CS5340: Uncertainty Modeling in AI

Tutorial 2: Solutions

Released: Jan. 26, 2023

Problem 1. (Uncorrelated Random Variables)

Consider two random variables X and Y, where $\text{Cov}[X, Y] = 0$. Recall that

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Problem 1.a. Your friend Donald says, "Since the covariance of X and Y is zero, then it must be case that that X and Y are independent!" Is Donald correct? If yes, provide a proof. If not, give a counterexample. *Hint:* Can you find some function f such that $Y = f(X)$ and $\text{Cov}[X, Y] = 0$?

Solution: In general, $\text{Cov}(X, Y) = 0$ does not imply X and Y are independent (however, if X and Y are independent, then $\text{Cov}(X, Y) = 0$). It is important to note that covariance is a measure of a *linear* relationship between the variables. It is possible for two variables X and Y to be related in a non-linear way.

Consider the following example: suppose $X \sim \text{Uniform}(-1, 1)$ and $Y = X^2$ (for this counterexample to work, you can pick a different distribution for X as long as $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^3] = 0$, e.g., $X \sim \mathcal{N}(0, 1)$). Clearly, X and Y are dependent; in fact, Y is *deterministically* related to X. Let's look at $\text{Cov}(X, Y)$

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] && (1)\\ &= \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] && (2) \end{aligned}$$

Since,

$$\mathbb{E}[X] = \int_{-1}^{1} xp(x)dx = \int_{-1}^{1} \frac{x}{2}dx = 0 \tag{3}$$

and

$$\mathbb{E}[X^3] = \int_{-1}^{1} x^3 p(x)dx = \int_{-1}^{1} \frac{x^3}{2}dx = 0 \tag{4}$$

Thus,

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] && (5)\\ &= 0 && (6) \end{aligned}$$

Problem 1.b. Consider we obtain samples $x_1, x_2, \ldots, x_N$ drawn iid from $p(X)$ and another batch of samples $y_1, y_2, \ldots, y_N$ drawn iid from $p(Y)$. Suppose we want to model the joint distribution $p_\theta(X, Y)$. Donald suggests we perform maximum likelihood estimation by finding

$$\arg\max_\theta \sum_i^N \log p_\theta(x_i, y_i).$$

Is Donald correct? Justify your answer.

Solution: Donald is incorrect in general. If X and Y are dependent, $p(X, Y)$ *cannot* be factorized into $p(X, Y) = p(X)p(Y)$. Since the pairs (x_i, y_i) are not sampled from joint distribution $(x_i, y_i) \sim p(X, Y)$, performing MLE will not result in learning the correct distribution parameters.

Note: MLE can work if X and Y happen to be *independent*.

Problem 2. (Exponential Family)

In Lecture 2, we learned about the Exponential Family (ExpFam). In this tutorial, we'll make the general concept more concrete with some examples. Recall that ExpFam distributions have the following (natural) form:

$$p_\eta(x) = \frac{h(x)\exp[\eta^\top s(x)]}{Z(\eta)} \quad \text{or} \quad p_\theta(x) = h(x)\exp[\eta^\top s(x) - A(\eta)]$$

where $\eta = \eta(\theta)$ are the natural parameters, $s(x)$ are the sufficient statistics, $h(x)$ is the base measure, and $Z(\eta)$ is the partition function. In an alternative form (on the right above), $A(\eta)$ is the log partition function or cumulant function.

Problem 2.a. We'll begin with the **exponential distribution**[1]. Consider a process where events occur continuously and independently at some average rate λ ($\lambda > 0$). Real-world examples include radioactive decay, customer arrival times, and machine failure times. Let x be the time between events. The exponential distribution models the probability distribution of x:

$$p(x|\lambda) = \begin{cases} \lambda\exp(-\lambda x) & x \geq 0, \\ 0 & x < 0, \end{cases} \tag{7}$$

where $\lambda > 0$. Show that the Exponential distribution is ExpFam.
Hint: By rearranging elements, try to rewrite the exponential distribution in terms of the natural parameters, sufficient statistics, base measure, and (log) partition function. To get you started, the base measure is $h(x) = 1$ for $x \geq 0$, which we will assume for this example.

