Question 1

1. **[2pts] Marginalization and Log-Sum-Exp.** In this question you will learn how to compute log marginal probabilities in a numerically stable way. Suppose that you have a generative model $p(x, i)$ for labeled data (x, i) where i is a label that can be one of $0, 1, 2, \ldots, k$. Recall that the marginal probability $p(x)$ can be computed using the following formula:

$$p(x) = \sum_{i=0}^{k} p(x, i) \tag{0.1}$$

When x is a high-dimensional data point, it is typical for the marginal $p(x)$ and the joint $p(x, i)$ to be extremely small numbers that cannot be represented in floating point. For this reason, we usually report and compute with *log probabilities*.

If we want to compute $\log(p(x))$ only given access to $\log(p(x, i))$, then we can use what is called a *log-sum-exp*:

$$\log\left(\sum_{i=0}^{k} \exp(a_i)\right) \tag{0.2}$$

where $a_i \in \mathbb{R}$ are real numbers. In our example, if $a_i = \log(p(x, i))$, then expression (0.2) would correspond to the log marginal probability $\log(p(x))$ of x.

Unfortunately, computing log-sum-exp naively can lead to numerical instabilities. The numerical instabilities in log-sum-exp are caused by problems that arise when trying to compute exponentials using floating point numbers. Two things can go wrong:

(a) **Underflow.** If `a[i]` is very small, then `np.exp(a[i])` will evaluate to `0`.

(b) **Overflow.** If `a[i]` is very large, then `np.exp(a[i])` will evaluate to `inf`.

The cause of underflow and overflow is that floating point numbers cannot represent numbers arbitrarily close to 0 nor arbitrarily large numbers.

(a) **[1pt]** We have provided code in `q1.py` with two implementations of log-sum-exp: a naive, numerically unstable implementation and a numerically stable one. Modify the elements of `a` so that `logsumexp_unstable` returns `-inf`, and modify the elements of `b` so that `logsumexp_unstable` returns `inf`. Report the two vectors, `a` and `b`, in your write-up.

```
a = np.array([-10**10, -10**10, -10**10])
b = np.array([10 ** 10, 0, 0])
```

```
In[2]: runfile('/Users/ChangyanXu/Desktop/code/q1.py', wdir='/Users/ChangyanXu/Desktop/code')
Computing the log-sum-exp of a=[-10000000000 -10000000000 -10000000000]
    Unstable: -inf
    Stable: -9999999998.981388
Computing the log-sum-exp of b=[10000000000           0           0]
    Unstable: inf
    Stable: 10000000000.0
```

(b) **[1pt]** Prove that our numerically stable implementation is correct by proving that

$$\log\left(\sum_{i=0}^{k} \exp(a_i)\right) = \log\left(\sum_{i=0}^{k} \exp\left(a_i - \max_{j=0}^{k}\{a_j\}\right)\right) + \max_{j=0}^{k}\{a_j\} \tag{0.3}$$

Briefly explain why the numerically stable version is robust to underflow and overflow.

let $m = \max_{j=0}^{k}\{a_j\}$

then LHS $= \log\left(\sum_{i=0}^{k} \exp(a_i - m + m)\right)$

$= \log\left(\sum_{i=0}^{k} \exp(a_i - m) \cdot \exp(m)\right)$

$= \log\left(\exp(m) \sum_{i=0}^{k} \exp(a_i - m)\right)$

$= \log\left(\exp(m)\right) + \log\left(\sum_{i=0}^{k} \exp(a_i - m)\right)$

$= m + \log\left(\sum_{i=0}^{k} \exp(a_i - m)\right)$

$= \max_{j=0}^{k}\{a_j\} + \log\left(\sum_{i=0}^{k} \exp(a_i - \max_{j=0}^{k}\{a_j\})\right) = \text{RHS}$

Hence proved

If any a_i is very large: now as we subtract the maximum number in vector a from a_i, the (a_i - max{a_j}) would be <= 0, then 0 <= exp(a_i - max{a_j}) <= 1. So when we apply the log to the summation of exponentials, we may get very large numbers because of log, but we would not get an overflow problem.

