In the name of GOD.



**Sharif University of Technology**

# Machine Learning

**Spring 2025**

Hamid R. Rabiee, Zahra Dehghanian

| Homework 1 | Regression, Bias and Variance, ML/MAP | Deadline: 1403/12/10 |
|---|---|---|

1. [18] In the real world, we are often faced with problems where the data follows the regression function $h(x)$. That is, for any input value $x$, the corresponding label is $y = h(x) + \epsilon$, where $\epsilon$ represents noise. Using the given data, we create $M$ predictive models $y_1(x), y_2(x), \ldots, y_M(x)$. If $H_M(x)$ is defined as:

$$H_M(x) = \frac{1}{M} \sum_{i=1}^{M} y_i(x),$$

prove that:

$$\mathbb{E}_x \left[ (H_M(x) - h(x))^2 \right] \leq \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_x \left[ (y_i(x) - h(x))^2 \right].$$

hint:Cauchy-Schwarz!

Solution:
We define the quantities $E_{\text{com}}$ and $E_{\text{avg}}$ as follows:

$$E_{\text{com}} = \mathbb{E}_x \left[ (H_M(x) - h(x))^2 \right]$$

$$E_{\text{avg}} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_x \left[ (y_i(x) - h(x))^2 \right]$$

Using the definition $\epsilon_i(x) = y_i(x) - h(x)$, for $E_{\text{com}}$, we have:

$$\begin{aligned} E_{\text{com}} &= \mathbb{E}_x \left[ (H_M(x) - h(x))^2 \right] \\ &= \mathbb{E}_x \left[ \left( \left( \frac{1}{M} \sum_{i=1}^{M} y_i(x) \right) - h(x) \right)^2 \right] \\ &= \mathbb{E}_x \left[ \left( \frac{1}{M} \sum_{i=1}^{M} (y_i(x) - h(x)) \right)^2 \right] \\ &= \mathbb{E}_x \left[ \left( \frac{1}{M} \sum_{i=1}^{M} \epsilon_i(x) \right)^2 \right] \end{aligned}$$

$$= \frac{1}{M^2}\mathbb{E}_x \left[ \left( \sum_{i=1}^{M} \epsilon_i(x) \right)^2 \right].$$

Similarly, using the definition $\epsilon_i(x) = y_i(x) - h(x)$, for $E_{\text{avg}}$, we have:

$$E_{\text{avg}} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_x \left[ (y_i(x) - h(x))^2 \right]$$

$$= \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_x \left[ \epsilon_i^2(x) \right].$$

Using the Cauchy-Schwarz inequality, we have:

$$\left( \sum_{i=1}^{M} \epsilon_i(x) \right)^2 \leq M \sum_{i=1}^{M} \epsilon_i^2(x),$$

which gives:

$$\frac{1}{M^2}\mathbb{E}_x \left[ \left( \sum_{i=1}^{M} \epsilon_i(x) \right)^2 \right] \leq \frac{1}{M^2}\mathbb{E}_x \left[ M \sum_{i=1}^{M} \epsilon_i^2(x) \right].$$

Simplifying:

$$\frac{1}{M^2} M \sum_{i=1}^{M} \mathbb{E}_x \left[ \epsilon_i^2(x) \right] = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_x \left[ \epsilon_i^2(x) \right].$$

Thus:

$$E_{\text{com}} \leq E_{\text{avg}}.$$

2. [20] Consider a standard linear regression setting where there exists a true parameter vector $\theta^* \in \mathbb{R}^{d\times 1}$ that generates outputs $y_i \in \mathbb{R}$ from inputs $x_i \in \mathbb{R}^{d\times 1}$ as follows:

$$y_i = {\theta^*}^T x_i + \epsilon_i, \quad \text{where } \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Suppose we have fixed the inputs $x_1, \ldots, x_n$. We construct the matrix X by placing $x_i$ in the i-th row ($X \in \mathbb{R}^{n\times d}$). The training error for a given parameter $\theta$ can be written in matrix form as $\hat{R}(\theta) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\theta\|_2^2$. We define the statistical error of parameter $\theta$, denoted as $R(\theta)$, as the expectation of the training error over the randomness in $\mathbf{y}$, that is:

$$R(\theta) = \mathbb{E}_{\mathbf{y}}[\hat{R}(\theta)].$$

Since the design matrix $\mathbf{X}$ is fixed, the only source of randomness comes from $\epsilon$. Let $R^*$ be the statistical error corresponding to the true parameter $\theta^*$, which generates the outputs $\mathbf{y}$. Imagine that the matrix $\mathbf{X}$ is full rank, and the OLS estimator is written as $\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Prove the following:

