1. (a) $g(z) = J(w) = \sum_{i=1}^{n}\left[y_i - \frac{1}{w_0 + \sum_{j=1}^{d} w_j x_{ij}} \right]^2$.

$z = Aw$

$$A = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & \cdots & \cdots & x_{nd} \end{bmatrix} \qquad w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \qquad z = Aw = \begin{bmatrix} w_0 + \sum_{j=1}^{d} x_{1j} w_j \\ w_0 + \sum_{j=1}^{d} x_{2j} w_j \\ \vdots \\ w_0 + \sum_{j=1}^{d} x_{nj} w_j \end{bmatrix}$$

$\therefore J(w) = \sum_{i=1}^{n} (y_i - \frac{1}{z_i})^2 = g(z)$. factorizable.

(b) $\nabla_w J(w) = \begin{bmatrix} \partial f(w)/\partial w_0 \\ \vdots \\ \partial f(w)/\partial w_d \end{bmatrix} = \nabla g(z) \cdot \frac{dz}{dw}$

$\nabla_z g(z) = 2\sum_{i=1}^{n} (y_i - \frac{1}{z_i}) \cdot (\frac{1}{z_i^2})$.

$dz/dw = \begin{cases} 1 & (j=0) \\ x_{ij} & (j>0) \end{cases}$.

$\therefore \nabla J(w) = 2\sum_{i=1}^{n} (y_i - \frac{1}{z_i}) \cdot (\frac{1}{z_i^2}) \cdot \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix}$.

(c). $f(w^{k+1}) = f(w^k) + \nabla f(w^k) \cdot (w^{k+1} - w^k) + O\|w^{k+1} - w^k\|^2$

$= f(w^k) + \alpha \cdot \nabla f(w^k) \cdot \nabla f(w^k) + O(\alpha^2)$

$= f(w^k) + \alpha \|\nabla f(w^k)\|^2 + O(\alpha^2)$.

where $w^{k+1} = w^k - \alpha \| dJ/dw \|$

$\|\nabla J(w)\|^2 = 4\sum_{i=1}^{n} (y_i - \frac{1}{w_0 + \sum_{j=1}^{d} x_{ij} w_j})^2 \cdot \frac{1}{(w_0 + \sum_{j=1}^{d} x_{ij} w_j)^4} \cdot (1 + \sum_{j=1}^{d} x_{ij}^2)$.

hw5_1

March 18, 2018

```
In [ ]: import numpy as np
        def getf(A,Y,w):
            z=np.def(A,w)
            J=np.sum((Y-1/z)**2)
            return J
        def getGrad(A,Y,w):
            z=np.dot(A,w)
            Grad=np.zeros(w.shape[0],1)
            for i in range(A.shape[0]):
                Grid[i]=2*np.sum((Y-1/z)*(1/z**2)*A[i,;])
            return Grad
```

2. (a) $\nabla J(w) = [b_1 w_1 \; b_2 w_2]^T$

(b) In order to get the minimum w^*, we set $\nabla J(w) = 0$. $\therefore w^* = 0$.

(c). $w^{k+1} = w^k - \alpha \nabla J(w^k)$

$w^{k+1} = w_i^k - \alpha b_i w_i^k = p_i w_i^k$. where $p_i = 1 - b_i \alpha$

(d) $\because w^k \to w^* = 0$. we need $|p_i| < 1$ for $i = 1, 2$.

$\therefore |1 - b_i \alpha| < 1$. $\therefore \alpha < \frac{2}{b_i}$.

(e). $\alpha = 2/(b_1 + b_2)$.

$\therefore p_1 = 1 - b_1 \alpha = \frac{b_2 - b_1}{b_1 + b_2}$. $p_2 = \frac{b_1 - b_2}{b_1 + b_2}$.

~~非 we~~ $w_i^k = p_i^k w_i^0$ so $|w_i^k| = C|w_i^0|$

$\therefore \|w^k\|^2 = |w_1^k|^2 + |w_2^k|^2 = C^{2k}[|w_1^0|^2 + |w_2^0|^2] = C^{2k} \|w^0\|^2$

$\therefore \|w^k\| = C\|w^0\|$.