If all a_i are very small: now as we subtract the maximum number in vector a from a_i, the (a_i - max{a_j}) would be close to 0, then exp(a_i - max{a_j}) for all a_i are close to 1. So when we apply the log to the summation of exponentials, we would not get an underflow problem.

Question 2

2. **[3pts] Gaussian Discriminant Analysis.** For this question you will build classifiers to label images of handwritten digits. Each image is 8 by 8 pixels and is represented as a vector of dimension 64 by listing all the pixel values in raster scan order. The images are grayscale and the pixel values are between 0 and 1. The labels y are $0, 1, 2, \ldots, 9$ corresponding to which character was written in the image. There are 700 training cases and 400 test cases for each digit; they can be found in `a4digits.zip`.

   A skeleton (`q2.py`) is is provided for each question that you should use to structure your code. Starter code to help you load the data is provided (`data.py`). Note: the `get_digits_by_label` function in `data.py` returns the subset of digits that belong to a given class.

   Using maximum likelihood, fit a set of 10 class-conditional Gaussians with a separate, full covariance matrix for each class. Remember that the conditional multivariate Gaussian probability density is given by,

$$p(\mathbf{x} \mid y = k, \boldsymbol{\mu}, \Sigma_k) = (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \tag{0.4}$$

   You should take $p(y = k) = \frac{1}{10}$. You will compute parameters μ_{kj} and Σ_k for $k \in (0...9), j \in (1...64)$. You should implement the covariance computation yourself (i.e. without the aid of 'np.cov'). *Hint: To ensure numerical stability you may have to add a small multiple of the identity to each covariance matrix. For this assignment you should add* $0.01\mathbf{I}$ *to each matrix.*

Sample mean:
$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)}$$

Sample covariance:
$$\begin{aligned} \hat{\boldsymbol{\Sigma}} &= \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})^\top \\ &= \frac{1}{N} (\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^\top)^\top (\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^\top) \end{aligned}$$

(a) **[1pt]** Using the parameters you fit on the training set and Bayes rule, compute the average conditional log-likelihood, i.e. $\frac{1}{N} \sum_{i=1}^{N} \log(p(y^{(i)} \mid \mathbf{x}^{(i)}, \theta))$ on both the train and test set and report it. *Hint: you will want to use the log-sum-exp we discussed in Question 1 to your code.*

conditional likelihood $\log p(y|x)$

generative likelihood $\log p(x|y)$

let $a_i = -\frac{1}{2}(\mathbf{x}^{(i)} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \mu_k)$

let $m = \max_{x^{(i)} \in D} \{-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \mu_k)\} = \max_{x^{(i)} \in D} \{a_i\}$

$$\log p(x \mid y=k, \mu, \Sigma_k) = \log\left[(2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\{ a_i - m + m \} \right]$$

$$= \log\left[(2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\{ a_i - m \} \right] + m$$

$$\log \sum_{k=0}^{9} \underbrace{p(x|y=k)}_{n \times 1} \underbrace{p(y=k)}_{1 \times 10} = \log \sum_{k=0}^{9} \left[(2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\{ a_i - m + m \} \right] \cdot \frac{1}{10}$$

$$= \log \sum_{k=0}^{9} \left[(2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\{ a_i - m + m \} \right] \cdot \frac{1}{10}$$

$$= m + \log \sum_{k=0}^{9} \left[(2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\{ a_i - m \} \right] \cdot \frac{1}{10}$$

$$= m + \log \frac{1}{10} + \log \sum_{k=0}^{9} \left[(2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\{ a_i - m \} \right]$$

