HW4

1. The pdf of a $\Gamma(2,1)$ random variable is $p(z) = z\exp(-z),\ z > 0$, and the pmf of a Poisson random variable X is $p_X(x) = \lambda^x e^{-\lambda}/x!,\ \lambda > 0,\ x = 0, 1, \ldots$. Assuming that $X_1, X_2, \ldots, X_n$ is an i.i.d Poisson sample *given* that λ has a $\Gamma(2,1)$ prior distribution, find the MAP estimate of λ and prove that what you find is actually a value that maximizes the posterior. (10 pts)

$\because$ iid

$\therefore p(X_1, X_2, \cdots, X_n \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!}$

$\therefore P(\lambda \mid X_1, X_2, \cdots, X_n) \propto p(X_1, X_2, \cdots, X_n \mid \lambda) P(\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \cdot \lambda e^{-\lambda}$

$\therefore P(\lambda \mid X_1, X_2, \cdots, X_n) \propto \lambda^{1+\sum_{i=1}^n x_i} e^{-(n+1)\lambda} = \lambda^{(2+\sum_{i=1}^n x_i)-1} e^{-(n+1)\lambda}$

$\therefore$ posterior is a $\Gamma(2+\sum_{i=1}^n x_i,\ n+1)$ distribution

$\therefore \hat{\lambda}^{MAP} = \underset{\lambda}{\text{argmax}}\ P(\lambda \mid X_1, X_2, \cdots, X_n) =$ mode of $\Gamma(2+\sum_{i=1}^n x_i,\ n+1)$

$\therefore \hat{\lambda}^{MAP} = \frac{\alpha - 1}{\beta} = \frac{1+\sum_{i=1}^n x_i}{n+1}$

proof: let $g(\lambda) = \lambda^{(2+\sum_{i=1}^n x_i)-1} e^{-(n+1)\lambda}$

$\therefore \ln g(\lambda) = (1+\sum_{i=1}^n x_i)\ln\lambda - (n+1)\lambda$

$\therefore$ let $\frac{d}{d\lambda} g(\lambda) = \frac{1+\sum_{i=1}^n x_i}{\lambda} - (n+1) = 0$

$\therefore \lambda = \frac{1+\sum_{i=1}^n x_i}{n+1} = \hat{\lambda}^{MAP}$

$\because \frac{d^2}{d\lambda^2} g(\lambda) = \frac{-(1+\sum_{i=1}^n x_i)}{\lambda^2} < 0,\ \forall \lambda > 0$

$\therefore \hat{\lambda}^{MAP}$ maximizes the posterior

2. Assume that you have an i.i.d sample from a population with Poisson pmf, i.e. $p_X(x) = \lambda^x e^{-\lambda}/x!,\ \lambda > 0,\ x = 0, 1, \ldots$. Calculate the MLE of λ and its asymptotic distribution by calculating Fisher information and compare the results with those of the Central Limit Theorem. (10 pts)

$\because L(\lambda) = L(\lambda|D) = \prod p(x_i|\lambda) = \prod \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod x_i!}$

$\therefore \ell(\lambda) = (\sum x_i) \ln \lambda - n\lambda - \sum \ln x_i!$

$\therefore$ let $\frac{d}{d\lambda} \ell(\lambda) = \frac{\sum x_i}{\lambda} - n = 0$

$\therefore \hat{\lambda}^{MLE} = \frac{1}{n} \sum x_i = \bar{x}$

$\because \frac{d^2}{d\lambda^2} \ell(\lambda) = -\frac{\sum x_i}{\lambda^2}$ and $E[x_i] = \lambda$

$\therefore I(\lambda) = -E\left[\frac{d^2}{d\lambda^2} \ell(\lambda)\right] = -E\left[-\frac{\sum x_i}{\lambda^2}\right] = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$

$\therefore \hat{\lambda}^{MLE} \sim N(\lambda, \frac{1}{I(\lambda)}) = N(\lambda, \frac{\lambda}{n})$

$\because$ in CLT $\bar{x} \sim N(E[x], \frac{V[x]}{n}) = N(\lambda, \frac{\lambda}{n})$

