Christopher Hase
christopher_hase@g.harvard.edu
CS281-F17

Assignment #1 v 1.2

Due: 5:00pm September 22, 2017

Collaborators: None

Problem 1 (A Classic on the Gaussian Algebra, 10pts)

Let X and Y be independent univariate Gaussian random variables. In the previous problem set, you likely used the closure property that $Z = X + Y$ is also a Gaussian random variable. Here you'll prove this fact.

(a) Suppose X and Y have mean 0 and variances σ_X^2 and σ_Y^2 respectively. Write the pdf of $X + Y$ as an integral.

(b) Evaluate the integral from the previous part to find a closed-form expression for the pdf of $X+Y$, then argue that this expression implies that $X + Y$ is also Gaussian with mean 0 and variance $\sigma_X^2 + \sigma_Y^2$. Hint: what is the integral, over the entire real line, of

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right),$$

i.e., the pdf of a univariate Gaussian random variable?

(c) Extend the above result to the case in which X and Y may have arbitrary means.

(d) Univariate Gaussians are supported on the entire real line. Sometimes this is undesirable because we are modeling a quantity with positive support. A common way to transform a Gaussian to solve this problem is to exponentiate it. Suppose X is a univariate Gaussian with mean μ and variance σ^2. What is the pdf of e^X?

(a) Let $X \sim \mathcal{N}(0, \sigma_X^2)$, $Y \sim \mathcal{N}(0, \sigma_Y^2)$, and $Z = X + Y$. $f_Z(z)$ can be found by convolving $f_X(x)$ and $f_Y(y)$.

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{1}{2\sigma_X^2}x^2\right) \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{1}{2\sigma_Y^2}(z-x)^2\right) dx$$

(b) Continuing from (a):

$$= \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\left(\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)}} \exp\left(-\frac{1}{2\sigma_X^2}x^2 - \frac{1}{2\sigma_Y^2}(z-x)^2\right) dx$$

We need to rearrange the expression in the exponent:

$$-\frac{1}{2\sigma_X^2}x^2 - \frac{1}{2\sigma_Y^2}(z-x)^2$$

$$= -\frac{1}{2}\left(\frac{x^2}{\sigma_X^2} + \frac{z^2}{\sigma_Y^2} - \frac{2xz}{\sigma_Y^2} + \frac{x^2}{\sigma_Y^2}\right)$$

$$= -\frac{1}{2}\left(\frac{x^2(\sigma_X^2+\sigma_Y^2) + z^2\sigma_X^2 - 2xz\sigma_X^2}{\sigma_X^2\sigma_Y^2}\right)$$

$$= -\frac{1}{2}\left(\frac{x^2(\sigma_X^2+\sigma_Y^2)}{\sigma_X^2\sigma_Y^2} + \frac{z^2 - 2xz(\sigma_X^2+\sigma_Y^2)\sigma_X^2}{\sigma_X^2\sigma_Y^2(\sigma_X^2+\sigma_Y^2)}\right)$$

$$= -\frac{1}{2}\left(\frac{z^2}{\sigma_X^2+\sigma_Y^2} + \frac{z^2\sigma_X^4 - 2xz(\sigma_X^2+\sigma_Y^2)\sigma_X^2}{\sigma_X^2\sigma_Y^2(\sigma_X^2+\sigma_Y^2)} + \frac{x^2(\sigma_X^2+\sigma_Y^2)}{\sigma_X^2\sigma_Y^2}\right)$$

$$= -\frac{1}{2}\left(\frac{z^2}{\sigma_X^2+\sigma_Y^2} + \frac{\left(\frac{z\sigma_X^2}{\sigma_X^2+\sigma_Y^2}\right)^2 - \frac{2xz\sigma_X^2}{\sigma_X^2+\sigma_Y^2} + x^2}{\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}}\right)$$

$$= -\frac{z^2}{2(\sigma_X^2+\sigma_Y^2)} - \frac{\left(x - \frac{z\sigma_X^2}{\sigma_X^2+\sigma_Y^2}\right)^2}{2\left(\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)}$$

Plugging this into the exponent of the integral, we have:

$$= \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\left(\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)}}\exp\left(-\frac{z^2}{2(\sigma_X^2+\sigma_Y^2)} - \frac{\left(x - \frac{z\sigma_X^2}{\sigma_X^2+\sigma_Y^2}\right)^2}{2\left(\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)}\right)dx$$

$$= \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}}\exp\left(-\frac{z^2}{2(\sigma_X^2+\sigma_Y^2)}\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\left(\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)}}\exp\left(-\frac{\left(x - \frac{z\sigma_X^2}{\sigma_X^2+\sigma_Y^2}\right)^2}{2\left(\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)}\right)dx$$

$$= \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}}\exp\left(-\frac{z^2}{2(\sigma_X^2+\sigma_Y^2)}\right)$$

$$\Rightarrow Z \sim \mathcal{N}(0, \sigma_X^2+\sigma_Y^2)$$

(c) Let $X \sim \mathcal{N}(0,\sigma_X^2)$, $Y \sim \mathcal{N}(0,\sigma_Y^2)$, $Z = X+Y$ so that $Z \sim \mathcal{N}(0,\sigma_X^2+\sigma_Y^2)$ (this is the result from (b)), $X' \sim \mathcal{N}(\mu_X,\sigma_X^2)$, and $Y' \sim \mathcal{N}(\mu_Y,\sigma_Y^2)$. We want to find the distribution of $Z' = X'+Y'$. First we'll

show that X' and $X + \mu_X$ have the same distribution.

