Problem 1:

$$1-\delta(a)$$
$$=1-\frac{1}{1+e^{-a}}=\frac{1+e^{-a}-1}{1+e^{-a}}$$
$$=\frac{e^{-a}}{1+e^{-a}}=\frac{1}{e^{a}+1}=\delta(-a)$$

Let $y=\delta(a)=\frac{1}{1+e^{-a}}$

$$\to 1+e^{-a}=\frac{1}{y}$$
$$\to e^{-a}=\frac{1}{y}-1$$
$$\to e^{-a}=\frac{1-y}{y}$$
$$\to -a=\ln\frac{1-y}{y}$$
$$\to a=\ln\frac{y}{1-y}=\delta^{-1}(y)$$

Problem 2:

In the following deduction, for convenience I assume q equals 2, since $\| w \|_q^q$ will appear in both of the two deduction, it doesn't matter what does q equals to.

The minimization of the regularized error function can be deduced as follows:

$$\frac{\partial[(w^T X-y)(w^T X-y)^T+\lambda w w^T]}{\partial w}$$
$$=\frac{\partial[w^T X X^T w-w^T X y^T-y X^T w+y y^T+\lambda w w^T]}{\partial w}$$
$$=2w^T X X^T-y X^T-y X^T+2\lambda w^T=0$$

So

$$w^T X X^T+\lambda w^T=y X^T$$
$$\to w^T(X X^T+\lambda I_d)=y X^T$$
$$\to (X X^T+\lambda I_d)w=X y^T$$

The minimization of the unregularized sum-of-squares error can be deduced as follows:

$$\frac{\partial[(w^T X - y)(w^T X - y)^T + \alpha(ww^T - \eta)]}{\partial w}$$
$$= \frac{\partial[w^T XX^T w - w^T Xy^T - yX^T w + yy^T + \alpha ww^T - \alpha\eta]}{\partial w}$$

Because $\alpha\eta$ is not relevant to w, so the following deduction is the same as above. So they are equivalent.

As for the relationship, suppose we choose a specific value of λ (where $\lambda > 0$) to optimize unregularized error function $w = \arg\min_{w \in R^d}(\| w^T X - y \|_2^2 + \lambda(\| w \|_q^q - \eta))$. According to the method of Lagrange multipliers that $\nabla_\lambda L(w,\lambda) = 0$, we have $\| w \|_q^q - \eta = 0$, therefore, parameter $\eta = \| w^*(\lambda) \|_q^q$, where $w^*(\lambda)$ is notation for the result value of w.

Problem 3:

First to separate the X^T into blocks according to its rows, namely, we represent X^T as:

$$X^T = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_n \end{pmatrix}$$ where $\alpha_1, \alpha_2, \ldots \alpha_n$ are row vectors of X^T.

Likely, we represent XX^T as:

$$XX^T = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{pmatrix}$$ where $\beta_1, \beta_2, \ldots \beta_n$ are the row vectors of XX^T

Also, we use $a_{i1}, a_{i2} \ldots a_{in} (i = 1,2,\ldots,d)$ to represent the elements in the row i of matrix X.

Since $\beta_i = a_{i1}\alpha_1 + a_{i2}\alpha_2 + \ldots + a_{in}\alpha_n (i = 1,2,\ldots,d)$

So the row vectors of XX^T can be represented as the linear combination of the row vectors of X^T.

$Rank(XX^T) < Rank(X^T)$ **(1)**

Also because X^T is a n by d matrix, so $Rank(X^T) \le \min(n,d) = n$ **(2)**

By inequality **(1)** and **(2)**, we can deduce that

$Rank(XX^T) < Rank(X^T) \le n < d$

Finally, since XX^T is a d by d matrix, and $Rank(XX^T) < d$, we can conclude that XX^T is not invertible.

Problem 4

(1)
Let
$f(w) = \| w^T X - y \|_2^2$
$= (w^T X - y)(w^T X - y)^T + \lambda w w^T$
Then let
$$\frac{\partial f(w)}{\partial w} = \frac{\partial[(w^T X - y)(w^T X - y)^T + \lambda w w^T]}{\partial w}$$
$$= \frac{\partial[w^T XX^T w - w^T X y^T - yX^T w + yy^T + \lambda w w^T]}{\partial w}$$
$$= 2w^T XX^T - yX^T - yX^T + 2\lambda w^T = 0$$

We can get
$w^T XX^T + \lambda w^T = yX^T$
$\to w^T (XX^T + \lambda I_d) = yX^T$
$\to (XX^T + \lambda I_d) w = Xy^T$

(2) Since XX^T is positive semi-define, so all its eigenvalues are nonnegative, so we have all the eigenvalues of $XX^T + \lambda I_d$ are no less than λ, so $XX^T + \lambda I_d$ is invertible.