$\therefore$ same

3. Assume that $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Show that the MLE and least squares estimates of the β vector are the same, which means MLE is also BLUE according to Gauss-Markov. Remember that the log-likelihood function $l(\beta_0, \beta_1, \ldots, \beta_p)$ here is based on the conditional density $p(Y|X_1, \ldots, X_p)$. (10 pts)

$\because \varepsilon \sim N(0, \sigma^2)$

$\therefore Y = X\beta + \varepsilon \sim N(X\beta, \sigma^2 I)$

$$\therefore L(\beta) = p(Y|X,\beta) = \frac{1}{(2\pi)^{\frac{n}{2}} |\det \sigma^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(Y-X\beta)^T (\sigma^2 I)^{-1} (Y-X\beta)\right)$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\|Y-X\beta\|^2\right)$$

$$\therefore \ell(\beta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|Y-X\beta\|^2$$

$$\therefore \hat{\beta}_{MLE} = \underset{\beta}{\operatorname{argmax}}\, \ell(\beta) = \underset{\beta}{\operatorname{argmin}} \|Y-X\beta\|^2 = \hat{\beta}_{LS}$$

$$\therefore \hat{\beta}_{MLE} = \hat{\beta}_{LS} = (X^T X)^{-1} X^T Y$$

$\therefore \hat{\beta}_{MLE}$ is BLUE

4. Find the MAP estimate of β under the assumption that $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and that the prior distribution of (independent) β_i, $i = 1, 2, \ldots, p$ is $\mathcal{N}(0, \sigma^2/\lambda)$. Interpret your results. (15 pts)

according to Q3 $p(Y|X,\beta) \propto \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right)$

$\because \beta_i \sim N(0, \sigma^2/\lambda)$, $i = 1, 2, \cdots, p$

$\therefore p(\beta) \propto \exp\left(-\frac{1}{2\sigma^2}\beta^T \Lambda \beta\right)$, $\Lambda = \text{diag}(0, \lambda, \cdots, \lambda)$

$\therefore p(\beta|X,Y) \propto p(Y|X,\beta)\, p(\beta) \propto \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta\|^2 - \frac{1}{2\sigma^2}\beta^T \Lambda \beta\right)$

$\therefore \ln p(\beta|X,Y) \propto -(\|Y - X\beta\|^2 + \beta^T \Lambda \beta)$ similar to ridge regression

$\therefore \hat{\beta}_{MAP} = (X^T X + \Lambda)^{-1} X^T Y$

$\therefore \hat{\beta}_{MAP}$ corresponds to ridge regression.
use Gaussian prior to achieve l_2-penalty
but only regularize non-intercept coefficients

5. Find the MAP estimate of β under the assumption that $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and that the prior distribution of (independent) β_i, $i = 1, 2, \ldots, p$ is $\text{Lap}(0, \sigma^2/\lambda)$. Interpret your results. (15 pts)

Similarly, $p(Y|X,\beta) \propto \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right)$

$\because \beta_i \sim \text{Lap}(0, \sigma^2/\lambda)$, $i = 1, 2, \cdots, p$

$\therefore p(\beta_i) = \frac{1}{2\sigma^2/\lambda} \exp\left(-\frac{|\beta_i|}{\sigma^2/\lambda}\right) = \frac{\lambda}{2\sigma^2} \exp\left(-\frac{\lambda}{\sigma^2}|\beta_i|\right)$, $i = 1, 2, \cdots, p$

$\therefore p(\beta) \propto \exp\left(-\frac{\lambda}{\sigma^2}\sum_{i=1}^{p}|\beta_i|\right)$

$\therefore p(\beta|X,Y) \propto p(Y|X,\beta)\, p(\beta) \propto \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta\|^2 - \frac{\lambda}{\sigma^2}\sum_{i=1}^{p}|\beta_i|\right)$

$\therefore \ln p(\beta|X,Y) \propto -\frac{1}{2}\|Y - X\beta\|^2 - \lambda\sum_{i=1}^{p}|\beta_i|$