$$f_{X'}(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{1}{2\sigma_X^2}(x-\mu_X)^2 \right)$$

$$F_{X+\mu_X}(x) = P(X + \mu_X \leq x) = P(X \leq x - \mu_X) = F_X(x - \mu_X)$$

$$\Rightarrow f_{X+\mu_X}(x) = f_X(x-\mu_X) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{1}{2\sigma_X^2}(x-\mu_X)^2 \right)$$

Thus, we can set $X' = X+\mu_X$. By the same argument, we can set $Y' = Y+\mu_Y$. Then $Z' = X+Y+\mu_X+\mu_Y = Z + \mu_X + \mu_Y$.

$$F_{Z'}(z) = P(Z' \leq z) = P(Z + \mu_X + \mu_Y \leq z) = P(Z \leq z - \mu_X - \mu_Y) = F_Z(z - \mu_X - \mu_Y)$$

$$\Rightarrow f_{Z'}(z) = f_Z(z - \mu_x - \mu_Y) = \frac{1}{\sqrt{2\pi(\sigma_X^2 + \sigma_Y^2)}} \exp\left(-\frac{(z-\mu_X-\mu_Y)^2}{2(\sigma_X^2+\sigma_Y^2)} \right)$$

$$\Rightarrow Z' \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

(d) Let $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$F_{e^X}(x) = P(e^X \leq x) = P\big(X \leq \ln(x)\big) = F_X\big(\ln(x)\big)$$

$$\Rightarrow f_{e^X}(x) = \frac{1}{x} f_X\big(\ln(x)\big) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\big(\ln(x)-\mu\big)^2}{2\sigma^2} \right)$$

$$\Rightarrow e^X \sim \text{Lognormal}(\mu, \sigma^2)$$

Problem 2 (Regression, 13pts)

Suppose that $X \in \mathbb{R}^{n \times m}$ with $n \geq m$ and $Y \in \mathbb{R}^n$, and that $Y \sim \mathcal{N}(Xw, \sigma^2 I)$. You learned in class that the maximum likelihood estimate $\hat{w}$ of w is given by

$$\hat{w} = (X^T X)^{-1} X^T Y$$

(a) Why do we need to assume that $n \geq m$?

(b) Define $H = X(X^T X)^{-1} X^T$, so that the "fitted" values $\hat{Y} = X\hat{w}$ satisfy $\hat{Y} = HY$. Show that H is an orthogonal projection matrix that projects onto the column space of X, so that the fitted y-values are a projection of Y onto the column space of X.

(c) What are the expectation and covariance matrix of $\hat{w}$?

(d) Compute the gradient with respect to w of the log likelihood implied by the model above, assuming we have observed Y and X.

(e) Suppose we place a normal prior on w. That is, we assume that $w \sim \mathcal{N}(0, \tau^2 I)$. Show that the MAP estimate of w given Y in this context is

$$\hat{w}_{MAP} = (X^T X + \lambda I)^{-1} X^T Y$$

where $\lambda = \sigma^2/\tau^2$. (You may employ standard conjugacy results about Gaussians without proof in your solution.)

[Estimating w in this way is called *ridge regression* because the matrix λI looks like a "ridge". Ridge regression is a common form of *regularization* that is used to avoid the overfitting (resp. underdetermination) that happens when the sample size is close to (resp. higher than) the output dimension in linear regression.]

(f) Do we need $n \geq m$ to do ridge regression? Why or why not?

(g) Show that ridge regression is equivalent to adding m additional rows to X where the j-th additional row has its j-th entry equal to $\sqrt{\lambda}$ and all other entries equal to zero, adding m corresponding additional entries to Y that are all 0, and then computing the maximum likelihood estimate of w using the modified X and Y.

(a) We showed in problem 4a of homework 0 that $X^T X$ and XX^T have the same non-zero eigenvalues and that if $m > n$, $X^T X$ will have at least one zero-valued eigenvalue. Thus, if $m > n$, $X^T X$ is not invertible. Then $\hat{w}$ does not have a unique maximum likelihood solution (if $X^T X$ not invertible $\hat{w}$ cannot have the unique maximum likelihood solution $(X^T X)^{-1} X^T Y$).

