COMS 4771 Machine Learning (Spring 2015)
Problem Set #4

Solutions - `coms4771@machinelearning.nyc`
Discussants: `None`

April 14, 2015

Problem 1

(a) Minimize $\mathbb{E}[(Y - \boldsymbol{w}^\top \boldsymbol{X})^2]$ as a function of $\boldsymbol{w}$. The solution is $\boldsymbol{w} = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\mathbb{E}(Y\boldsymbol{X})$, so $\hat{Y} = \mathbb{E}(Y\boldsymbol{X})^\top \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}$.

(b)

$$\begin{aligned}
\mathbb{E}((Y - \hat{Y})\boldsymbol{X}) &= \mathbb{E}((Y - \mathbb{E}(Y\boldsymbol{X})^\top \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X})\boldsymbol{X}) \\
&= \mathbb{E}(Y\boldsymbol{X}) - \mathbb{E}(Y\boldsymbol{X})\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\mathbb{E}(\boldsymbol{X}\boldsymbol{X}) \\
&= \mathbb{E}(Y\boldsymbol{X}) - \mathbb{E}(Y\boldsymbol{X}) \\
&= \boldsymbol{0}.
\end{aligned}$$

(c) $\mathbb{E}(Z_i \boldsymbol{X}_{(-i)}) = \boldsymbol{0}$, for essentially the same reason as in (b).

(d) We first show $\mathbb{E}(Z_i^2) = \mathbb{E}(X_i Z_i)$:

$$\begin{aligned}
\mathbb{E}(Z_i^2) &= \mathbb{E}((X_i - \mathbb{E}(X_i \boldsymbol{X}_{(-i)})^\top \mathbb{E}(\boldsymbol{X}_{(-i)}\boldsymbol{X}_{(-i)}^\top)^{-1}\boldsymbol{X}_{(-i)})Z_i) \\
&= \mathbb{E}(X_i Z_i) - \mathbb{E}(X_i \boldsymbol{X}_{(-i)})^\top \mathbb{E}(\boldsymbol{X}_{(-i)}\boldsymbol{X}_{(-i)}^\top)^{-1}\mathbb{E}(\boldsymbol{X}_{(-i)} Z_i) \\
&= \mathbb{E}(X_i Z_i)
\end{aligned}$$

where the last step follows since $\mathbb{E}(\boldsymbol{X}_{(-i)} Z_i) = \boldsymbol{0}$ by the result in part (c).

Now we show that $\mathbb{E}(\langle \boldsymbol{w}, \boldsymbol{X} \rangle Z_i) = w_i \mathbb{E}(X_i Z_i)$:

$$\begin{aligned}
\mathbb{E}(\langle \boldsymbol{w}, \boldsymbol{X} \rangle Z_i) &= \mathbb{E}\left(w_i X_i Z_i + \sum_{j \neq i} w_j X_j Z_i \right) \\
&= w_i \mathbb{E}(X_i Z_i) + \sum_{j \neq i} w_j \mathbb{E}(X_j Z_i) \\
&= w_i \mathbb{E}(X_i Z_i)
\end{aligned}$$

where the last step again uses the same result from (c).

Now we show that $\mathbb{E}((Y - \hat{Y})Z_i) = 0$: this is true because Z_i is a linear combination of the X_i, so each term $(Y - \hat{Y})Z_i$ is a linear combination of the $(Y - \hat{Y})X_i$, each of which has expectation zero by the result from part (b).