(a) Show that

$$R(\theta) - R^* = \|\theta - \theta^*\|_{\hat{\Sigma}}^2,$$

where $\|\theta - \theta^*\|_{\hat{\Sigma}}^2 = (\theta - \theta^*)^T\hat{\Sigma}(\theta - \theta^*)$ and $\hat{\Sigma} = \frac{\mathbf{X}^T\mathbf{X}}{n}$ is the empirical covariance matrix.

(b) Prove that the OLS estimator is unbiased. Additionally, show that the variance of the OLS estimator $\hat{\theta}$ is:

$$\text{var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)^T] = \frac{\sigma^2}{n}\hat{\Sigma}^{-1}.$$

Solution:

(a) Here we denote $\epsilon$ as an $n$-dimensional vector, where each index $\epsilon_i$ represents the noise of each output $y_i$.

$$\begin{aligned} R(\theta) &= \mathbb{E}_y\left[\frac{1}{n}\|\mathbf{y} - \mathbf{X}\theta\|_2^2\right] = \mathbb{E}_\epsilon\left[\frac{1}{n}\|\mathbf{X}\theta^* + \epsilon - \mathbf{X}\theta\|_2^2\right] \\ &= \frac{1}{n}\mathbb{E}_\epsilon\left[\|\mathbf{X}(\theta^* - \theta)\|_2^2 + \|\epsilon\|_2^2 + 2(\mathbf{X}(\theta^* - \theta))^T\epsilon\right] \\ &= \sigma^2 + \frac{1}{n}(\theta - \theta^*)^T\mathbf{X}^T\mathbf{X}(\theta - \theta^*). \end{aligned}$$

Note that $\mathbb{E}[\|\epsilon\|_2^2] = \mathbb{E}\left[\sum_{i=1}^n \epsilon_i^2\right] = n\sigma^2$. Writing the above expression for $R^*$ results in only $\sigma^2$. Therefore, the expression:

$$R(\theta) - R^* = \frac{1}{n}(\theta - \theta^*)^T\mathbf{X}^T\mathbf{X}(\theta - \theta^*) = \|\theta - \theta^*\|_{\hat{\Sigma}}^2$$

(b)

$$\begin{aligned} \mathbb{E}[\hat{\theta}] &= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta^* + \epsilon)] \\ &= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\theta^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon] \\ &= \theta^*. \end{aligned}$$

Therefore, the estimator is unbiased. Now, we write the expression for variance. Note that

$$\mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\epsilon] = 0.$$

We showed that $\hat{\theta} - \theta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$. Now, we write:

$$\begin{aligned}\text{var}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)^T] \\ &= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}].\end{aligned}$$

Because the random noise for each output is uncorrelated, $\mathbb{E}[\epsilon\epsilon^T] = \sigma^2 I$. Also, since $\mathbf{X}^T\mathbf{X}$ is symmetric, its inverse is also symmetric, so $(\mathbf{X}^T\mathbf{X})^{-T} = (\mathbf{X}^T\mathbf{X})^{-1}$. Thus:

$$\text{var}(\hat{\theta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 = \frac{\sigma^2}{n}\hat{\Sigma}^{-1}.$$

Note that $\hat{\Sigma}^{-1} = n(\mathbf{X}^T\mathbf{X})^{-1}$.