Solution: The exponential distribution is already in ExpFam form. Just relate $h(x) = 1$ for $x \geq 0$ and $h(x) = 0$ else, $\eta = \lambda$, $s(x) = -x$, and $Z(\eta) = 1/\lambda$ (or $A(\eta) = \log\frac{1}{\lambda}$).

[1] Note that the exponential distribution is *not* the same as the exponential family. The exponential distribution is a specific distribution, but the exponential family is a more general form.

Problem 2.b. Assume that you have access to data $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$. Use facts about ExpFam to help you derive the MLE of the natural parameters of the exponential distribution. Recall that:

$$\mathbb{E}[s(x)] = \nabla \log Z(\eta) = \nabla A(\eta) \tag{8}$$

and that the maximum likelihood estimator η_{MLE} satisfies the condition that

$$\nabla A(\eta_{\text{MLE}}) = \frac{1}{N} \sum_{i=1}^{N} s(x_i) \tag{9}$$

Hint: The MLE estimator for the exponential satisfies $\mathbb{E}_{\lambda_{\text{MLE}}}[s(x)] = \frac{1}{N} \sum_{i=1}^{N} s(x_i)$.

Solution: We specialize the general equations above for the exponential distribution. so, we let

$$\mathbb{E}_{\lambda_{\text{MLE}}}[s(x)] = \frac{1}{N} \sum_{i=1}^{N} s(x_i) \tag{10}$$

Applying eqn. (8) above to the case of the exponential distribution:

$$\frac{1}{N} \sum_{i=1}^{N} s(x_i) = \frac{d}{d\eta} A(\eta) \tag{11}$$

$$= \frac{1}{\frac{1}{\eta}} \cdot \left(-\frac{1}{\eta^2} \right) \tag{12}$$

$$= -\frac{1}{\eta} \tag{13}$$

Since $s(x_i) = -x_i$,

$$-\frac{1}{\eta} = -\frac{1}{N} \sum_{i=1}^{N} x_i \tag{14}$$

so,

$$\eta_{\text{MLE}} = \frac{1}{\frac{1}{N} \sum_{i=1}^{N} x_i} = \frac{1}{\bar{x}} \tag{15}$$

where $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ is the sample mean. Similarly, $\lambda_{\text{MLE}} = \eta_{\text{MLE}} = \frac{1}{\frac{1}{N} \sum_{i=1}^{N} x_i} = \frac{1}{\bar{x}}$

Problem 2.c. Repeat the two problems above for the (univariate) **Gaussian** distribution, i.e.,

1. show that the Gaussian is exponential family, and
2. derive the MLE of its natural parameters by leveraging the properties of ExpFam distributions.

(*Challenge*) If you are feeling like a challenge, show that the *multivariate Gaussian* is also exponential family, and derive the MLE of its natural parameters.

Solution:
Solutions for univariate Gaussian: We first prove that univariate Gaussian is in exponential family

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{16}$$

$$= \frac{1}{\sqrt{2\pi}} \exp(\log(\frac{1}{\sigma}) \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{17}$$

$$= \frac{1}{\sqrt{2\pi}} \exp(-\log(\sigma) - \frac{x^2+\mu^2-2x\mu}{2\sigma^2}) \tag{18}$$

$$= \frac{1}{\sqrt{2\pi}} \exp(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma) \tag{19}$$

$$= \frac{1}{\sqrt{2\pi}} \exp(\begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}^T \begin{pmatrix} x \\ x^2 \end{pmatrix} - (\frac{\mu^2}{2\sigma^2} + \log\sigma)) \tag{20}$$

