# EE5907/EE5027 Week 2: Probabilistic Estimation + Conjugate Priors

The following questions are from Kevin Murphy's (KM) book "Machine Learning: A Probabilistic Perspective".

### Exercise 3.1 MLE for the Bernoulli/ binomial model

Derive

$$\hat{\theta}_{MLE} = \frac{N1}{N} \tag{1}$$

by optimizing the log of the likelihood in Eq. (2)

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{2}$$

### Exercise 3.6 MLE for the Poisson distribution

The Poisson pmf is defined as $\text{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, for $x \in \{0, 1, 2, \cdots\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE.

### Exercise 3.7 Bayesian analysis of the Poisson distribution

In the previous exercise, we defined the Poisson distribution with rate $\lambda$ and derived its MLE. Here we perform a conjugate Bayesian analysis.

a. Derive the posterior $p(\lambda|\mathcal{D})$ assuming a conjugate prior $p(\lambda) = Ga(\lambda|a, b) \propto \lambda^{a-1}e^{-\lambda b}$. Hint: the posterior is also a Gamma distribution.

b. What does the posterior mean tend to as $a \to 0$ and $b \to 0$? (Recall that the mean of a $\text{Ga}(a, b)$ distribution is $a/b$.)

## Exercise 3.12 MAP estimation for the Bernouli with non-conjugate Priors

We discussed Bayesian inference of a Bernoulli rate parameter with the prior $p(\theta) = Beta(\theta|\alpha, \beta)$. We know that, with this prior, the MAP estimate is given by

$$\hat{\theta} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2} \tag{3}$$

where $N_1$ is the number of heads, $N_0$ is the number of tails, and $N = N_0 + N_1$ is the total number of trials.

a. Now consider the following prior, that believes the coin is fair, or is slightly biased towards tails:

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Derive the MAP estimate under the prior as a function of $N_1$ and $N$.

b Suppose the true parameter is $\theta = 0.41$. Which prior leads to a better estimate when $N$ is small? Which leads to a better estimate when $N$ is large?

## Exercise 3.14 Posterior predictive for Dirichlet-multinomial

a. Suppose we compute the empirical distribution over letters of the Roman alphabet plus the space character (a distribution over 27 values) from 2000 samples. Suppose we see the letter "e" 260 times. What is $p(x_{2001} = e| \mathcal{D})$, if we assume $\theta \sim \text{Dir}(\alpha_1, \cdots, \alpha_{27})$, where $\alpha_k = 10$ for all $k$?

b. Suppose, in the 2000 samples, we saw "e" 260 times, "a" 100 times, and "p" 87 times. What is $p(x_{2001} = p, x_{2002} = a|\mathcal{D})$, if we assume $\theta \sim \text{Dir}(\alpha_1, \cdots, \alpha_{27})$, where $\alpha_k = 10$ for all $k$? Show your work.