.

This homework is due **December 16th (Fri.) 23:59 PM**

Send your homework at mcuturi@i.kyoto-u.ac.jp

Preliminary questions on convexity

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and **bounded**, that is, there exists $M \in \mathbb{R}$ such that $\forall x, f(x) \leq M$. What can you say about f?
- **Kullback-Leibler Divergence**: Let D_{kl} be the Kullback-Leibler divergence between two vectors $x, y \in \mathbb{R}^d_+$, defined as

$$D_{\text{kl}}(x, y) = \sum_{i=1^n} \left(x_i \log(\frac{x_i}{y_i}) - x_i + y_i \right).$$

  Prove the inequality

$$D_{\text{kl}}(x, y) \geq 0$$

  and show that $D_{\text{kl}}(x, y) = 0$ if and only if $x = y$. *hint* use the fact that $D_{\text{kl}}(x, y) = f(x) - f(y) - \nabla f(y)^T(x - y)$ with $f(x) = \sum_{i=1}^d x_i \log x_i$.

- Give examples of two functions f and g from $\mathbb{R}$ to $\mathbb{R}$, where f is strictly convex and g is strictly concave, with (1) $f+g$ convex, (2) $f+g$ concave and (3) $f+g$ neither convex nor concave respectively.

Gradient Descent, Logistic Regression

- **Gradient Descent**: Implement gradient descent (Boyd and Vandenberghe, 2005, Algo.9.3) with a backtracking line search (Boyd and Vandenberghe, 2005, Algo.9.2). Your function should consider six parameters:
  - f (the function to minimize)
  - ∇f (the gradient)
  - η (the stopping threshold);
  - α and β (the parameters of the backtracking line search);
  - x_0 (the starting point).

  Your function should check that the parameters $\alpha < \frac{1}{2}$ and $0 < \beta < 1$. Your function should terminate once the stopping condition

  $$\|\nabla f(x)\|^2 \leq \eta$$

  is verified, or whenever the algorithm has reached the maximal number of iterations $t_{\max} = 200$. Your function should return a list of triplets $(x, f(x), i)$ for each iteration which describes:
  - the current point x,
  - its objective function value $f(x)$,
  - the total number of calls to f or ∇f which have taken place during this iteration (including inside the line search).
- **Logistic Regression as a Convex Programming Problem**: We consider now a simple binary classification problem where we are given a database of pairs of observations $\{(\mathbf{x}_j, y_j)\}_{j=1,\ldots,N}$ (also called a **training set**) which we study using logistic regression (you can read the short reminder on logistic regression in the appendix of this homework if you are not familiar with this technique).

In such a setting, the training set is used to estimate the parameters $\mathbf{c}$ and b by defining the likelihood function of $(\mathbf{c}, b)$ with respect to the dataset,

$$\mathcal{L}(\mathbf{c}, b) = \prod_{j=1}^{N} g(\mathbf{c}^T \mathbf{x}_j + b)^{y_j} (1 - g(\mathbf{c}^T \mathbf{x}_j + b))^{1-y_j}.$$

The parameters $\mathbf{c}$ and b can be estimated using the Maximum Likelihood principle, that is by maximizing $\mathcal{L}(\mathbf{c}, b)$.

- Show that maximizing the likelihood is equivalent to minimizing the following function

$$\min_{\mathbf{c},b} - \sum_{j=1}^{N} y_j (\mathbf{c}^T \mathbf{x}_j + b) - \log(1 + e^{\mathbf{c}^T \mathbf{x}_j + b}). \tag{1}$$

- Show that the function in Equation (1) is a **convex** function of $\mathbb{R}^{d+1}$.

- **Estimating Logistic Parameters Using Gradient Descent**: Download the Wisconsin Breast Cancer[1] binary classification dataset.
  - Split randomly the dataset into 2 subsets, the training subset and the test subset.
  - Compute estimates for vectors $\mathbf{c}$ and b using your implementation of gradient descent to obtain a solution to Equation (1). Use arbitrary values for your gradient descent parameters.
  - Provide now a report of the following quantities,
    - log-likelihood on train set,
    - train classification error,
    - test classification error,
    - total number of function calls,

    for different parameters of α, β.

[1] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

Appendix: A short reminder on Logistic Regression

We consider pairs $(\mathbf{x}, y)$ where a feature vector $\mathbf{x} \in \mathbb{R}^n$ is paired with a label $y \in \{0, 1\}$. The feature vector might for instance describe the behavior of a user on a website (number of visited pages, time spent *etc.*) and the label describe whether the user has bought a product or not. We would like to define a tool that can predict y based only on $\mathbf{x}$.

In order to do so, we suppose that there is a probability density $p(X, Y)$ on couples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{0, 1\}$ which can quantify this relationship. The ratio

$$r(\mathbf{x}) = \frac{p(Y = 1|X = \mathbf{x})}{p(Y = 0|X = \mathbf{x})}$$

is called the odds-ratio of a given point $\mathbf{x}$ and measures how times more likely it is that, given a point $\mathbf{x}$, this point has a label 1 rather than a label 0. Obviously, if $r(\mathbf{x}) > 1$, then it is more likely that $y = 1$ than $y = 0$ for that particular feature vector $\mathbf{x}$. On the contrary, if $r(\mathbf{x}) < 1$, then one is tempted to guess that $y = 0$ than $y = 1$. In other words, if for a given observation $\mathbf{x}$,

$$\log \frac{p(Y = 1|X = \mathbf{x})}{p(Y = 0|X = \mathbf{x})}, \quad \begin{cases} > 0 \text{ then } y = 1 \text{ is the likely answer} \\ < 0 \text{ then } y = 0 \text{ is the likely answer} \end{cases}$$

Logistic regression is a classification tool which assumes that the log-odds ratio r follows a linear relationship

$$\log \frac{p(Y = 1|X = \mathbf{x})}{p(Y = 0|X = \mathbf{x})} \approx \mathbf{c}^T \mathbf{x} + b,$$

where $\mathbf{c} \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

Since $p(Y = 0|X = \mathbf{x}) = 1 - p(Y = 1|X = \mathbf{x})$, we hence have

$$\log \frac{p(Y = 1|X = \mathbf{x})}{1 - p(Y = 1|X = \mathbf{x})} = \mathbf{c}^T \mathbf{x} + b,$$

which implies that, using the notation $g(u) = \frac{1}{1+e^{-u}}$,

$$p(Y = 1|X = \mathbf{x}) = \frac{1}{e^{-(\mathbf{c}^T \mathbf{x} + b)} + 1} = g(\mathbf{c}^T \mathbf{x} + b).$$