Then we can set

$$h(\mathbf{x}) = \frac{1}{\sqrt{2\pi}} \tag{21}$$

$$\eta = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \tag{22}$$

$$s(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \tag{23}$$

$$A(\eta) = \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) \tag{24}$$

$$\tag{25}$$

And we get the form of exponential family Now let's derive the MLE of its natural parameters. Following eqn.9, we have

$$\frac{\partial A}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{26}$$

Then

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{27}$$

$$\frac{\partial A}{\partial \eta_2} = \frac{\eta_{1_{MLE}}^2}{4\eta_2^2} - \frac{1}{2\eta_2} = \mu_{MLE}^2 - \frac{1}{2\eta_2} = \mu_{MLE}^2 + \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{28}$$

$$\tag{29}$$

From the above two equation, we can get

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - \mu^2_{MLE}) \tag{30}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2\mu^2_{MLE} + \mu^2_{MLE}) \tag{31}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2\mu_{MLE}x_i + \mu^2_{MLE}) \tag{32}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{MLE})^2 \tag{33}$$

Substituting them into η, then we have $\eta_{MLE} = \begin{pmatrix} \frac{\bar{x}}{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2} \\ -\frac{1}{2\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2} \end{pmatrix}$, where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the sample mean.

Solution for multivariate Gaussian: The density function of multivariate Gaussian can be written as

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-u)^T\Sigma^{-1}(x-u)\right\} \tag{34}$$

Define $\Lambda = \Sigma^{-1}$ and $\eta = \Sigma^{-1}\mu$. Substitute them into the density function, we get

$$p(x) = (2\pi)^{-\frac{d}{2}}\exp\left\{-\frac{1}{2}x^T\Lambda x + x^T\eta - \frac{1}{2}\eta^T\Lambda^{-1}\eta + \frac{1}{2}\log|\Lambda|\right\} \tag{35}$$

By using Frobenius inner product and the property that

$$x^TAx = tr\left[Axx^T\right] \tag{36}$$

we can get

$$-\frac{1}{2}x^T\Lambda x = \mathrm{tr}\left[-\frac{1}{2}\Sigma^{-1}xx^T\right] \tag{37}$$

$$= -\frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\right)^T\cdot\mathrm{vec}\left(xx^T\right) \tag{38}$$

$$= -\frac{1}{2}\mathrm{vec}\left(\Lambda\right)^T\cdot\mathrm{vec}\left(xx^T\right) \tag{39}$$

where $\mathrm{vec}(A)$ reshapes a matrix A into a vector (eq.40).

$$\mathrm{vec}\left(A\right) = \mathrm{vec}\left(\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}\right) = \begin{bmatrix} a_{11} \\ \vdots \\ a_{1m} \\ \vdots \\ a_{nm} \end{bmatrix} \tag{40}$$

By taking transpose of the scalar $x^T\eta$, we can also get

$$x^T\eta = \left(x^T\eta\right)^T = \eta^T x \tag{41}$$

Then, we can get

$$-\frac{1}{2}x^T\Lambda x + x^T\eta = \begin{bmatrix}\mathrm{vec}\left(\Lambda\right)\\ \eta\end{bmatrix}^T \cdot \begin{bmatrix}-\frac{1}{2}\mathrm{vec}\left(xx^T\right)\\ x\end{bmatrix} \tag{42}$$

Therefore, we can set

$$\hat{\eta} = \begin{bmatrix}\mathrm{vec}\left(\Lambda\right)\\ \eta\end{bmatrix}, s(x) = \begin{bmatrix}-\frac{1}{2}\mathrm{vec}\left(xx^T\right)\\ x\end{bmatrix} \tag{43}$$

and also

$$A(\hat{\eta}) = \frac{1}{2}\eta^T\Lambda^{-1}\eta - \frac{1}{2}\log|\Lambda| \tag{44}$$

Now, we look at how to derive the MLE for model parameters. Recall that

$$\nabla A(\hat{\eta}) = \frac{1}{N}\sum_{i=1}^{N} s(x_i) \tag{45}$$

We first look at the second part of natural parameter

$$\frac{\partial A}{\partial \eta} = \Lambda^{-1}\eta = \Sigma\Sigma^{-1}\mu = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{46}$$