(b) By construction, $\forall\ Y \in \mathbb{R}^n\ \exists\ \hat{w} \in \mathbb{R}^m$ such that $X\hat{w} = HY \Rightarrow \text{Im}(H) \subseteq \text{Im}(X)$. Also, $\hat{Y} \in \text{Im}(X)$ and $\hat{Y} = HY$. Thus, if H is an orthogonal projection matrix, it projects onto the column space of X with $\hat{Y}$ being the projection of Y by H onto the column space of X. So all we need to show is that H is an orthogonal projection matrix. To do this, we need to show (**1**) that multiplying a vector that is already in the column space of H by H yields the same vector and (**2**) that multiplying a vector that is perpendicular to the column space of H by H yields a 0 vector.

(**1**) $\hat{Y}$ is in the column space of H.

$\hat{Y} = HY \Rightarrow H\hat{Y} = HHY$

$\Rightarrow H\hat{Y} = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^TY$

$\Rightarrow H\hat{Y} = X(X^TX)^{-1}X^TY$

$\Rightarrow H\hat{Y} = HY$

$\Rightarrow H\hat{Y} = \hat{Y}$

(**2**) By construction, $Y - \hat{Y}$ is perpendicular to the column space of H.

$H(Y - \hat{Y}) = HY - H\hat{Y} = HY - \hat{Y} = 0.$

(c) $E(\hat{w}) = E\big((X^TX)^{-1}X^TY\big)$

$= (X^TX)^{-1}X^TE(Y)$

$= (X^TX)^{-1}(X^TX)w$

$= w$

$\mathrm{Var}(\hat{w}) = E\Big(\big((X^TX)^{-1}X^TY\big)\big((X^TX)^{-1}X^TY\big)^T\Big) - E\Big((X^TX)^{-1}X^TY\Big)E\Big(\big((X^TX)^{-1}X^TY\big)^T\Big)$

$= (X^TX)^{-1}X^TE(YY^T)X(X^TX)^{-T} - (X^TX)^{-1}X^TE(Y)E(Y^T)X(X^TX)^{-T}$

$= (X^TX)^{-1}X^T\big(E(YY^T - E(Y)E(Y^T)\big)X(X^TX)^{-T}$

$= \sigma^2(X^TX)^{-1}(X^TX)(X^TX)^{-T}$

$= \sigma^2(X^TX)^{-1}$

(d) $p(Y) = \dfrac{1}{\sqrt{|2\pi\sigma^2 I|}}\exp\left(-\dfrac{1}{2\sigma^2}(Y - Xw)^T(Y - Xw)\right)$

$\Rightarrow \ln\big(p(y)\big) = -\ln\big(\sqrt{|2\pi\sigma^2 I|}\big) - \dfrac{1}{2\sigma^2}(Y - Xw)^T(Y - Xw)$

$= -\ln\big(\sqrt{|2\pi\sigma^2 I|}\big) - \dfrac{1}{2\sigma^2}(Y^TY - 2w^TX^TY + w^TX^TXw)$

$\Rightarrow \dfrac{d\Big(\ln\big(p(y)\big)\Big)}{dw} = -\dfrac{1}{2\sigma^2}(-2X^TY + 2X^TXw)$

$= \dfrac{X^TY - X^TXw}{\sigma^2}$

(e) $\underset{w}{\text{argmax}}\ p(w|Y) = \underset{w}{\text{argmax}}\ p(Y|w)p(w)$

$$= \underset{w}{\text{argmax}}\ \frac{1}{\sqrt{|2\pi\sigma^2 I|}} \exp\left(-\frac{1}{2\sigma^2}(Y - Xw)^T(Y - Xw) \right) \frac{1}{\sqrt{|2\pi\tau^2 I|}} \exp\left(-\frac{1}{2\tau^2} w^T w \right)$$

$$= \underset{w}{\text{argmax}} - \ln\left(\sqrt{|2\pi\sigma^2 I|}\right) - \ln\left(\sqrt{|2\pi\tau^2 I|}\right) + \frac{-Y^TY + 2w^TX^TY - w^TX^TXw}{2\sigma^2} - \frac{w^Tw}{2\tau^2}$$

$$= \underset{w}{\text{argmax}}\ \frac{2w^TX^TY - w^TX^TXw}{2\sigma^2} - \frac{w^Tw}{2\tau^2}$$

$$\frac{d\Big(\ln\big(p(Y|w)p(w)\big)\Big)}{dw} = \frac{X^TY - X^TX\hat{w}}{\sigma^2} - \frac{\hat{w}}{\tau^2} \equiv 0$$

$$\Rightarrow \left(\frac{X^TX}{\sigma^2} + \frac{1}{\tau^2}I\right)\hat{w} = \frac{X^TY}{\sigma^2}$$

$$\Rightarrow \left(X^TX + \frac{\sigma^2}{\tau^2}I\right)\hat{w} = X^TY$$

$$\Rightarrow \hat{w}_{MAP} = \left(X^TX + \frac{\sigma^2}{\tau^2}I\right)^{-1} X^TY$$

$$\Rightarrow \hat{w}_{MAP} = \Big(X^TX + \lambda I\Big)^{-1} X^TY \text{ where } \lambda = \frac{\sigma^2}{\tau^2}$$

(f) We do not need $n \geq m$ to do ridge regression. X^TX is positive semi-definite, so its eigenvalues are ≥ 0. Let γ be an eigenvalue of X^TX associated with vector a.