3. (a) $\cancel{\#}\ z_i = x_i^T P x_i = \sum_{j,k} x_{ij} x_{ik} P_{jk}$

$\therefore \partial z_i / \partial P_{jk} = x_{ij} x_{ik}$.

$\therefore \nabla_P z_i = [x_{ij} x_{ik}]_{jk} = x_i x_i^T$

(b) By the chain rule

$$\partial J / \partial P_{jk} = \sum_{i=1}^{n} \frac{\partial J}{\partial z_i} \cdot \frac{\partial z_i}{\partial P_{jk}} = \sum_{i=1}^{n} \left[\frac{1}{y_i} - \frac{1}{z_i} \right] \frac{\partial z_i}{\partial P_{jk}}.$$

$$\therefore \nabla_P J = \sum_{i=1}^{n} \left[\frac{1}{y_i} - \frac{1}{z_i} \right] \cdot x_i x_i^T$$

hw5_3

March 18, 2018

In [1]:
```
import numpy as np
n=X.shape[0]
z=np.zeros(n)
for i in range(n):
    z[i]=X[i,:].dot(P.dot(X[i,:]))
J=np.sum(z/y-np.log(z))
g=1/y-1/z
Jgrad=np.zeros((n,n))
for i in range(n):
    xi=X[i,:]
    Jgrad += g[i]*xi[:,None]*xi[None,:]
```

```
---------------------------------------------------------------------------

NameError                                 Traceback (most recent call last)

<ipython-input-1-fb5251638f50> in <module>()
      1 import numpy as np
----> 2 n=X.shape[0]
      3 z=np.zeros(n)
      4 for i in range(n):
      5     z[i]=X[i,:].dot(P.dot(X[i,:]))


NameError: name 'X' is not defined
```

In [ ]:
```
Xp=X.dot(P)
z=np.sum(Xp*X,axis=1)
J=np.sum(z/y-np.log(z))
g=1/y-1/z
Gx=g[:,None]*X
Jgrad=X.T.dot(Gx)
```

4. (a) We set $\hat{w}_2 = \arg\min_{w_2} J(w_1, w_2)$.

$\therefore J_1(w_1) = J(w_1, \hat{w}_2)$

$$\nabla_{w_1} J_1(w_1) = \frac{\partial J(w_1, \hat{w}_2)}{\partial w_1}$$

$$= \left.\frac{\partial J(w_1, w_2)}{\partial w_1}\right|_{w_2=\hat{w}_2} + \left.\frac{\partial J(w_1, w_2)}{\partial w_2}\right|_{w_2=\hat{w}_2} \cdot \frac{\partial w_2}{\partial w_1}$$

Meanwhile $\left.\nabla_{w_2} J(w_1, w_2)\right|_{w_2=\hat{w}_2} = 0$, since $\hat{w}_2$ minimize $J(w_1, w_2)$ over w_2.

$$\therefore \nabla_{w_1} J(w_1) = \nabla_{w_1} J(w_1, w_2)\Big|_{w_2=\hat{w}_2}.$$

(b). We could see the loss function as $J = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

As we learn in question 1

$$\hat{y}_i = Ab. \quad \text{where} \quad A = \begin{bmatrix} e^{-a_1 x_1} & \cdots & e^{-a_d x_1} \\ \vdots & & \vdots \\ e^{-a_1 x_n} & & e^{-a_d x_n} \end{bmatrix}$$

and $\hat{b} = \arg\min_b J(a, b) = (A^T A)^{-1} A^T Y$.

(c). $$\nabla_{a_j} J(a, b) = \sum_{i=1}^{n} \frac{\partial (y_i - \hat{y}_i)^2}{\partial a_j} = -2\sum_{i=1}^{n} (y_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial a_j}.$$

$$= 2\sum_{i=1}^{n} (y_i - \hat{y}_i) b_j x_i e^{-a_j x_i}$$

$$\therefore \nabla_a J(a, b) = \left[\frac{\partial J(a, b)}{\partial a_1}, \cdots, \frac{\partial J(a, b)}{\partial a_d}\right]^T.$$