Therefore, $\mu_{\mathrm{MLE}} = \frac{1}{N}\sum_{i=1}^{N} x_i$. Then we look at the first part of natural parameter, due to the definition of matrix derivative, we can calculate $\frac{\partial A}{\partial \Lambda}$ instead of $\frac{\partial A}{\partial \mathrm{vec}(\Lambda)}$, since they are only different at their shapes.

$$\begin{aligned}
\frac{\partial A}{\partial \Lambda} &= \frac{\partial}{\partial \Lambda}(\frac{1}{2}\eta^T\Lambda^{-1}\eta - \frac{1}{2}\log|\Lambda|) && (47)\\
&= \frac{\partial}{\partial \Lambda}(\frac{1}{2}\Lambda^{-1}\eta\eta^{\mathrm{T}} - \frac{1}{2}\log|\Lambda|) && (48)\\
&= -\frac{1}{2}\Lambda^{-1}\eta\eta^T\Lambda^{-1} - \frac{1}{2}\cdot\frac{1}{|\Lambda|}\cdot\frac{\partial|\Lambda|}{\Lambda} && (49)\\
&= -\frac{1}{2}\Lambda^{-1}\Lambda\eta\eta^T\Lambda^T\Lambda^{-1} - \frac{1}{2}\cdot\frac{1}{|\Lambda|}\cdot|\Lambda|\cdot\Lambda^{-1} && (50)\\
&= -\frac{1}{2}\mu\mu^T - \frac{1}{2}\Lambda^{-1} = -\frac{1}{2}\cdot\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T && (51)
\end{aligned}$$

Then, we get

$$\Lambda_{\mathrm{MLE}}^{-1} = \Sigma_{\mathrm{MLE}} = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T - \mu\mu^T \tag{52}$$

We can write precision matrix into a more elegant way. Suppose $(x_i)_k$ is the kth element of vector x_i.

$$
\begin{aligned}
(\Sigma)_{k,l} &= \frac{1}{N}\sum_{i=1}^{N}(x_i)_k\,(x_i)_l - (\mu)_k(\mu)_l && (53)\\
&= \frac{1}{N}\left(\sum_{i=1}^{N}(x_i)_k\,(x_i)_l - \sum_{i=1}^{N}(\mu)_k(\mu)_l\right) && (54)\\
&= \frac{1}{N}\left(\sum_{i=1}^{N}(x_i)_k\,(x_i)_l - 2\sum_{i=1}^{N}(\mu)_k(\mu)_l + \sum_{i=1}^{N}(\mu)_k(\mu)_l\right) && (55)\\
&= \frac{1}{N}\left(\sum_{i=1}^{N}(x_i)_k\,(x_i)_l - \sum_{i=1}^{N}(x_i)_k(\mu)_l - \sum_{i=1}^{N}(\mu)_k(x_i)_l + \sum_{i=1}^{N}(\mu)_k(\mu)_l\right) && (56)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left((x_i)_k\,(x_i)_l - (x_i)_k(\mu)_l - (\mu)_k(x_i)_l + (\mu)_k(\mu)_l\right) && (57)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left((x_i)_k - (\mu)_k\right)\left((x_i)_l - (\mu)_l\right) && (58)
\end{aligned}
$$