$X^TXa = \gamma a$

$\Rightarrow X^TXa + \lambda a = \gamma a + \lambda a$ where $\lambda = \dfrac{\sigma^2}{\tau^2}$ as in (e) $\Rightarrow \lambda > 0$

$\Rightarrow (X^TX + \lambda I)a = (\gamma + \lambda)a$ where $(\gamma + \lambda) > 0$ since $\gamma \geq 0$ and $\lambda > 0$

Then all of the eigenvalues of $(X^TX + \lambda I)$ are positive, so $(X^TX + \lambda I)^{-1}$ exists. Then under the conditions set in this problem excluding that $n \geq m$, $\hat{w}_{MAP}$ always has the unique solution $\Big(X^TX + \lambda I\Big)^{-1}X^TY$.

(g) Let X be the original matrix and $X_r \in \mathbb{R}^{(n+m)\times m}$ be the newly constructed matrix. Let Y be the original vector and $Y_r \in \mathbb{R}^{n+m}$ be the newly constructed vector. Using the formula in the problem, we know that the maximum likelihood solution using both the newly constructed matrix and vector is:

$\hat{w} = (X_r^TX_r)^{-1}X_r^TY_r$

$X_r^T X_r$ is an $m \times m$ matrix with $x_i^T x_i + \lambda$ as the ith diagonal entry $\forall\ i \in \{1, ..., m\}$ where x_i is the ith column of X. $X^T X$ is the same as $X_r^T X_r$ everywhere but the diagonal where $X^T X$ has $x_i^T x_i$ as its ith diagonal entry $\forall\ i \in \{1, ..., m\}$. Then $X_r^T X_r = X^T X + \lambda I$.

$X_r^T Y_r$ is an m-dimensional vector with $x_i^T Y$ as its ith entry $\forall\ i \in \{1, ..., m\}$ where x_i is the ith column of X. Then $X_r^T Y_r = X^T Y$.

Plugging in $X^T X + \lambda I$ for $X_r^T X_r$ and $X^T Y$ for $X_r^T Y_r$ in the maximum likelihood solution using both the newly constructed matrix and vector gives us:

$\hat{w} = \left(X^T X + \lambda I\right)^{-1} X^T Y$, which is the ridge regression solution for w.

Problem 3 (The Dirichlet and Multinomial Distributions, 12pts)

The Dirichlet distribution over K categories is a generalization of the beta distribution. It has a shape parameter $\alpha \in \mathbb{R}^K$ with non-negative entries and is supported over the set of K-dimensional positive vectors whose components sum to 1. Its density is given by

$$f(\theta_{1:K}|\alpha_{1:K}) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

(Notice that when $K = 2$, this reduces to the density of a beta distribution.) For the rest of this problem, assume a fixed $K \geq 2$.

(a) Suppose θ is Dirichlet-distributed with shape parameter α. Without proof, state the value of $E(\theta)$. Your answer should be a vector defined in terms of either α or K or potentially both.

(b) Suppose that $\theta \sim \text{Dir}(\alpha)$ and that $X \sim \text{Cat}(\theta)$, where *Cat* is a Categorical distribution. That is, suppose we first sample a K-dimensional vector θ with entries in $(0, 1)$ from a Dirichlet distribution and then roll a K-sided die such that the probability of rolling the number k is θ_k. Prove that the posterior $p(\theta|X)$ also follows a Dirichlet distribution. What is its shape parameter?

(c) Now suppose that $\theta \sim \text{Dir}(\alpha)$ and that $X^{(1)}, X^{(2)}, \ldots \overset{iid}{\sim} \text{Cat}(\theta)$. Show that the posterior predictive after $n-1$ observations is given by,

$$P(X^{(n)} = k|X^{(1)}, \ldots, X^{(n-1)}) = \frac{\alpha_k^{(n)}}{\sum_k \alpha_k^{(n)}}$$

where for all k, $\alpha_k^{(n)} = \alpha_k + \sum_{i=1}^{n-1} \mathbf{1}\{X^{(i)} = k\}$. (Bonus points if your solution does not involve any integrals.)

(d) Consider the random vector $Z_k = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X^{(i)} = k\}$ for all k. What is the mean of this vector? What is the distribution of the vector? (If you're not sure how to rigorously talk about convergence of random variables, give an informal argument. Hint: what would you say if θ were fixed?) What is the marginal distribution of a single class $p(Z_k)$?

