HW Lesson 2: Probability and Training Algorithms

Remark:

- You can choose any one format of Handwriting, LaTex, Microsoft Word, Mac Pages or Jupyter to finish this homework and then update to Canvas. (If you use Microsoft Word, Mac Pages or other softwares, please transfer your file to PDF version.)
- Please update the file with name "HWLesson2_YourName.pdf" or "HWLesson2_YourName.ipynb" (if you use Jupyter).

Problem 1 The definition of variance for random variable is as follows

$$\mathrm{Var}[X] = \mathbb{E}[X - \mathbb{E}[X]]^2 = \mathbb{E}[X^2] - (E[X])^2.$$

Assume that X and Y are independent random variables, verify thee following properties:

1. **(10 %)**
$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$
2. **(10 %)**
$$\mathrm{Var}[XY] = \mathbb{E}[X^2]\mathbb{E}[Y^2] - (\mathbb{E}[X])^2(\mathbb{E}[Y])^2.$$
3. **(10 %)** Furthermore, if $\mathbb{E}[X] = 0$,
$$\mathrm{Var}[XY] = \mathrm{Var}[X]\mathbb{E}[Y^2].$$

Hint: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X and Y are independent random variables.

Problem 2 Verify the following properties of convex functions:

1. **(10 %)** If $f(x)$, $g(x)$ are both convex, then $\alpha f(x) + \beta g(x)$ is also convex, if $\alpha, \beta \geq 0$.
2. **(10 %)** Linear function is both convex and concave. Here, $f(x)$ is concave if and only if $-f(x)$ is convex.
3. **(10 %)** If $f(x)$ is a convex function on $\mathbb{R}^n$, then $g(y) = f(Wy + b)$ is a convex function on $\mathbb{R}^m$, where $W \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$.

Problem 3

1. **(5 %)** Given an infinitely differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$, its Hessian matrix is given by

$$H(x) := \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial {x_1}^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial {x_n}^2} \end{bmatrix}.$$

   Prove the following identity:

$$f(x) = f(y) + \nabla f(y) \cdot (x - y) + \frac{1}{2}(x - y)^T H(y_\beta)(x - y),$$

   where

$$y_\beta = \beta x + (1 - \beta) y,$$

   for some $\beta \in [0, 1]$.

   Hint: Given x and y, consider the following one-variable function of t:

$$g(t) = f(tx + (1 - t)y) \tag{1}$$

   and use the following Taylor expansion (which you do not need to prove):

$$g(1) = g(0) + g'(0) + \frac{1}{2} g''(\beta), \tag{2}$$

   which holds for some $\beta \in [0, 1]$.

2. **(5 %)** Based on the previous result, prove that if $f(x)$ is infinitely differentiable and $v^T H(x) v \geq 0$ for any $v \in \mathbb{R}^n$, then $f(x)$ must be a convex function.

   Hint: $f(x)$ is differentiable on $\mathbb{R}^n$ and $f(x) \geq f(y) + \nabla f(y) \cdot (x - y)$, $\forall x, y \in \mathbb{R}^n$, then $f(x)$ is convex.

3. **(5 %)** Consider these infinitely differentiable functions:

$$g : \mathbb{R} \mapsto \mathbb{R}, \quad h : \mathbb{R}^k \mapsto \mathbb{R},$$

   and their composition function

$$f(x) = g \circ h(x) = g(h(x)).$$

   Use Chair rule to verify that the Hessian matrix of $f(x)$ is

$$\nabla^2 f(x) = g''(h(x)) \nabla h(x) \left(\nabla h(x)\right)^T + g'(h(x)) \nabla^2 h(x).$$

4. **(5 %)** In view of previous results, define

$$g(y) = \log y, \quad h(x) = \sum_{i=1}^{k} e^{x_i},$$

and

$$f(x) = g(h(x)) = \log\left(\sum_{i=1}^{k} e^{x_i}\right).$$

Verify the following identity:
(3)

$$\nabla^2 f(x) = -\frac{1}{h^2(x)} \begin{pmatrix} e^{x_1} \\ e^{x_2} \\ \vdots \\ e^{x_k} \end{pmatrix} \begin{pmatrix} e^{x_1} & e^{x_2} & \cdots & e^{x_k} \end{pmatrix} + \frac{1}{h(x)} \begin{pmatrix} e^{x_1} & 0 & \cdots & 0 \\ 0 & e^{x_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{x_k} \end{pmatrix}.$$

Hint: Apply the result for last problem.