Then, we get

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)\,(x_i-\mu)^T \tag{59}$$

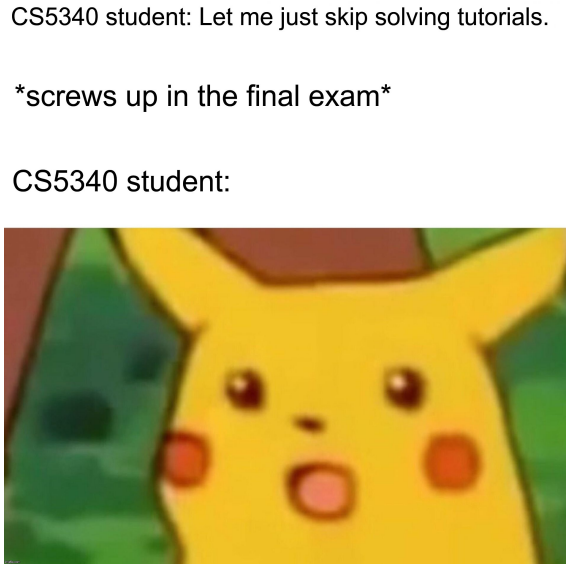
Finally, we can get

$$
\begin{aligned}
\mu &= \frac{1}{N}\sum_{i=1}^{N}x_i && (60)\\
\Sigma &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)\,(x_i-\mu)^T && (61)
\end{aligned}
$$

And we can get MLE for natural parameters as

$$\hat{\eta} = \begin{bmatrix}\mathrm{vec}\,(\Lambda)\\ \eta\end{bmatrix} = \begin{bmatrix}N\cdot\mathrm{vec}\left(\left(\sum_{i=1}^{N}(x_i-\mu)\,(x_i-\mu)^T\right)^{-1}\right)\\ \left(\sum_{i=1}^{N}(x_i-\mu)\,(x_i-\mu)^T\right)^{-1}\cdot\sum_{i=1}^{N}x_i\end{bmatrix} \tag{62}$$

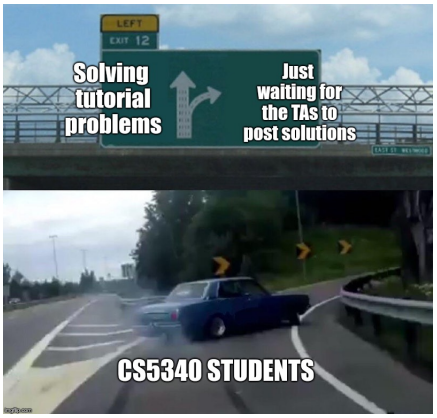
Problem 3. (Meme of the Year)
A poll was conducted amongst CS5340 students to pick the best meme template in 2018. The four meme templates that were in the run for *meme of the year* 2018 are shown in Fig. 1. The votes received by each meme template are tabulated in Table 1. Denote the vote of i-th student by a one-hot vector $\mathbf{x}_i$ and the entire dataset by $\mathcal{X} = \{\mathbf{x}\}_{i=1}^{N}$. For example, someone who voted for "Surprised Pikachu" will have $\mathbf{x} = \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}^\top$.



(a) Surprised Pikachu

(b) Two Buttons Dilemma

(c) Distracted Boyfriend



(d) Left Exit 12

Figure 1: The four meme templates in the run for meme of the year 2018

ID	Template Name	# Votes
1	Surprised Pikachu	25
2	Two Buttons Dilemma	12
3	Distracted Boyfriend	30
4	Left Exit 12	10

Table 1: Votes received by each template by CS5340 students

Problem 3.a. Fit an appropriate distribution to the data given above and compute the parameters of the distribution using maximum likelihood estimation.

Solution: Since the given problem involves a choice from a set of categories and we observe individual votes, a suitable distribution to model this problem is the **categorical distribution**. If you only observed counts (rather than individual votes), then a multinomial distribution would be a better model.

A categorical distribution is defined by parameters $\lambda_1, \ldots, \lambda_k$ that represent the probability of each category. As all the probabilities must sum to 1, we have the constraint that $\sum_{j=1}^{k} \lambda_j = 1$.

The likelihood of the categorical distribution is by

$$p(\mathcal{X}|\lambda_1, \ldots, \lambda_k) = \prod_{i=1}^{n} \prod_{j=1}^{k} \lambda_j^{[\mathbf{x}_{i,j}=1]} \qquad \text{such that } \sum_{j=1}^{k} \lambda_j = 1 \tag{63}$$

where $[\mathbf{x}_{i,j} = 1]$ means that the j-th element of the vector $\mathbf{x}_i$ is equal to 1.

For our specific problem, n (the number of data points) = 77 and k (the number of categories) = 4. However, let's derive the general estimators of the parameters of the categorical distribution.