(e) Suppose we have K distinct colors and an urn with α_k balls of color k. At each time step, we choose a ball uniformly at random from the urn and then add into the urn an additional new ball of the same color as the chosen ball. (So if at the first time step we choose a ball of color 1, we'll end up with $\alpha_1 + 1$ balls of color 1 and α_k balls of color k for all $k > 1$ at the start of the second time step.) Let $\rho_k^{(n)}$ be the fraction of all the balls that are of color k at time n. What is the distribution of $\lim_{n\to\infty} \rho_k^{(n)}$? Prove your answer.

(a) $E(\theta) = \dfrac{\alpha}{\sum_{i=1}^{K} \alpha_i}$

(b) Let $X \in \mathbb{R}^K$ be a one-hot encoded random vector such that $X_i = 1$ if the ith side of the die shows and $X_i = 0$ otherwise.

$p(\theta|X = x) \propto P(X = x|\theta)p(\theta)$

$$= \prod_{i=1}^{K} \theta_i^{x_i} \frac{\Gamma\Big(\sum_{j=1}^{K} \alpha_j\Big)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{l=1}^{K} \theta_l^{\alpha_l - 1}$$

$$\propto \prod_{i=1}^{K} \theta_l^{\alpha_i + x_i - 1}$$

$\Rightarrow \theta | X = x \sim \text{Dir}(\alpha + x)$ since the final expression has the form of the Dirichlet distribution (unnormalized).

(c) Let $X^{(i)}$ be a K-dimensional random vector distributed categorically with parameter θ that is one-hot encoded such that $X_j^{(i)} = 1$ if the jth side of the die shows and $X_j^{(i)} = 0$ otherwise $\forall\ i \in \{1, ..., n-1\}$. Let X be a $K \times (n-1)$ random matrix where $X^{(i)} \in \mathbb{R}^K$ is the ith column $\forall\ i \in \{1, ..., n-1\}$ and $X_j \in \mathbb{R}^{1 \times (n-1)}$ is the jth row $\forall\ j \in \{1, ..., K\}$.

$p(\theta | X = x) \propto P(X = x | \theta) p(\theta)$

$$= \prod_{i=1}^{n-1} \prod_{j=1}^{K} \theta_j^{x_j^{(i)}} \frac{\Gamma\Big(\sum_{l=1}^{K} \alpha_l\Big)}{\prod_{l=1}^{K} \Gamma(\alpha_l)} \prod_{m=1}^{K} \theta_m^{\alpha_m - 1}$$

$$\propto \prod_{j=1}^{K} \theta_j^{\sum_{i=1}^{n-1} x_j^{(i)} + \alpha_j - 1}$$

$\Rightarrow \theta | X = x \sim \text{Dir}(\alpha + x\mathbf{1})$ where $\mathbf{1}$ is an $(n-1)$-dimensional vector with each of its entries being 1.

This gives is a similar result to the one obtained in part (b) for the posterior. We will use this result to get the posterior predictive $P(X_j^{(n)} = 1 | X = x)$.

$$P(X_j^{(n)} = 1 | X = x) = E(X_j^{(n)} | X = x) = E\big(E(X_j^{(n)} | \theta, X = x) | X = x\big) = E\big(E(X_j^{(n)} | \theta) | X = x\big)$$

$$= E(\theta_j | X = x) = \frac{\alpha_j + x_j \mathbf{1}}{\sum_{i=1}^{K} \alpha_i + x_i \mathbf{1}} \text{ (this step comes from the result for part (a))}$$

$$= \frac{\alpha_j + x_j \mathbf{1}}{\sum_{i=1}^{K} \alpha_i + n - 1} = \frac{\alpha_j^{(n)}}{\sum_{i=1}^{K} \alpha_i^{(n)}}$$

(d) Let $Z = [Z_1, ..., Z_K]^T$.

$$E(Z_k) = E\left(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_k^{(i)} \right)$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E\Big(X_k^{(i)}\Big)$$

$= \lim_{n\to\infty} \frac{1}{n} n\theta_k = \theta_k \ \forall \ k \in \{1, ..., K\}$

Thus, $E(Z) = \theta$. By the law of large numbers, after infinitely many trials, the proportion of times that the kth side of the K-sided die shows up will converge to $\theta_k \ \forall \ k \in \{1, ..., K\}$. This implies that Z_k will have the same distribution as θ_k and Z will have the same distrubtion as $\left[\theta_1, ..., \theta_K\right]^T$. Since $\left[\theta_1, ..., \theta_K\right]^T \sim \text{Dir}(\alpha)$, $Z \sim \text{Dir}(\alpha)$.