5. **(5 %)** Consider the Hessian matrix $H(x) = \nabla^2 f(x)$ given by (3), verify that

$$v^T H(x) v = -\frac{\left(\sum_{i=1}^{k} v_i e^{x_i}\right)^2}{h^2(x)} + \frac{\sum_{i=1}^{k} v_i^2 e^{x_i}}{h(x)},$$

for any $v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix} \in \mathbb{R}^k$.

Hint: Use the properties of matrix multiplication: $v^T(xx^T)v = (v^T x)(v^T x)$.

6. **(5 %)** Prove that

$$v^T H(x) v = -\frac{\left(\sum_{i=1}^{k} v_i e^{x_i}\right)^2}{h^2(x)} + \frac{\sum_{i=1}^{k} v_i^2 e^{x_i}}{h(x)} \geq 0,$$

for any $v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix} \in \mathbb{R}^k$.

Hint: The following inequality can be verified by Cauchy-Schwartz inequality

$$\left(\sum_{i=1}^{k} v_i e^{x_i}\right)^2 \leq \sum_{i=1}^{k} e^{x_i} \sum_{i=1}^{k} v_i^2 e^{x_i}.$$

7. **(5 %)** Prove that the following function is convex

$$f(x) = \log\left(\sum_{i=1}^{k} e^{x_i}\right).$$

Hint: Apply the results from previous problems

8. **(5 %)** For any given vector $X \in \mathbb{R}^d$, let consider the function $f(\theta)$ given by

$$f(\theta) = \log\left(\sum_{i=1}^{k} e^{w_i \cdot X + b_i}\right), \tag{4}$$

where $\theta = \begin{pmatrix} w_1 \\ b_1 \\ \vdots \\ w_k \\ b_k \end{pmatrix} \in \mathbb{R}^n$ with $w_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ for $i = 1, 2, \cdots, k$ and we know that $n = k(d+1)$. Prove that $f(\theta)$ is a convex function.

Hint: Consider the last result for $f(x) = \log\left(\sum_{i=1}^{k} e^{x_i}\right)$ and the last properties in Problem 2 in this homework.

Optional Problems

Problem 1 (10 pts) Prove that if $f(x)$ is differentiable and convex on $\mathbb{R}^n$, then $f(x) \geq f(y) + \nabla f(y) \cdot (x - y), \forall x, y \in \mathbb{R}^n$.

Problem 2 (25 pts) Assume we have a data set $\{(X_j, Y_j)\}_{j=1}^{N}$ where

$$X_j \in \mathbb{R}^d, \quad Y_j \in R^k,$$

and all component of Y_j are non-negative for all j. Consider the logistic regression model $\boldsymbol{p}(X; \theta)$ with

$$\boldsymbol{p}(X; \theta) = \frac{1}{\sum_{i=1}^{k} e^{w_i \cdot X + b_i}} \begin{pmatrix} e^{w_1 \cdot X + b_1} \\ e^{w_2 \cdot X + b_2} \\ \vdots \\ e^{w_k \cdot X + b_k} \end{pmatrix} = \begin{pmatrix} p_1(X; \theta) \\ p_2(X; \theta) \\ \vdots \\ p_k(X; \theta) \end{pmatrix}. \tag{5}$$

Prove that

$$L(\theta) = \sum_{j=1}^{N} H(Y_j, \boldsymbol{p}(X_j; \theta)) = \sum_{j=1}^{N} \sum_{i=1}^{k} -[Y_j]_i \log(p_i(X_j; \theta)) \tag{6}$$

is a convex function where $[Y_j]_i \geq 0$. (Here $[Y_j]_i$ means the i-th component of Y_j.)

Hint: Use these properties in Problem 2 and results in Problem 3.

Problem 3 (15 pts) Furthermore, consider the loss function of logistic regression model with regularization term as

$$L_\lambda(\theta) = L(\theta) + \lambda\|\theta\|_F^2,$$

where $L(\theta)$ is defined in (6). Prove that $L_\lambda(\theta)$ is a λ-strongly convex function.

Hint: Recall the theorem in the notes that $L(x)$ is a differentiable and convex function on $\mathbb{R}^n$ if and only if $L(x) \geq L(y) + \nabla L(y) \cdot (x - y), \forall x, y \in \mathbb{R}^n$. Also use the example of $f(x) = \|x\|^2$ for λ-strongly convex functions.