3. [12] Consider a linear regression model where the cost function is given by:

$$J = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \mathbf{w}^T \mathbf{w}$$

where:

- $\mathbf{y}$ is the target vector,
- $\mathbf{X}$ is the feature matrix,
- $\mathbf{w}$ is the weight vector,
- $\lambda$ is the regularization parameter,
- $\|\cdot\|_2$ denotes the $L_2$ norm.

(a) Explain what $\lambda \mathbf{w}^T \mathbf{w}$ is and why we add it to the loss function. How does it work?

(b) Explain the importance of selecting an appropriate value for the regularization parameter $\lambda$. What happens if we choose $\lambda$ very large (i.e., $\lambda \to \infty$) or very small (i.e., $\lambda \to 0$)?

(c) Consider the general form of regularization:

$$J = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda r(\mathbf{w}),$$

where $r : \mathbb{R}^d \to \mathbb{R}$ is the regularization function. Suppose for two weight vectors $\mathbf{w}_1 = [1, 0, 6]^T$ and $\mathbf{w}_2 = [-2, 3, 3]^T$, the squared error term $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$ is equal for both. Explain which weight vector ($\mathbf{w}_1$ or $\mathbf{w}_2$) the model will prefer if we replace $r(\mathbf{w})$ with each of the following functions:

- $r(\mathbf{w}) = 1$,
- $r(\mathbf{w}) = \sum_{i=1}^{d} |w_i|$ ($L_1$ regularization),
- $r(\mathbf{w}) = \sum_{i=1}^{d} |w_i|^2$ ($L_2$ regularization),
- $r(\mathbf{w}) = \max(|w_1|, |w_2|, \ldots, |w_d|)$ ($L_\infty$ regularization).

(d) Suppose we are using **Stochastic Gradient Descent (SGD)** to optimize the cost function:

$$J = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \mathbf{w}^T \mathbf{w}.$$

Derive the update rule for the weight $w_i$, using learning rate $\eta$ and data point $(\mathbf{x}^{(j)}, y^{(j)})$ in SGD.

Solution:

(a) 1. $\lambda \mathbf{w}^T \mathbf{w}$ is the **L2 regularization term**, also known as **Ridge regularization**.
2. It is added to the loss function to:
   - Penalize large weights, preventing overfitting by discouraging the model from relying too heavily on any single feature.
   - Improve generalization by shrinking the weights toward zero, leading to a simpler model.

3. It works by adding a penalty proportional to the squared magnitude of the weights, which balances the trade-off between fitting the training data and keeping the model weights small.

(b) If $\lambda \to \infty$:

- The regularization term dominates the loss function.

- The model will prioritize minimizing $\mathbf{w}^T\mathbf{w}$, forcing all weights to approach zero.
- This results in an overly simplistic model (underfitting).

If $\lambda \to 0$:

- The regularization term becomes negligible.
- The model focuses solely on minimizing the squared error $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$.
- This may lead to overfitting if the model is complex.

(c) For $\mathbf{w}_1 = [1, 0, 6]^T$ and $\mathbf{w}_2 = [-2, 3, 3]^T$:

- $r(\mathbf{w}) = 1$:
  - The regularization term is constant for all $\mathbf{w}$.
  - The model has no preference between $\mathbf{w}_1$ and $\mathbf{w}_2$.
- $r(\mathbf{w}) = \sum |w_i|$:
  - Compute $r(\mathbf{w}_1) = |1| + |0| + |6| = 7$,
  - Compute $r(\mathbf{w}_2) = |-2| + |3| + |3| = 8$.
  - The model prefers $\mathbf{w}_1$ because it has a smaller $L1$ penalty.
- $r(\mathbf{w}) = \sum |w_i|^2$:
  - Compute $r(\mathbf{w}_1) = 1^2 + 0^2 + 6^2 = 37$,
  - Compute $r(\mathbf{w}_2) = (-2)^2 + 3^2 + 3^2 = 22$.
  - The model prefers $\mathbf{w}_2$ because it has a smaller $L2$ penalty.
- $r(\mathbf{w}) = \max(|w_1|, |w_2|, |w_3|)$:
  - Compute $r(\mathbf{w}_1) = \max(1, 0, 6) = 6$,
  - Compute $r(\mathbf{w}_2) = \max(2, 3, 3) = 3$.
  - The model prefers $\mathbf{w}_2$ because it has a smaller $L\infty$ penalty.