The marginal distribution of a single class Z_k is $\text{Beta}\left(\alpha_k, \sum_{i=1}^{K} \alpha_i - \alpha_k\right)$. We can show this by decomposing the Dirichlet PDF $p(z)$. Note that we can write the density of Z in terms of $(K-1)$ of $Z_1, ..., Z_K$ since $\sum_{i=1}^{K} Z_i = 1$. Without loss of generality, suppose we choose to write the joint density of $Z_1, ..., Z_K$ in terms of $Z_1, ..., Z_{K-1}$ with $Z_K = 1 - \sum_{i=1}^{K-1} Z_i$ and find the marginal distribution of Z_1. Then we can write $p(z) = p(z_1)p(z_2|z_1)p(z_3|z_1, z_2)...p(z_{K-1}|z_1, ...z_{K-2})$ where $p(z_1)$ is a beta PDF parametrized by α_1 and $\sum_{i=1}^{K} \alpha_i - \alpha_1$.

For purposes of demonstration, let $K = 3$. We choose to write the joint density of Z_1, Z_2, Z_3 in terms of Z_1 and Z_2 with $Z_3 = 1 - Z_1 - Z_2$ and find the marginal distribution Z_1.

$$p(z) = \frac{\Gamma\left(\sum_{j=1}^{3} \alpha_j\right)}{\prod_{j=1}^{3} \Gamma(\alpha_j)} z_1^{\alpha_1 - 1} z_2^{\alpha_2 - 1} (1 - z_1 - z_2)^{\sum_{j=1}^{3} \alpha_j - \alpha_1 - \alpha_2 - 1}$$

$$= \left(\frac{\Gamma\left(\alpha_1 + \sum_{j=1}^{3} \alpha_j - \alpha_1\right)}{\Gamma(\alpha_1)\Gamma\left(\sum_{j=1}^{3} \alpha_j - \alpha_1\right)} z_1^{\alpha_1 - 1} (1 - z_1)^{\sum_{j=1}^{3} \alpha_j - \alpha_1 - 1}\right)\left(\frac{\Gamma\left(\sum_{j=1}^{3} \alpha_j - \alpha_1\right)}{\Gamma(\alpha_2)\Gamma\left(\sum_{j=1}^{3} \alpha_j - \alpha_1 - \alpha_2\right)} \frac{z_2^{\alpha_2 - 1}(1 - z_1 - z_2)^{\sum_{j=1}^{3} \alpha_j - \alpha_1 - \alpha_2 - 1}}{(1 - z_1)^{\sum_{j=1}^{3} \alpha_j - \alpha_1 - 1}}\right)$$

$$= p(z_1)p(z_2|z_1) \Rightarrow Z_1 \sim \text{Beta}\left(\alpha_1, \sum_{i=1}^{3} \alpha_i - \alpha_1\right)$$

(e) Although we update our knowledge about the balls in the urn after each time-step, at time 0 the only information we have is the original set of balls. When we take the perspective of being at time 0, we have no reason to think that the fraction of all the balls that are of color k at any time n will not be approximately the same as the fraction of all the balls that are of color k at time 0. Thus, we would think that $\lim_{n\to\infty} \rho_k^{(n)}$ has the same distribution as θ_k. We showed in part (d) that $\theta_k \sim \text{Beta}\left(\alpha_k, \sum_{i=1}^{K} \alpha_i - \alpha_k\right)$. Then $\lim_{n\to\infty} \rho_k^{(n)} \sim \text{Beta}\left(\alpha_k, \sum_{i=1}^{K} \alpha_i - \alpha_k\right)$.

Physicochemical Properties of Protein Tertiary Structure

In the following problems we will code two different approaches for solving linear regression problems and compare how they scale as a function of the dimensionality of the data. We will also investigate the effects of linear and non-linear features in the predictions made by linear models.

We will be working with the regression data set Protein Tertiary Structure: `https://archive.ics.uci.edu/ml/machine-learning-databases/00265/` (download CASP.csv). This data set contains information about predicted conformations for 45730 proteins. In the data, the target variable y is the root-mean-square deviation (RMSD) of the predicted conformations with respect to the true properly folded form of the protein. The RMSD is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed proteins. The features $\mathbf{x}$ are physico-chemical properties of the proteins in their true folded form. After downloading the file CASP.csv we can load the data into python using

```
>>> import numpy as np
>>> data = np.loadtxt("CASP.csv", delimiter = ",", skiprows = 1)
```

We can then obtain the vector of target variables and the feature matrix using

```
>>> y = data[:, 0]
>>> X = data[:, 1:]
```

We can then split the original data into a training set with 90% of the data entries in the file CASP.csv and a test set with the remaining 10% of the entries. Normally, the splitting of the data is done at random, but here **we ask you to put into the training set the first 90% of the elements from the file CASP.csv** so that we can verify that the values that you will be reporting are correct. (This should not cause problems, because the rows of the file are in a random order.)