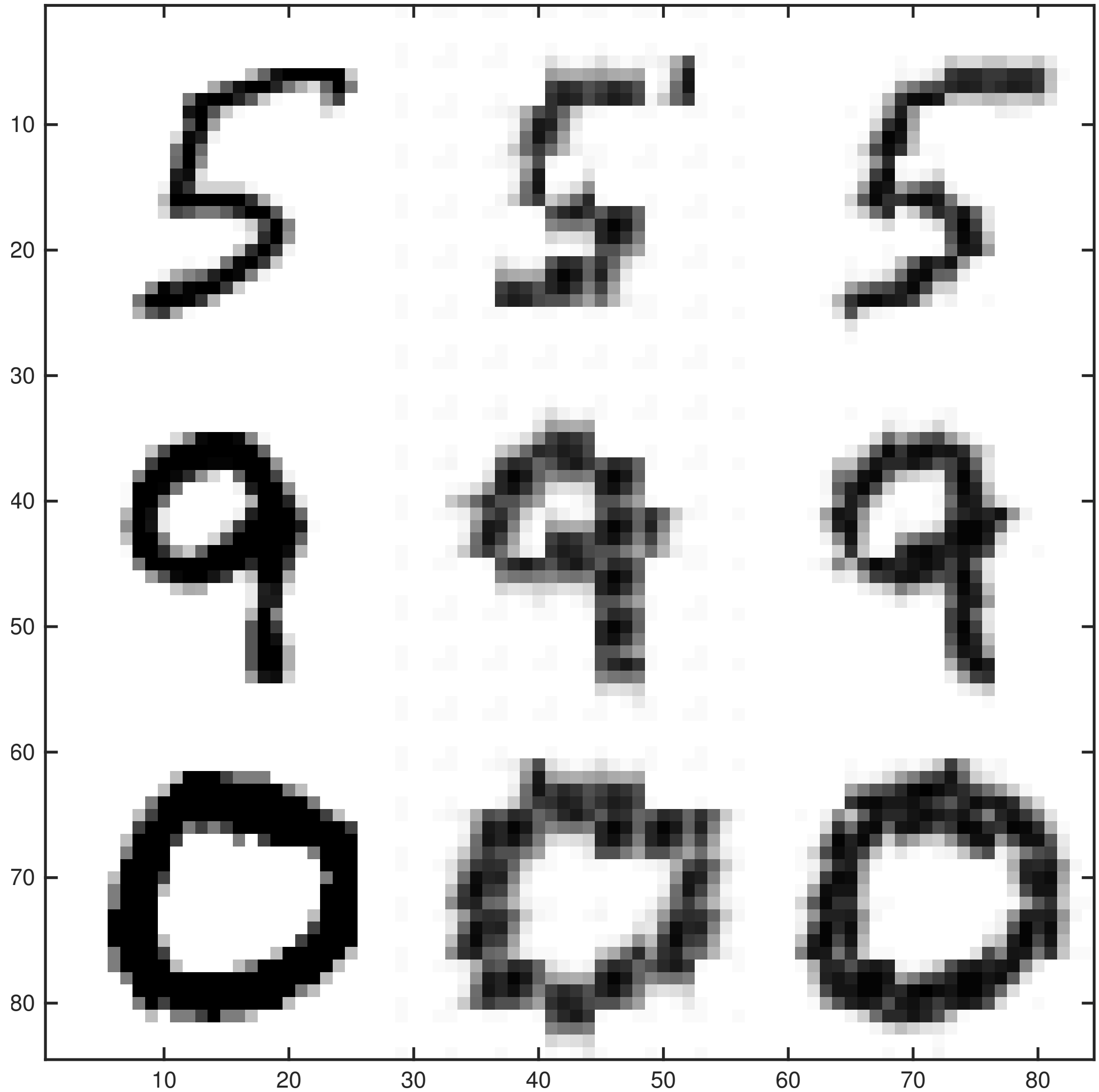
Since $YZ_i = \langle \boldsymbol{w}, \boldsymbol{X} \rangle Z_i + (Y - \hat{Y})Z_i$, we have

$$\mathbb{E}(YZ_i) = \mathbb{E}(\langle \boldsymbol{w}, \boldsymbol{X} \rangle Z_i) + \mathbb{E}((Y - \hat{Y})Z_i) = w_i \mathbb{E}(X_i Z_i) + 0,$$

which is the same as $\mathbb{E}(Z_i^2)w_i$.