The distribution in Eq. (63) can also be written as

$$p(\mathcal{X}|\lambda_1, \ldots, \lambda_k) = \prod_{j=1}^{k} \lambda_j^{N_j} \qquad \text{such that } \sum_{j=1}^{k} \lambda_j = 1 \tag{64}$$

where N_j is number of times the j-th element is 1 in the dataset or the count of each category.

We can apply log to the likelihood and write the constraint $\sum_{j=1}^{k} \lambda_j = 1$ with a lagrange multiplier ν to get the final auxiliary function

$$\mathcal{L} = \sum_{j=1}^{k} N_j \log \lambda_j + \nu \left(\sum_{j=1}^{k} \lambda_j - 1 \right) \tag{65}$$

For finding the maximum likelihood parameters we differentiate Eq. (65) w.r.t each parameter λ_j and set it equal to 0.

$$\frac{\partial \mathcal{L}}{\partial \lambda_j} = \frac{N_j}{\lambda_j} + \nu = 0 \tag{66}$$

$$\Rightarrow \lambda_j = -\frac{N_j}{\nu} \tag{67}$$

but we also have the constraint $\sum_{j=1}^{k} \lambda_j = 1$, which means $\nu = -\sum_{j=1}^{k} N_j$ and finally

$$\lambda_j = \frac{N_j}{\sum_{l=1}^{k} N_l} \tag{68}$$

Parameters	Values
λ_1	$\frac{25}{77} = 0.324$
λ_2	0.156
λ_3	0.390
λ_4	0.130

Table 2: Parameters estimated using MLE

Problem 3.b. Imagine that before looking at the poll results from the class you had access to poll results from another such poll that was conducted on Reddit, the results of which are shown in Table 3.

ID	Template Name	# Votes
1	Surprised Pikachu	250
2	Two Buttons Dilemma	110
3	Distracted Boyfriend	280
4	Left Exit 12	140

Table 3: Votes received by each template on Reddit

- Choose an appropriate distribution to incorporate this prior knowledge into your model. Derive an expression for the posterior distribution.
- Derive an expression for posterior predictive distribution $p(\mathbf{x}^*|\mathcal{X})$. Specifically, compute the probability $p(\mathbf{x}^* = \begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix}^\top |\mathcal{X})$.

Solution:

- As we're dealing with the categorical distribution, a suitable prior would be the Dirichlet distribution which is the conjugate prior for categorical distribution. The Dirichlet distribution is parameterized by the vector $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_{k-1} & \alpha_k\end{bmatrix}^\top$ and the PDF is given by

$$p(\lambda_1, \dots, \lambda_k) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{j=1}^{k} \lambda_j^{\alpha_j - 1} \qquad \text{where } \mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)} \tag{69}$$

  As Dirichlet is a conjugate prior, the posterior $p(\lambda_1, \dots, \lambda_k|\mathcal{X})$ is also a Dirichlet distribution. It can be shown that (similar to the derivation of posterior of Beta-Binomial from tutorial 1) the posterior Dirichlet distribution has the parameter vector

$$\tilde{\boldsymbol{\alpha}} = \begin{bmatrix}\alpha_1 + N_1 & \alpha_2 + N_2 & \dots & \alpha_{k-1} + N_{k-1} & \alpha_k + N_k\end{bmatrix}^\top$$

  where N_j is the count of category j as defined earlier.

  The parameter vector of the prior $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2 & \alpha_3 & \alpha_4\end{bmatrix}^\top$ for this specific scenario can be initialized using counts from Table 3. Recall that $\alpha_j - 1$ can be thought of as a pseudo-count of category j. However, here we have actual counts from a previous poll – even better. Finally, we have the parameters of the prior and the posterior as follows

$$\boldsymbol{\alpha} = \begin{bmatrix}251 & 111 & 281 & 141\end{bmatrix}^\top$$
$$\tilde{\boldsymbol{\alpha}} = \begin{bmatrix}276 & 123 & 311 & 151\end{bmatrix}^\top$$