We then ask that you **normalize** the features so that they have zero mean and unit standard deviation in the training set. This is a standard step before the application of many machine learning methods. After these steps are done, we can concatenate a **bias feature** (one feature which always takes value 1) to the observations in the normalized training and test sets.

We are now ready to apply our machine learning methods to the normalized training set and evaluate their performance on the normalized test set. In the following problems, you will be asked to report some numbers and produce some figures. Include these numbers and figures in your assignment report. **The numbers should be reported with up to 8 decimals**.

Problem 4 (7pts)

Assume that the targets y are obtained as a function of the normalized features $\mathbf{x}$ according to a Bayesian linear model with additive Gaussian noise with variance $\sigma^2 = 1.0$ and a Gaussian prior on the regression coefficients $\mathbf{w}$ with *precision* matrix $\Sigma^{-1} = \tau^{-2}\mathbf{I}$ where $\tau^{-2} = 10$. Code a routine using the **QR decomposition** (see Section 7.5.2 in Murphy's book) that finds the Maximum a Posteriori (MAP) value $\hat{\mathbf{w}}$ for $\mathbf{w}$ given the normalized training data

- Report the value of $\hat{\mathbf{w}}$ obtained.
- Report the root mean squared error (RMSE) of $\hat{\mathbf{w}}$ in the normalized test set.

$\hat{\mathbf{w}} = [7.74153395, 5.55782079, 2.25190765, 1.07880135, -5.91177796, -1.73480336, -1.63875478, -0.26610556,$

$0.81781409, -0.65913397]^T$ with the first value in the vector as the bias.

Normalized test set RMSE = 5.20880461

Problem 5 (14pts)

L-BFGS is an iterative method for solving general nonlinear optimization problems. For this problem you will use this method as a black box that returns the MAP solution by sequentially evaluating the objective function and its gradient for different input values. The goal of this problem is to use a built-in implementation of the L-BFGS algorithm to find a point estimate that maximizes our posterior of interest. Generally L-BFGS requires your black box to provide two values: the current objective and the gradient of the objective with respect to any parameters of interest. To use the optimizer, you need to first write two functions: (1) to compute the loss, or the *negative* log-posterior and (2) to compute the gradient of the loss with respect to the weights w.

As a preliminary to coming work in the class, we will use the L-BFGS implemented in PyTorch. [Warning: For this assignment we are using a small corner of the PyTorch world. Do not feel like you need to learn everything about this library.]

There are three parts to using this optimizer:

1. Create a vector of weights in NumPy, wrap in a pytorch `Tensor` and `Variable`, and pass to the optimizer.

```
from torch import Tensor
from torch.autograd import Variable

# Construct a PyTorch variable array (called tensors).
weights = Variable(Tensor(size))

# Initialize an optimizer of the weights
optimizer = torch.optim.LBFGS([weights])

...
```

2. Write a python function that uses the current weights to compute the log-posterior **and** sets weights.grad to be the gradient of the log-posterior with respect to the current weights.

```
def black_box():
    # Access the value of the variable as a numpy array.
    weights_data = weights.data.numpy()

    ...

    # Set the gradient of the variable.
    weights.grad = Tensor({numpy})

    return {objective}
```

3. Repeatedly call `optimizer.step(black_box)` to optimize.

[If you are feeling adventurous, you might find it useful to venture into the land of autograd and check your computation with PyTorch's `torch.autograd.gradcheck.get_numerical_jacobian`.]

- After running for 100 iterations, report the value of $\hat{\mathbf{w}}$ obtained.
- Report the RMSE of the predictions made with $\hat{\mathbf{w}}$ in the normalized test set.

$\hat{\mathbf{w}} = [7.74153376, 5.55782127, 2.25190735, 1.07880151, -5.91177797, -1.73480356, -1.63875508, -0.2661055, 0.81781411, -0.65913397]^T$ with the first value in the vector as the bias.

Normalized test set RMSE = 5.20880461

Problem 6 (14pts)

Linear regression can be extended to model non-linear relationships by replacing the original features $\mathbf{x}$ with some non-linear functions of the original features $\boldsymbol{\phi}(\mathbf{x})$. We can automatically generate one such non-linear function by sampling a random weight vector $\mathbf{a} \sim \mathcal{N}(0, \mathbf{I})$ and a corresponding random bias $b \sim \mathrm{U}[0, 2\pi]$ and then making $\phi(\mathbf{x}) = \cos(\mathbf{a}^{\mathrm{T}}\mathbf{x} + b)$. By repeating this process d times we can generate d non-linear functions that, when applied to the original features, produce a non-linear mapping of the data into a new d dimensional space. We can encode these d functions into a matrix $\mathbf{A}$ with d rows, each one with the weights for each function, and a d-dimensional vector $\mathbf{b}$ with the biases for each function. The new mapped features are then obtained as $\boldsymbol{\phi}(\mathbf{x}) = \cos(\mathbf{A}\mathbf{x} + \mathbf{b})$, where cos applied to a vector returns another vector whose elements are the result of applying cos to the individual elements of the original vector.