(d) The cost function for a single data point $(\mathbf{x}^{(j)}, y^{(j)})$ is:

$$J^{(j)} = (y^{(j)} - \mathbf{x}^{(j)}\mathbf{w})^2 + \lambda \mathbf{w}^T\mathbf{w}.$$

To derive the update rule for $w_i$, we compute the gradient of $J^{(j)}$ with respect to $w_i$:

$$\frac{\partial J^{(j)}}{\partial w_i} = \frac{\partial}{\partial w_i}\left((y^{(j)} - \mathbf{x}^{(j)}\mathbf{w})^2 + \lambda \mathbf{w}^T\mathbf{w}\right).$$

First, we compute the gradient of the squared error term:

$$\frac{\partial}{\partial w_i}((y^{(j)} - \mathbf{x}^{(j)}\mathbf{w})^2) = 2(y^{(j)} - \mathbf{x}^{(j)}\mathbf{w}) \cdot (-x_i^{(j)}) = -2x_i^{(j)}(y^{(j)} - \mathbf{x}^{(j)}\mathbf{w}).$$

The regularization term is:

$$\lambda \mathbf{w}^T\mathbf{w} = \lambda \sum_{k=1}^{d} w_k^2.$$

The gradient with respect to $w_i$ is:

$$\frac{\partial}{\partial w_i}\left(\lambda \sum_{k=1}^{d} w_k^2\right) = 2\lambda w_i.$$

The total gradient of $J^{(j)}$ with respect to $w_i$ is:

$$\frac{\partial J^{(j)}}{\partial w_i} = -2x_i^{(j)}(y^{(j)} - \mathbf{x}^{(j)}\mathbf{w}) + 2\lambda w_i.$$

Using the gradient descent update rule with learning rate $\eta$:

$$w_i \leftarrow w_i - \eta \frac{\partial J^{(j)}}{\partial w_i}.$$

Substitute the gradient:

$$w_i \leftarrow w_i - \eta \left( -2x_i^{(j)}(y^{(j)} - \mathbf{x}^{(j)}\mathbf{w}) + 2\lambda w_i \right).$$

Simplify:

$$w_i \leftarrow w_i + 2\eta x_i^{(j)}(y^{(j)} - \mathbf{x}^{(j)}\mathbf{w}) - 2\eta\lambda w_i.$$

4. [35] Answer the following quesitons.

   (a) Suppose our samples are $\mathbf{x} = (0, 0, 1, 1, 0)$, from $\text{Ber}(\theta)$, where $\theta$ is unknown. Assume $\theta$ is unrestricted; that is, $\theta \in (0, 1)$. What is the MLE for $\theta$?

   (b) Suppose we impose the restriction that $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for $\theta$?

   (c) Assume $\Theta$ is restricted as in part (b) (but now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for $\theta$?

   (d) Show that we can make the MAP whatever we like, by finding a prior over $\{0.2, 0.5, 0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

   (e) Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\in (0, 1)$, not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$. So we need a (continuous) prior distribution with range $(0, 1)$ instead of our discrete one. We assign $\Theta \sim \text{Beta}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density

   $$\pi_\Theta(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } \theta \in (0, 1).$$

   Recall the mode of a $W \sim \text{Beta}(\alpha, \beta)$ random variable is

   $$\frac{\alpha - 1}{(\alpha - 1) + (\beta - 1)}$$

   (the mode is the value with highest density $\arg\max_w f_W(w)$).

   Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is $\frac{k}{n}$, where $k = \sum x_i$ (the total number of successes). Show that the posterior $\pi_\Theta(\theta \mid x)$ is proportional to $\text{Beta}(k+\alpha, n-k+\beta)$ distribution, and find the MAP estimator.

   (f) Recall that $\text{Beta}(1, 1) \equiv \text{Unif}(0, 1)$ (pretend we saw $1 - 1$ heads and $1 - 1$ tails ahead of time). If we used this as the prior, how would the MLE and MAP compare?