- The posterior predictive distribution is given by:

$$p(\mathbf{x}^*|\mathcal{X}) = \int p(\mathbf{x}^*|\lambda_1,\ldots,\lambda_k)p(\lambda_1,\ldots,\lambda_k|\mathcal{X})d\boldsymbol{\lambda} \tag{70}$$

$$= \int \prod_{j=1}^{k} \lambda_j^{[\mathbf{x}_j^*=1]} \cdot \frac{1}{\mathrm{B}(\tilde{\boldsymbol{\alpha}})} \prod_{j=1}^{k} \lambda_j^{\tilde{\alpha}_j-1} d\boldsymbol{\lambda} \tag{71}$$

multiplying and dividing by $\frac{\prod_{j=1}^{k}\Gamma(\tilde{\alpha}_j+[\mathbf{x}_j^*=1])}{\Gamma\left(\sum_{j=1}^{k}\tilde{\alpha}_j+[\mathbf{x}_j^*=1]\right)}$

$$= \frac{1}{\mathrm{B}(\tilde{\boldsymbol{\alpha}})} \cdot \frac{\prod_{j=1}^{k}\Gamma(\tilde{\alpha}_j+[\mathbf{x}_j^*=1])}{\Gamma\left(\sum_{j=1}^{k}\tilde{\alpha}_j+[\mathbf{x}_j^*=1]\right)} \int \underbrace{\frac{1}{\frac{\prod_{j=1}^{k}\Gamma(\alpha_j+[\mathbf{x}_j^*=1])}{\Gamma\left(\sum_{j=1}^{k}\alpha_j+[\mathbf{x}_j^*=1]\right)}} \prod_{j=1}^{k} \lambda_j^{\tilde{\alpha}_j+[\mathbf{x}_j^*=1]-1} d\boldsymbol{\lambda}}_{=1 \because \text{ it's Dirichlet PDF}} \tag{72}$$

$$= \frac{\Gamma\left(\sum_{j=1}^{k}\tilde{\alpha}_j\right)}{\prod_{j=1}^{k}\Gamma(\tilde{\alpha}_j)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\tilde{\alpha}_j+[\mathbf{x}_j^*=1])}{\Gamma\left(\sum_{j=1}^{k}\tilde{\alpha}_j+[\mathbf{x}_j^*=1]\right)} \tag{73}$$

$$= \frac{\Gamma\left(\sum_{j=1}^{k}\tilde{\alpha}_j\right)}{\prod_{j=1}^{k}\Gamma(\tilde{\alpha}_j)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\tilde{\alpha}_j+[\mathbf{x}_j^*=1])}{\sum_{j=1}^{k}\tilde{\alpha}_j \cdot \Gamma\left(\sum_{j=1}^{k}\tilde{\alpha}_j\right)} \tag{74}$$

now assume that $\mathbf{x}^*$ has 1 at m-th element

$$= \frac{1}{\prod_{j=1}^{k}\Gamma(\tilde{\alpha}_j)} \cdot \frac{\Gamma(\tilde{\alpha}_m+1)\cdot\prod_{j=1,j\neq m}^{k}\Gamma(\tilde{\alpha}_j)}{\sum_{j=1}^{k}\tilde{\alpha}_j} \tag{75}$$

$$= \frac{1}{\prod_{j=1}^{k}\Gamma(\tilde{\alpha}_j)} \cdot \frac{\tilde{\alpha}_m\cdot\prod_{j=1}^{k}\Gamma(\tilde{\alpha}_j)}{\sum_{j=1}^{k}\tilde{\alpha}_j} \tag{76}$$

$$= \frac{\tilde{\alpha}_m}{\sum_{j=1}^{k}\tilde{\alpha}_j} \tag{77}$$

Finally,

$$p(\mathbf{x}^* = \begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix}^\top|\mathcal{X}) = \frac{311}{276+123+311+151} = 0.361$$

Problem 3.c. Choose an appropriate prior distribution for the likelihood distribution chosen in 2.a. and estimate the parameters using maximum a posteriori (MAP) estimation.

