1. (**Math**) Nonlinear least-squares. Suppose that $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_m(\mathbf{x})) : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n, \mathbf{f} \in \mathbb{R}^m$ and some $f_i(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is a (are) non-linear function(s). Then, the problem,

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{f}(\mathbf{x})\|_2^2 = \arg\min_{\mathbf{x}} \frac{1}{2} (\mathbf{f}(\mathbf{x}))^T \mathbf{f}(\mathbf{x})$$

is a nonlinear least-squares problem. In our lecture, we mentioned that Levenberg-Marquardt algorithm is a typical method to solve this problem. In L-M algorithm, for each updating step, at the current **x**, a local approximation model is constructed as,

$$\begin{aligned} L(\mathbf{h}) &= \frac{1}{2}(\mathbf{f}(\mathbf{x}+\mathbf{h}))^T \mathbf{f}(\mathbf{x}+\mathbf{h}) + \frac{1}{2}\mu \mathbf{h}^T \mathbf{h} \\ &= \frac{1}{2}(\mathbf{f}(\mathbf{x}))^T \mathbf{f}(\mathbf{x}) + \mathbf{h}^T (\mathbf{J}(\mathbf{x}))^T \mathbf{f}(\mathbf{x}) + \frac{1}{2}\mathbf{h}^T (\mathbf{J}(\mathbf{x}))^T \mathbf{J}(\mathbf{x}) \mathbf{h} + \frac{1}{2}\mu \mathbf{h}^T \mathbf{h} \end{aligned}$$

where **J**(**x**) is **f**(**x**)'s Jacobian matrix, and $\mu > 0$ is the damped coefficient. Please prove that $L(\mathbf{h})$ is a strictly convex function. (Hint: If a function $L(\mathbf{h})$ is differentiable up to at least second order, L is strictly convex if its Hessian matrix is positive definite.)

we know $h \in R^n$, and $L(h)$ is differentiable at second order, $\operatorname{dom} L \in R^n$, then

$$\nabla L(h) = (J(x))^T f(x) + \frac{1}{2}\left(((J(x))^T J(x) + ((J(x))^T J(x))^T\right) h + \mu h$$

$$= (J(x))^T f(x) + (J(x))^T J(x) h + \mu h$$

so $\nabla^2 L(h) = J(x)^T J(x) + \mu I \quad (\mu > 0)$

note $J(x)$ as J, $J^T J \in S^n$, $I \in S^n$, so $J^T J + \mu I \in S^n$

we need to prove $J^T J + \mu I$ is positive definite

$\forall x \neq 0$, we have $x^T J^T J x = (Jx)^T (Jx) = \|Jx\|_2^2 \geq 0$

then $J^T J$ is positive semi-positive

for $J^T J$'s eigenvalues λ_i, $i = 1, ..., n$, we have $\lambda_i \geq 0$ for positive semi-positive matrix

$J^T J x_i = \lambda_i x_i$

$(J^T J + \mu I) x_i = (\lambda_i + \mu) x_i$, $\lambda_i + \mu$ is the eigenvalue for $J^T J + \mu I$, and $\mu > 0 \Rightarrow \lambda_i + \mu > 0$

thus $J^T J + \mu I$, the Hessian matrix of $L(h)$

is positive definite

then $L(h)$ is strictly convex

2. (**Math**) In our lecture, we mentioned that for logistic regression, the cost function is,

$$J(\boldsymbol{\theta}) = -\sum_{i=1}^{m} y_i \log\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right) + (1-y_i)\log\left(1-h_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right)$$

Please verify that the gradient of this cost function is

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{i=1}^{m} \boldsymbol{x}_i \left(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\right)$$

$$h_\theta(x) = \sigma(\theta^T x + b) = \frac{1}{1+e^{-(\theta^T x+b)}}$$

Let $t = \theta^T x + b$, $\frac{\partial t}{\partial \theta} = x$

$$\frac{\partial h_\theta(x)}{\partial t} = \frac{\partial \sigma(t)}{\partial t} = \sigma(t)(1-\sigma(t)) = h_\theta(x)(1-h_\theta(x))$$

so: $$\nabla_\theta J(\theta) = -\sum_{i=1}^{m}\Big(y_i \frac{\partial \log(h_\theta(x_i))}{\partial h_\theta(x_i)} \cdot \frac{\partial h_\theta(x_i)}{\partial t} \cdot \frac{\partial t}{\partial \theta} + (1-y_i)\frac{\partial \log(1-h_\theta(x_i))}{\partial h_\theta(x_i)} \cdot \frac{\partial h_\theta(x_i)}{\partial t} \cdot \frac{\partial t}{\partial \theta}\Big)$$

$$= -\sum_{i=1}^{m}\Big(y_i \frac{1}{h_\theta(x_i)} \cdot h_\theta(x_i)(1-h_\theta(x_i))\, x_i + (1-y_i)\frac{-1}{(1-h_\theta(x_i))} h_\theta(x_i)(1-h_\theta(x_i))\, x_i\Big)$$

$$= -\sum_{i=1}^{m} y_i(1-h_\theta(x_i))\, x_i - (1-y_i)\, h_\theta(x_i)\, x_i$$

$$= -\sum_{i=1}^{m} x_i(y_i - h_\theta(x_i))$$

$$= \sum_{i=1}^{m} x_i(h_\theta(x_i) - y_i)$$