   (g) Since the posterior is also a Beta Distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution's parameter $p$. Interpret $\alpha, \beta$ as to how they affect our estimate. This is a really special property: if the prior distribution multiplied by the likelihood results in a posterior distribution in the same family (with different parameters), then we say that distribution is the conjugate prior to the distribution we are estimating.

   (h) As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our **prior** when $n$ is small, or $n$ is large?

   (i) Which do you think is "better", MLE or MAP?

Solution:

   (a) Suppose our samples are $\mathbf{x} = (0, 0, 1, 1, 0)$, from $\text{Ber}(\theta)$, where $\theta$ is unknown. Assume $\theta$ is unrestricted; that is, $\theta \in (0, 1)$. What is the MLE for $\theta$?

   - **Answer:** $\frac{2}{5}$. We just find the likelihood of the data, which is the probability of observing 2 heads and 3 tails, and find the $\theta$ that maximizes it.

   $$L(\mathbf{x} \mid \theta) = \theta^2(1-\theta)^3$$

   $$\hat{\theta}_{MLE} = \arg\max_{\theta \in [0,1]} \theta^2(1-\theta)^3 = \frac{2}{5}$$

(b) Suppose we impose the restriction that $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for $\theta$?

- **Answer:** 0.5. We need to find which of the three acceptable $\theta$ values maximizes the likelihood, and since there are only finitely many, we can just plug them all in and compare!

$$L(\mathbf{x} \mid 0.2) = (0.2)^2(0.8)^3 = 0.02048$$

$$L(\mathbf{x} \mid 0.5) = (0.5)^2(0.5)^3 = 0.03125$$

$$L(\mathbf{x} \mid 0.7) = (0.7)^2(0.3)^3 = 0.01323$$

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \{0.2, 0.5, 0.7\}} L(\mathbf{x} \mid \theta) = 0.5$$

(c) Assume $\Theta$ is restricted as in part (b) (but now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1, \pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for $\theta$?

- **Answer:** 0.7. Instead of maximizing just the likelihood, we need to maximize the likelihood times the prior. Again, since there are only finitely many values, we just plug them in!

$$\pi_\Theta(0.2 \mid x) = L(\mathbf{x} \mid 0.2)\pi_\Theta(0.2) = (0.2)^2(0.8)^3(0.1) = 0.0020480$$

$$\pi_\Theta(0.5 \mid x) = L(\mathbf{x} \mid 0.5)\pi_\Theta(0.5) = (0.5)^2(0.5)^3(0.01) = 0.0003125$$

$$\pi_\Theta(0.7 \mid x) = L(\mathbf{x} \mid 0.7)\pi_\Theta(0.7) = (0.7)^2(0.3)^3(0.89) = 0.0117747$$

Note the effect of this prior - by setting $\pi_\Theta(0.7)$ so high and the other two values, we actually get a different maximizer. This is the effect of the prior on the MAP estimate (which was completely arbitrary)!

(d) Show that we can make the MAP whatever we like, by finding a prior over $\{0.2, 0.5, 0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

- **Answer:** Choose $\pi_\Theta(\theta)$ to fit the $\theta$ you want! This shows that the prior really does make a difference, and that MAP and MLE are indeed different techniques.

(e) Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\theta \in (0, 1)$, not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$. So we need a (continuous) prior distribution with range $(0, 1)$ instead of our discrete one. We assign $\Theta \sim \text{Beta}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density

$$\pi_\Theta(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } \theta \in (0, 1).$$

Recall the mode of a $W \sim \text{Beta}(\alpha, \beta)$ random variable is

$$\frac{\alpha - 1}{\alpha - 1 + \beta - 1}$$

(the mode is the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is $\frac{k}{n}$, where $k = \sum x_i$ (the total number of successes). Show that the posterior $\pi_\Theta(\theta \mid x)$ is proportional to Beta$(k+\alpha, n-k+\beta)$ distribution, and find the MAP estimator.