Generate 4 sets of non-linear functions, each one with $d = 100, 200, 400, 600$ functions, respectively, and use them to map the features in the original normalized training and test sets into 4 new feature spaces, each one of dimensionality given by the value of d. After this, for each value of d, find the MAP solution $\hat{\mathbf{w}}$ for $\mathbf{w}$ using the corresponding new training set and the method from problem 4. Use the same values for σ^2 and τ^{-2} as before. You are also asked to record the time taken by the method QR to obtain a value for $\hat{\mathbf{w}}$. In python you can compute the time taken by a routine using the time package:

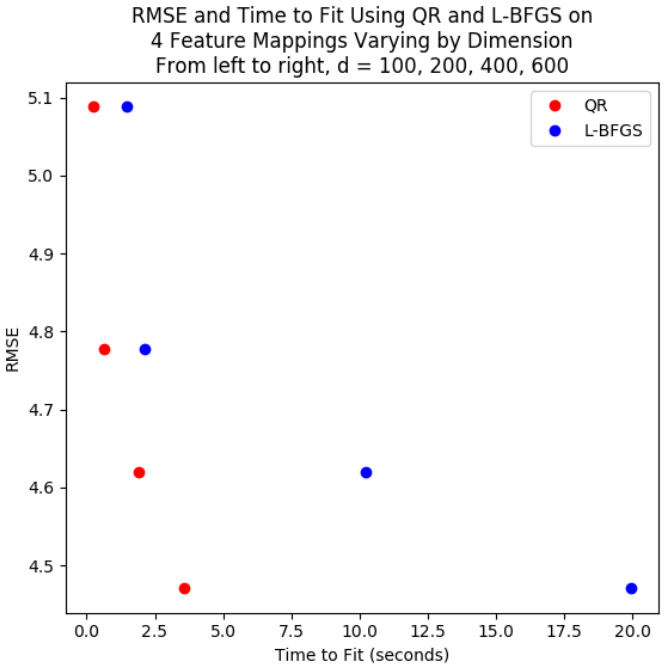
```
>>> import time
>>> time_start = time.time()
>>> routine_to_call()
>>> running_time = time.time() - time_start
```

Next, compute the RMSE of the resulting predictor in the normalized test set. Repeat this process with the method from problem 5 (L-BFGS).

- Report the test RMSE obtained by each method for each value of d.

You are asked to generate a plot with the results obtained by each method (QR and L-BFGS) for each value of d. In this plot the x axis should represent the time taken by each method to run and the y axis should be the RMSE of the resulting predictor in the normalized test set. The plot should contain 4 points in red, representing the results obtained by the method QR for each value of d, and 4 points in blue, representing the results obtained by the method L-BFGS for each value of d. Answer the following questions:

- Do the non-linear transformations help to reduce the prediction error? Why?
- What method (QR or L-BFGS) is faster? Why?
- (Extra Problem, Not Graded) Instead of using random $\mathbf{A}$, what if we treat $\mathbf{A}$ as another parameter for L-BFGS to optimize? You can do this by wrapping it as a variable and passing to the constructor. Compute its gradient as well in *black_box* either analytically or by using PyTorch *autograd*.




```
Test RMSE using QR when d = 100: 5.08817271221
Test RMSE using QR when d = 200: 4.77679289446
Test RMSE using QR when d = 400: 4.61913566655
Test RMSE using QR when d = 600: 4.47031426244
Test RMSE using L-BFGS when d = 100: 5.08817268403
Test RMSE using L-BFGS when d = 200: 4.77680654611
Test RMSE using L-BFGS when d = 400: 4.61914203587
Test RMSE using L-BFGS when d = 600: 4.47031404723
```

The results indicate that the non-linear transformations of the covariates reduced the prediction error. This is may be because our data follow a non-linear pattern with respect to some covariate(s), which is exactly when non-linear transformations of covariates are useful. When data follow a non-linear pattern, having no non-linear transformations of covariates will result in underfitting of the data. This was likely the case with our original set of covariates. Adding random noise to our original covariates and then using the cosine function to do a non-linear transformation worked well here, though I suspect that there are other transformations that would perform just as well.

The results also indicate that QR was faster than L-BFGS at coming up with the model fit for each d. For each increase in d, the time to complete model fitting using L-BFGS increased relative to the time it took to complete model fitting using QR.

Matrix inversion is normally the slowest part of coming up with the least squares or ridge solution to linear regression, but using QR decomposition means only having to invert an upper triangular matrix, which is relatively fast regardless of the value of d. L-BFGS involves using the gradient and Hessian matrix to update a d-dimensional vector of weights in order to find the set of d weights that minimize loss. When d becomes large, it makes sense that searching for the optimal weights would become slow. Note that it is possible that an alternative set of optimization parameters could speed up L-BFGS.