MAT 180 Theoretical Homework 2

Problem 1: Compute $E_P[\mathrm{x}]$ and $\mathrm{Var}_P(\mathrm{x})$ where P is the categorical distribution with sample space $\Omega = \{1, \ldots, k\}$ and $P(\mathrm{x} = i) = p_i$ for $1 \leq i \leq k$.

Problem 2: Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be a matrix whose rows are data points in some dataset. Let $\mathbf{y} \in \mathbb{R}^m$ be a label vector so the value y_i is associated to the data point $\mathbf{X}_{i,*}$ for each i. Let $\hat{\mathbf{X}}$ denote the matrix $\mathbf{X}$ with a column of 1's inserted as the new left-most column, so $\hat{\mathbf{X}} \in \mathbb{R}^{m \times (n+1)}$. Define the cost function $J(\mathbf{v}) = \frac{1}{m}||\hat{\mathbf{X}}\mathbf{v} - \mathbf{y}||_2^2$ and compute both the gradient and the Hessian

$$\frac{\partial}{\partial \mathbf{v}} J(\mathbf{v}) = \nabla_{\mathbf{v}} J(\mathbf{v})^\top \quad \text{and} \quad \frac{\partial^2}{\partial \mathbf{v}^2} J(\mathbf{v}) = H(J)(\mathbf{v})$$

using the matrix differentiation properties discussed in lecture. Show that the Hessian is positive definite as long as the columns of $\hat{\mathbf{X}}$ are linearly independent, and that otherwise the Hessian is positive semidefinite. Show that J is convex and that $\nabla_{\mathbf{v}} J(\mathbf{v})$ is Lipshchitz continuous. This result will be useful for one of the most basic machine learning algorithms: linear regression. Hint: use a formula for the 2-norm squared involving the transpose.

Problem 3: Let $\mathbf{X}, \mathbf{y}, \hat{\mathbf{X}}$ be as in the previous problem, but now the entries of $\mathbf{y}$ are in $\{0, 1\}$. Define the cost function

$$J(\mathbf{v}) = \frac{1}{m}[-(\mathbf{1} - \mathbf{y})^\top \mathbf{Log}(\mathbf{1} - \mathbf{G}(\hat{\mathbf{X}}\mathbf{v})) - \mathbf{y}^\top \mathbf{Log}(\mathbf{G}(\hat{\mathbf{X}}\mathbf{v}))]$$

where $\mathbf{G} : \mathbb{R}^m \to \mathbb{R}^m$ is defined by $\mathbf{G}(x_1, \ldots, x_m) = (g(x_1), \ldots, g(x_m))$ where g is the sigmoid function $g(z) = \frac{1}{1+e^{-z}}$, and $\mathbf{Log} : \mathbb{R}^m \to \mathbb{R}^m$ is defined by $\mathbf{Log}(x_1, \ldots, x_m) = (\log(x_1), \ldots, \log(x_m))$ where log is the natural log function. Note $\mathbf{1} \in \mathbb{R}^m$ is a vector of all 1's. Compute the gradient and the Hessian

$$\frac{\partial}{\partial \mathbf{v}} J(\mathbf{v}) = \nabla_{\mathbf{v}} J(\mathbf{v})^\top \quad \text{and} \quad \frac{\partial^2}{\partial \mathbf{v}^2} J(\mathbf{v}) = H(J)(\mathbf{v})$$

using the matrix differentiation properties discussed in lecture. Show that the Hessian is positive definite as long as the columns of $\hat{\mathbf{X}}$ are linearly independent and that otherwise the Hessian is positive semidefinite. Show that J is convex and that $\nabla_{\mathbf{v}} J(\mathbf{v})$ is Lipshchitz continuous. This result will be useful for another one of the most basic machine learning algorithms: logistic regression. Hints: It will be useful to use the fact that $g'(z) = g(z)(1 - g(z))$. You will also need to use the chain rule for matrix differentiation

$$\frac{\partial}{\partial \mathbf{v}}(f \circ g)(\mathbf{v}) = \frac{\partial f}{\partial \mathbf{v}}(g(\mathbf{v}))\frac{\partial g}{\partial \mathbf{v}}(\mathbf{v}).$$

You will need to compute the derivatives of **Log** and **G**, which in both cases should be diagonal matrices.