- **Answer:** $\hat{\theta}_{MAP} = \frac{k+\alpha-1}{n+(\alpha-1)+(\beta-1)}$. We first have to write out what the posterior distribution is, which is proportional to just the prior times the likelihood:

$$\pi_\Theta(\theta \mid x) \propto L(x \mid \theta) \cdot \pi_\Theta(\theta)$$

$$= \left(\binom{n}{k}\theta^k(1-\theta)^{n-k}\right) \cdot \left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$\propto \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$$

The first to second line comes from noticing $L(x \mid \theta)$ is just the probability of seeing exactly $k$ successes out of $n$ (binomial PMF), and plugging in our equation for $\pi_\Theta$ (beta density). The second to third line comes from dropping the normalizing constants (that don't depend on $\theta$), which we can do because we only care to maximize this over $\theta$. If you stare closely at that last equation, it actually proportional to the PDF of a Beta distribution with different parameters!

Our posterior is hence Beta$(k+\alpha, n-k+\beta)$ since PDFs uniquely define a distribution (there is only one normalizing constant that would make it integrate to 1). The MAP estimator is the mode of this posterior Beta distribution, which is given by the formula:

$$\hat{\theta}_{MAP} = \frac{k+\alpha-1}{(k+\alpha-1)+(n-k+\beta-1)} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$$

Try staring at this to see why this might make sense. We'll explain it more in part (g)!

(f) Recall that Beta$(1,1) \equiv$ Unif$(0,1)$ (pretend we saw $1-1$ heads and $1-1$ tails ahead of time). If we used this as the prior, how would the MLE and MAP compare?

- **Answer:** They would be the same! From our previous question, if $\alpha = \beta = 1$, then

$$\hat{\theta}_{MAP} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

This is because we don't have any prior information essentially, by saying each value is equally likely!

(g) Since the posterior is also a Beta Distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution's parameter $p$. Interpret $\alpha, \beta$ as to how they affect our estimate. This is a really special property: if the prior distribution multiplied by the likelihood results in a posterior distribution in the same family (with different parameters), then we say that distribution is the conjugate prior to the distribution we are estimating.

- **Answer:** The interpretation is: pretend we saw $\alpha - 1$ heads ahead of time, and $\beta - 1$ tails ahead of time. Then our total number of heads is $k + (\alpha - 1)$ (real + fake) and our total number of trials is $n + (\alpha + \beta - 2)$ (real + fake), so that's our estimate! That's how prior information was factored in to our estimator, rather than just using what we actually saw in the data.

(h) As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our **prior** when $n$ is small, or $n$ is large?

- **Answer:** They become equal! The prior is important if we don't have much data, but as we get more, the evidence overwhelms the prior. You can imagine that if we only flipped the coin 5 times, the prior would play a huge role in our estimate. But if we flipped the coin 10,000 times, any (small) prior wouldn't really change our estimate.

(i) Which do you think is "better", MLE or MAP?

- **Answer:** There is no right answer. There are two main schools in statistics: Bayesians and Frequentists.
- **Frequentists prefer MLE** since they don't believe you should be putting a prior belief on anything, and you should only make judgment based on what you've seen. They believe the parameter being estimated is a **fixed quantity**.
- **On the other hand, Bayesians prefer MAP**, since they can incorporate their prior knowledge into the estimation. Hence the parameter being estimated is a **random variable**, and we seek the mode - the value with the highest probability or density. An example would be estimating the probability of heads of a coin - is it reasonable to assume it is more likely fair than not? If so, what distribution should we put on the parameter space?
- **Anyway, in the long run, the prior "washes out"**, and the only thing that matters is the likelihood; the observed data. For small sample sizes like this, the prior significantly influences the MAP estimate. However, as the number of samples goes to infinity, the MAP and MLE are equal.

(a) [15] Complete the attached notebook. You are permitted to use chatbots or other resources for assistance, but you must ensure that you fully understand the code and implementations. During the online sessions, you may be asked to explain specific functions, code lines, and your overall approach. Be prepared to demonstrate your understanding in detail.