Solution: In the previous section we have seen that if we choose the Dirichlet distribution as the prior, the posterior is $\mathrm{Dir}(\tilde{\boldsymbol{\alpha}})$.

$$\mathrm{Dir}(\boldsymbol{\lambda}|\tilde{\boldsymbol{\alpha}}) \propto \prod_{j=1}^{k} \lambda_j^{\alpha_j+N_j-1}$$

Taking log probability and adding a lagrange multiplier for the constraint we get the auxiliary function

$$\mathcal{L} = \sum_{j=1}^{k} (\alpha_j + N_j - 1) \log \lambda_j + \nu \left(\sum_{j=1}^{k} \lambda_j - 1 \right) \tag{78}$$

Differentiating with respect to λ_j and proceeding as in MLE, we get

$$\lambda_j = \frac{\alpha_j + N_j - 1}{\sum_{l=1}^{k} \alpha_j + N_l - 1} \tag{79}$$

The estimated parameter values are tabulated in Table 4.

Parameters	**Values**
λ_1	$\frac{275}{275+122+310+150} = 0.321$
λ_2	0.142
λ_3	0.362
λ_4	0.175

Table 4: Parameters estimated using MAP

Problem 3.d. (*Challenge*) Show that the categorical distribution is exponential family.

Solution: The members of exponential family follow the following general form

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x}) \cdot \exp\{\boldsymbol{\eta}^\top \mathbf{s}(\mathbf{x}) - A(\boldsymbol{\eta})\} \tag{80}$$

The PMF of the categorical distribution is given by

$$p(\mathbf{x}|\lambda_1, \ldots, \lambda_k) = \prod_{j=1}^{k} \lambda_j^{x_j} \qquad \text{such that } \sum_{j=1}^{k} \lambda_j = 1 \tag{81}$$

Applying log and then exp (note that the function remains unchanged after this application):

$$p(\mathbf{x}|\lambda_1, \ldots, \lambda_k) = \exp\left\{ \sum_{j=1}^{k} x_j \log \lambda_j \right\} \tag{82}$$

$$= \exp\left\{ (\log \boldsymbol{\lambda})^\top \mathbf{x} - 0 \right\} \tag{83}$$

where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_k]^\top$.

Comparing Eq. (83) with Eq. (80) we see that

$$h(\mathbf{x}) = 1 \tag{84}$$

$$\boldsymbol{\eta} = \log \boldsymbol{\lambda} \quad \text{where } \sum_{j=1}^{k} e^{\eta_j} = 1 \tag{85}$$

$$\mathbf{s}(x) = \mathbf{x} \tag{86}$$

$$A(\boldsymbol{\eta}) = 0 \tag{87}$$

We have the inverse mapping $\lambda_i = e^{\eta_i}$ between the parameter vector and the natural parameter. Also, $\sum_{i=1}^{k} e^{\eta_i} = 1$.

Advanced: The form above is not quite right — it is a curved exponential family since there are actually only $k-1$ parameters but $\boldsymbol{\eta}$ lives in a k-dimensional real space. The constraint on $\boldsymbol{\lambda}$ makes $\boldsymbol{\eta}$ lie in a nonlinear subset of $\mathbb{R}^k$. The properties developed for exponential families do not often generalize to curved exponential familities (consider the gradient of $A(\eta)$). To fix this, we can define

$$\lambda_j = \frac{\exp(\eta_j)}{\sum_{l=1}^{k} \exp(\eta_l)} \tag{88}$$

so,

$$\log \lambda_j = \eta_j - \log \sum_{l=1}^{k} \exp(\eta_l) \tag{89}$$

and follow through with from (83) above to find that

$$p(\mathbf{x}|\lambda_1, \ldots, \lambda_k) = \exp\left\{ \boldsymbol{\eta}^\top \mathbf{x} - \log \sum_{l=1}^{k} \exp(\eta_l) \right\} \tag{90}$$

where $A(\boldsymbol{\eta}) = \log \sum_l^k \exp(\eta_l)$.