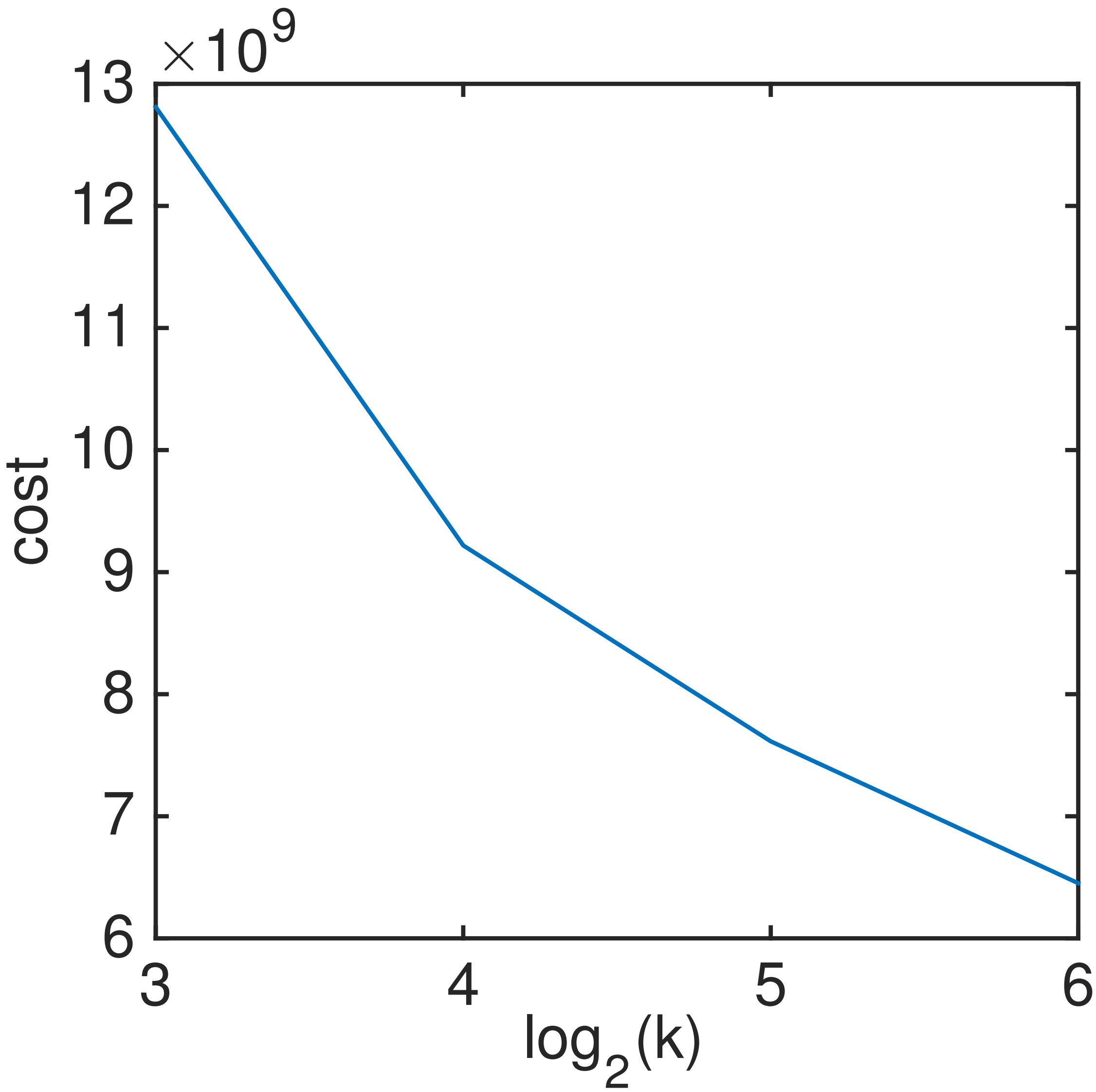
Problem 2

(a)



(b) The number of bits needed to encode all k representatives is $16 \times 64 \times k$, and the number of bits needed to encode all 10000 quantized test images is $10000 \times 49 \times \log_2(k)$. So the total number of bits is $1024k + 490000 \log_2(k)$.

Comment: If you wanted to be a little more clever, you could use a different encoding for $\{1, 2, \ldots, k\}$ based on the actual statistics of how these patches are used in the quantized test images. Then instead of $\log_2(k)$ bits per patch, the average number of bits per patch would be $\lceil H(X) \rceil$, where X is a random variable taking values in $\{1, 2, \ldots, k\}$ and whose probability of being i is equal to the empirical frequency of being used in all 49000 image patch locations.



(c)

Problem 3

This is just a little bit of calculus.

(a)
$$\mathbb{E}[\ell_{\text{sq}}(Y\hat{y})] = \eta(1-\hat{y})^2 + (1-\eta)(1+\hat{y})^2.$$

This is a convex function of $\hat{y}$, so we can take derivatives and set it equal to zero to solve for a minimizer. Indeed, this leads to the equation $2\eta(\hat{y}-1)+2(1-\eta)(\hat{y}+1) = 0$, which rearranges to $\hat{y} = 2\eta - 1$.

(b)
$$\mathbb{E}[\ell_{\log}(Y\hat{y})] = \eta\ln(1+\exp(-\hat{y})) + (1-\eta)\ln(1+\exp(\hat{y})).$$

This is a convex function of $\hat{y}$, so we can take derivatives and set it equal to zero to solve for a minimizer. Indeed, this leads to the equation $-\eta\frac{e^{-\hat{y}}}{1+e^{-\hat{y}}} + (1-\eta)\frac{e^{\hat{y}}}{1+e^{\hat{y}}} = 0$; letting $\hat{p} := \frac{e^{\hat{y}}}{1+e^{\hat{y}}}$, this can be re-written as $-\eta(1-\hat{p}) + (1-\eta)\hat{p} = 0$, which becomes $\hat{p} = \eta$. Thus $\frac{e^{\hat{y}}}{1+e^{\hat{y}}} = \eta$, which simplifes to $\hat{y} = \ln\frac{\eta}{1-\eta}$.

(c)
$$\mathbb{E}[\ell_{\exp}(Y\hat{y})] = \eta e^{-\hat{y}} + (1-\eta)e^{\hat{y}}.$$

This is a convex function of $\hat{y}$, so we can take derivatives and set it equal to zero to solve for a minimizer. Indeed, this leads to the equation $-\eta e^{-\hat{y}} + (1-\eta)e^{\hat{y}} = 0$, which rearranges to $\hat{y} = \frac{1}{2}\ln\frac{\eta}{1-\eta}$.