$$p(\theta \mid D) = \frac{p(\theta)\, p(D \mid \theta)}{\int p(\theta')\, p(D \mid \theta')\, d\theta'}$$

known $P(x|y=k)$

$P(y=k) = \frac{1}{10}$

$$\log \underbrace{p(y|x,\mu,\Sigma)}_{n\times 10} = \log \frac{p(x|y)\cdot p(y)}{p(x)}$$

$$= \log p(x|y) + \log p(y) - \log p(x)$$

$$= \log p(x|y) + \log p(y) - \log \sum_{k=0}^{9} p(x|y=k)\, p(y=k)$$

$$= \log\left[(2\pi)^{-d/2}|\Sigma_k|^{-1/2}\exp\{a_i\}\right] + \log\frac{1}{10} - \left[m + \log\frac{1}{10} + \log\sum_{k=0}^{9}\left[(2\pi)^{-d/2}|\Sigma_k|^{-1/2}\exp\{a_i - m\}\right]\right]$$

$$= \cancel{\log[2\pi^{-d/2}]} + \log[|\Sigma_k|^{-1/2}] + a_i + \cancel{\log\frac{1}{10}} - m - \cancel{\log\frac{1}{10}} - \cancel{\log[2\pi^{-d/2}|\Sigma_k|^{-1/2}]} + \log\sum_{k=0}^{9}[\exp\{a_i - m\}]$$

$$= a_i + \log\sum_{k=0}^{9}[\exp\{a_i - m\}]$$

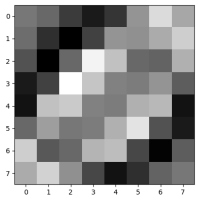
Average Conditional log-likelihood:

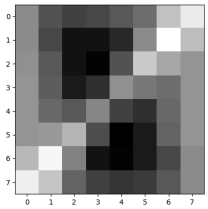
```
The avg_conditional_likelihood for TRAIN:
 -0.1824634072501818
The avg_conditional_likelihood for TEST:
 -0.19427133853939788
```

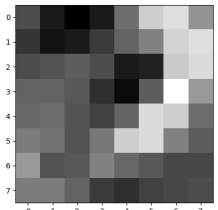
(b) **[1pt]** Select the most likely posterior class for each training and test data point as your prediction, and report your accuracy on the train and test set.

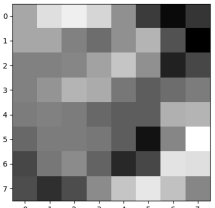
```
train accuracy:  0.9805714285714285
test accuracy:  0.96975
```

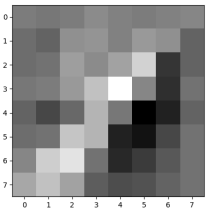
(c) **[1pt]** Compute the leading eigenvectors (largest eigenvalue) for each class covariance matrix (can use `np.linalg.eig`) and plot them side by side as 8 by 8 images.

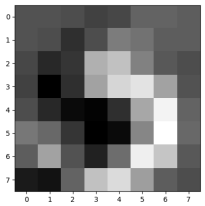

k=0

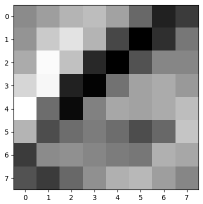

k=1

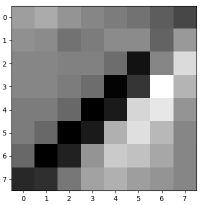

k=2

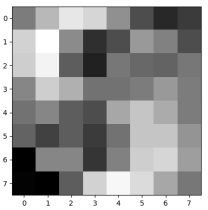

k=3

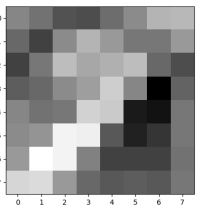

k=4


k=5


k=6


k=7


k=8


k=9

3. **[3pts] Categorial Distribution.** In this problem you will consider a Bayesian approach to modelling categorical outcomes. Let's consider fitting the categorical distribution, which is a discrete distribution over K outcomes, which we'll number 1 through K. The probability of each category is explicitly represented with parameter θ_k. For it to be a valid probability distribution, we clearly need $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. We'll represent each observation $\mathbf{x}$ as a 1-of-K encoding, i.e, a vector where one of the entries is 1 and the rest are 0. Under this model, the probability of an observation can be written in the following form:

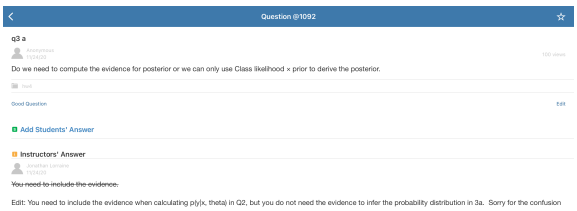
$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}.$$

Suppose you observe a dataset,

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}.$$

Denote the count for outcome k as $N_k = \sum_{i=1}^{n} x_k^{(i)}$. Recall that each data point is in the 1-of-K encoding, i.e., $x_k^{(i)} = 1$ if the ith datapoint represents an outcome k and $x_k^{(i)} = 0$ otherwise. In the previous assignment, you showed that the maximum likelihood estimate for the counts was:

$$\hat{\theta}_k = \frac{N_k}{N}.$$



(a) **[1pts]** For the prior, we'll use the Dirichlet distribution, which is defined over the set of probability vectors (i.e. vectors that are nonnegative and whose entries sum to 1). Its PDF is as follows:

$$p(\boldsymbol{\theta}) \propto \theta_1^{a_1-1} \cdots \theta_K^{a_k-1}.$$

What is the probability distribution of the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$?

$p(\theta) \propto \theta_1^{a_1-1} \cdots \theta_k^{a_k-1}$, $\quad p(D|\theta) = \prod_{k=1}^{K} \theta_k^{D_k}$

$$p(\theta \mid D) = \frac{p(\theta)\, p(D|\theta)}{\int p(\theta')\, p(D|\theta')\, d\theta'}$$

$$\propto \left[\theta_1^{a_1-1} \cdots \theta_k^{a_k-1}\right]\left[\prod_{k=1}^{K} \theta_k^{D_k}\right]$$

* evidence does not need to be included, as instructed from piazza.

$$= \theta_1^{a_1-1+D_1} \cdots \theta_k^{a_k-1+D_k}$$

$$= \prod_{k=1}^{K} \theta_k^{a_k-1+D_k}$$

(b) **[1pt]** Still assuming the Dirichlet prior distribution, determine the MAP estimate of the parameter vector $\boldsymbol{\theta}$. For this question, you may assume each $a_k > 1$.

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(\theta \mid D)$$

$$= \underset{\theta}{\operatorname{argmax}}\ p(\theta, D)$$

$$= \underset{\theta}{\operatorname{argmax}}\ p(\theta)\, p(D|\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \log p(\theta) + \log p(D|\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \log \prod_{k=1}^{K} \theta_k^{a_k-1+D_k}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \sum_{k=1}^{K} \log \theta_k^{a_k-1+D_k}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \sum_{k=1}^{K} (a_k - 1 + D_k) \log \theta_k$$

$$0 = \frac{d}{d\theta} \log p(\theta, \mathcal{D}) = \frac{d}{d\theta} \sum_{k=1}^{K} (a_k - 1 + D_k) \log \theta_k$$

$$\Rightarrow \quad \ell(\theta) = \sum_{i=1}^{k-1} (D_i + a_i - 1) \log \theta_i + (D_k + a_k - 1) \log \Big(1 - \sum_{j=1}^{k-1} \theta_j \Big)$$

$$\Rightarrow \quad \frac{\partial \ell(\theta)}{\partial \theta_i} = \frac{(D_i + a_i - 1)}{\theta_i} - \frac{(D_k + a_k - 1)}{1 - \sum_{j=1}^{k-1} \theta_j}$$