Problem 5

For GD:

BEGIN:
    Initialize $\varepsilon, \lambda, i, w$
    Repeat until w reach convergence
        $w_i = w_{i-1} - \varepsilon * (XX^T + \lambda I_d)^{-1} Xy^T$
    Output w_i
END

For SGD:

BEGIN:
    LOOP 1: Repeat until w reach convergence
        LOOP2: Repeat until w reach convergence
            (X,y) = Random(S)

$w_i = w_{i-1} - \varepsilon * (XX^T + \lambda I_d)^{-1} Xy^T$

$i \leftarrow i+1$

END LOOP 2

$\varepsilon = 0.5 * \varepsilon$

END LOOP 1

Output w_i

END

Problem 6

Assuming $w \in R^d$ denotes the vector coefficient we need to solve,
and $Y = (\gamma_1, \gamma_2, \ldots, \gamma_n) \in R^{1\times n}$.
So

$$P(w \mid Y)$$

$$= \frac{P(w,Y)}{P(Y)} \sim P(w)P(Y \mid w)$$

Since we assume that $\gamma_i = w^T x_i + \varepsilon_i \quad i = 1,2,\ldots,n$ and $\varepsilon^i \sim N(0,\sigma^2) \quad i = 1,2,\ldots,n$

So

$$P(Y \mid w)$$

$$= \prod_{i=1}^{n} [\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(\gamma_i - w^T x_i)^2}{2\sigma^2})]$$

$$= (\frac{1}{\sqrt{2\pi}\sigma})^n \exp[-\frac{\sum_{i=1}^{n} (\gamma_i - w^T x_i)^2}{2\sigma^2}]$$

In order to get Lasso, notice that the penalty in Lasso is a norm-1 of the coefficient w, so we can intuitively contrive the prior distribution of w as follows:

$$f(w_i) = \frac{1}{2\varsigma} \exp(-\lambda \mid w_i \mid) \quad i = 1,2,\ldots,d$$

So

$$P(w) = \prod_{i=0}^{d} \frac{1}{2\varsigma} \exp(-\lambda \mid w_i \mid)$$

$$= (\frac{1}{2\varsigma})^d \exp(-\lambda \sum_{i=0}^{d} \mid w_i \mid)$$

So

$P(w)P(Y|w)$

$$= (\frac{1}{2\varsigma})^d \exp(-\lambda\sum_{i=0}^{d}|w_i|)(\frac{1}{\sqrt{2\pi\sigma}})^n \exp[-\frac{\sum_{i=1}^{n}(\gamma_i - w^T x_i)^2}{2\sigma^2}]$$

$$= (\frac{1}{2\varsigma})^d (\frac{1}{\sqrt{2\pi\sigma}})^n \exp(-\lambda\sum_{i=0}^{d}|w_i| - \frac{\sum_{i=1}^{n}(\gamma_i - w^T x_i)^2}{2\sigma^2})$$

Since we want to maximize $P(w)P(Y|w)$, which is equivalent to minimize

$$\lambda\sum_{i=0}^{d}|w_i| + \frac{\sum_{i=1}^{n}(\gamma_i - w^T x_i)^2}{2\sigma^2}$$

Namely, to find $w = \arg\min_{w\in R^d} \| w^T X - \gamma \|_2^2 + \lambda \| w \|_1$.

So the proper prior distribution for w is

$$P(w) = \prod_{i=0}^{d}\frac{1}{2\varsigma}\exp(-\lambda|w_i|)$$

$$= (\frac{1}{2\varsigma})^d \exp(-\lambda\sum_{i=0}^{d}|w_i|)$$

Problem 7

1. $f = \Theta(g)$

2. $f = \Omega(g)$

3. $f = \Theta(g)$

4. $f = \Omega(g)$

5. $f = \Omega(g)$

6. $f = \Omega(g)$

7. $f = \Omega(g)$

8. $f = \Theta(g)$

9. $f = \Omega(g)$

Problem 8

C1: the first coin
C2: the second coin
P(C1)=P(C2)=0.5

$$P(C_1 \mid HHT)$$
$$= \frac{P(HHT \mid C_1)P(C_1)}{P(HHT)}$$
$$= \frac{P(HHT \mid C_1)P(C_1)}{P(HHT \mid C_1)P(C_1) + P(HHT \mid C_2)P(C_2)}$$
$$\frac{(\frac{1}{2})^4}{\frac{1}{2}(\frac{1}{2})^3 + \frac{1}{2}(\frac{2}{3}(\frac{1}{3})^2)} = \frac{27}{43}$$