$\therefore \hat{\beta}_{MAP} = \underset{\beta}{\text{argmax}} \ln p(\beta|X,Y) = \underset{\beta}{\text{argmin}}\left(\|Y - X\beta\|^2 + 2\lambda\sum_{i=1}^{p}|\beta_i|\right)$

$\therefore \hat{\beta}_{MAP}$ corresponds to Lasso regression.

use Laplace prior to achieve ℓ_1-penalty

6. In the regularized least squares problem, assume that the singular value decomposition of $\mathbf{X}$ is $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

   (a) Show that the vector of predicted values is: (10 pts)

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}^{\text{Ridge}} = \sum_{j=1}^{p} \mathbf{u}_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \mathbf{u}_j^T \mathbf{y}$$

   where $\mathbf{u}_j$ are the columns of $\mathbf{U}$. Conclude that greater amount of shrinkage is applied to basis vectors $\mathbf{u}_j$ that have smaller singular values σ_j, for a fixed $\lambda \geq 0$.

(a) $\because \hat{\beta}^{\text{Ridge}} = (X^TX + \lambda I)^{-1} X^T y$

$= ((U\Sigma V^T)^T (U\Sigma V^T) + \lambda I)^{-1} (U\Sigma V^T)^T y$

$= (V\Sigma U^T U \Sigma V^T + \lambda I)^{-1} V\Sigma U^T y$ $\quad (\because U^{-1} = U^T,\ V^{-1} = V^T)$

$= (V\Sigma^2 V^T + \lambda I)^{-1} V\Sigma U^T y$

$= (V^T)^{-1} (\Sigma^2 + \lambda I)^{-1} V^{-1} V \Sigma U^T y$

$= V(\Sigma^2 + \lambda I)^{-1} \Sigma U^T y$

$\therefore \hat{y} = X\hat{\beta}^{\text{Ridge}} = (U\Sigma V^T) V (\Sigma^2 + \lambda I)^{-1} \Sigma U^T y$

$= U\Sigma (\Sigma^2 + \lambda I)^{-1} \Sigma U^T y$

$= U \frac{\Sigma^2}{\Sigma^2 + \lambda I} U^T y$

$\because \frac{\Sigma^2}{\Sigma^2 + \lambda I} = \text{diag}\left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right)$

$\therefore \hat{y} = \sum_{j=1}^{p} u_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j^T y$

(b) Use SVD to show that (10 pts)

$$\mathrm{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T] = \sum_{j=1}^{p} \frac{\sigma_j^2}{\sigma_j^2+\lambda}$$

This quantity is equal to the degrees of freedom p when $\lambda = 0$ and is called the effective degrees of freedom for the Ridge-regularized model.

(b) $\because X(X^TX+\lambda I)^{-1}X^T = (U\Sigma V^T)(V\Sigma^2V^T+\lambda I)^{-1}(U\Sigma V^T)^T$

$= U\Sigma V^T V(\Sigma^2+\lambda I)^{-1}V^{-1}V\Sigma U^T$

$= U\Sigma(\Sigma^2+\lambda I)^{-1}\Sigma U^T$

$= U\frac{\Sigma^2}{\Sigma^2+\lambda I}U^T$

$\therefore \mathrm{tr}[X(X^TX+\lambda I)^{-1}X^T] = \mathrm{tr}\left[U\frac{\Sigma^2}{\Sigma^2+\lambda I}U^T\right]$

$= \mathrm{tr}\left[\frac{\Sigma^2}{\Sigma^2+\lambda I}U^TU\right]$ $(\because \mathrm{tr}(ABC) = \mathrm{tr}(BCA))$

$= \mathrm{tr}\left[\frac{\Sigma^2}{\Sigma^2+\lambda I}\right]$

$\because \frac{\Sigma^2}{\Sigma^2+\lambda I} = \mathrm{diag}\left(\frac{\sigma_j^2}{\sigma_j^2+\lambda}\right)$

$\therefore \mathrm{tr}[X(X^TX+\lambda I)^{-1}X^T] = \sum_{j=1}^{p}\frac{\sigma_j^2}{\sigma_j^2+\lambda}$