$$= \frac{(D_i + a_i - 1)}{\theta_i} - \frac{(D_k + a_k - 1)}{\theta_k} = 0$$

$$\Rightarrow \quad \frac{\hat{\theta}_k}{\hat{\theta}_i} = \frac{(D_k + a_k - 1)}{(D_i + a_i - 1)}$$

$$\sum_{i=1}^{K} \theta_i = 1 \quad \Rightarrow \quad \hat{\theta}_k + \sum_{i=1}^{k-1} \hat{\theta}_i = 1$$

$$\Rightarrow \quad \hat{\theta}_k + \sum_{i=1}^{k-1} \hat{\theta}_k \frac{(D_i + a_i - 1)}{(D_k + a_k - 1)} = 1$$

$$= \hat{\theta}_k \sum_{i=1}^{k} \frac{(D_i + a_i - 1)}{(D_k + a_k - 1)} = 1$$

$$= \frac{\hat{\theta}_k \sum_{i=1}^{k} (D_i + a_i - 1)}{(D_k + a_k - 1)} = 1$$

$$= \frac{\hat{\theta}_i \sum_{i=1}^{k} (D_i + a_i - 1)}{(D_i + a_i - 1)} = 1$$

$$\Rightarrow \quad \hat{\theta}_i = \frac{(D_i + a_i - 1)}{\sum_{i=1}^{k} (D_i + a_i - 1)}$$

$$= \frac{(D_i + a_i - 1)}{D + \sum_{i=1}^{k} a_i - K}$$

$$\Rightarrow \quad \boxed{\hat{\theta}_k = \frac{(D_k + a_k - 1)}{D + \sum_{k=1}^{k} a_k - K}}$$

$$\boxed{\hat{\theta}_{MAP} = \frac{1}{D + \sum_{k=1}^{k} a_k - K} \begin{bmatrix} D_1 + a_1 - 1 \\ \vdots \\ D_k + a_k - 1 \end{bmatrix}}$$

(c) [**1pts**] Now, suppose that your friend said that they had a hidden $N+1$st outcome, $\mathbf{x}^{(N+1)}$, drawn from the same distribution as the previous N outcomes. Your friend does not want to reveal the value of $\mathbf{x}^{(N+1)}$ to you. So, you want to use your Bayesian model to predict what *you* think $\mathbf{x}^{(N+1)}$ is likely to be. The "proper" Bayesian predictor is the so-called *posterior predictive distribution*:

$$p(\mathbf{x}^{(N+1)}|\mathcal{D}) = \int p(\mathbf{x}^{(N+1)}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}$$

What is the probability that the $N+1$ outcome was k, i.e., the probability that $x_k^{(N+1)} = 1$, under your posterior predictive distribution? *Hint: A useful fact is that if* $\boldsymbol{\theta} \sim$ Dirichlet$(a_1, \ldots, a_K)$, *then*

$$\mathbb{E}[\theta_k] = \frac{a_k}{\sum_{k'} a_{k'}}.$$

$$p(x^{(N+1)}=k|D) = \int p(x^{(N+1)}=k|\theta)\, p(\theta|D)\, d\theta$$

$$= \int p(x^{(N+1)}=k|\theta_k)\; p(\theta_k|D)\; d\theta$$

$$= \int \theta_k \prod_{k=1}^{K} \theta_k^{a_k+\alpha_k-1}\, d\theta_k$$

$$= \int \prod_{k=1}^{K} \theta_k^{a_k+\alpha_k}\, d\theta_k$$

$$= \mathbb{E}[\theta_k|D]$$

$$= \frac{a_k+N_k}{\sum_{k'}(a_{k'}+N_{k'})}$$

since we previously defined

$$\theta_k = p(x^{(i)}=k|\theta)$$