CO902
Solutions to Problem Set 3

1. *Bounding a classification error rate.* Note, this problem was inconsistent: It defined X_i in terms of correct classifications but then asked for the error rate. It is much easier (and what I intended) to have $X_i = 1$ correspond to an *misclassification*, so $\mathbb{E}(X_i) = \theta$ is the error rate.

   (a) To use Chebyshev's inequality we need the mean and variance of the random variable of interest, here, $\hat{\theta}_{\text{MLE}}$. From lecture, we know that the MLE of a sample of n iid Bernoulli's is $\hat{\theta}_{\text{MLE}} = n_1/n$ where n_1 is the number of successes, and (from lecture and the next problem) $\mathbb{E}(\hat{\theta}_{\text{MLE}}) = \theta$ and $\mathbb{V}(\hat{\theta}_{\text{MLE}}) = \theta(1-\theta)/n$.

   Of course, since the "true error rate" is θ, we're Chebyshev's inequality immediately gives the result we want:

$$P(|\hat{\theta}_{\text{MLE}} - \theta| \geq a) \leq \frac{\theta(1-\theta)/n}{a^2}$$

   (b) For $\theta = 0.85$ and $a = 0.085$, the table below gives the bounds on the prediction error accuracy.

n	10	100	1000
$P(\lvert\hat{\theta}_{\text{MLE}} - \theta\rvert \geq 0.085) \leq \cdots$	1.765	0.176	0.017

   The bound is useless for $n = 10$ (probability bounded above one!), but for 100 says the probability of the error being more than 10% away from 0.85 is 0.176. Not great, but OK; with 1000, we can be fairly confident that the estimated accuracy is close to the true accuracy (though see next part).

   (c) The independence assumption is dodgy. Leave one out cross validation (LOOCV) or k-fold cross validation both produce predictions for held out data. However, the estimator for each held-out observation is highly dependent with other held-out predictions. To see this, consider LOOCV: For observation $i = 1$, the predictor is (reverting back to usual $\{X_i, Y_i\}$ notation) $\hat{Y}_{-1}(X_1) = f(X_1, \{X_i, Y_i\}_{i=2,3,\ldots,n})$, while for $i = 2$ the predictor is $\hat{Y}_{-2}(X_2) = f(X_2, \{X_i, Y_i\}_{i=1,3,\ldots,n})$, and thus there is a huge overlap in the information in $\hat{Y}_{-1}(X_1)$ and $\hat{Y}_{-1}(X_1)$ and thus correlated.

   Note this correlation doesn't corrupt the unbiasedness, because the expectation of a sum of predictions is the sum of expected predictions, regardless of correlation. *Variance* computations, on the other hand, are made hugely difficult by this correlation. In fact, getting good estimated variability of accuracy estimates is notriously hard if not impossible[1].

[1] See Bengio & Grandvalet. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *J Mach Learn Res*, 5, 1089-1105.)

2. *Bernoulli MAP properties.* If $X_i \sim \mathrm{Ber}(\theta)$, iid, $i = 1, \ldots, n$, then the MLE is $\hat{\theta}_{\mathrm{MLE}} = n_1/n$, where $n_1 = \sum_{i=1}^n X_i$. If $\theta \sim \mathrm{Beta}(\alpha, \beta)$ a priori, then

$$\hat{\theta}_{\mathrm{MAP}} = \frac{n_1 + \alpha - 1}{n + \alpha + \beta - 2}.$$

(a)

$$\begin{aligned}
\mathbb{E}(\hat{\theta}_{\mathrm{MLE}}) &= \mathbb{E}(n_1/n) \\
&= \frac{1}{n}\mathbb{E}(\left(\sum_{i=1}^n X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) \\
&= \frac{1}{n}\sum_{i=1}^n \theta = \theta
\end{aligned}$$

That is, $\hat{\theta}_{\mathrm{MLE}}$ is unbiased.

(b) The bias of the MAP[2] is

$$\begin{aligned}
\mathbb{E}(\hat{\theta}_{\mathrm{MAP}}) &= \mathbb{E}\left(\frac{n_1 + \alpha - 1}{n + \alpha + \beta - 2}\right) \\
&= \frac{\mathbb{E}(n_1) + \alpha - 1}{n + \alpha + \beta - 2} \\
&= \frac{n\theta + \alpha - 1}{n + \alpha + \beta - 2}
\end{aligned}$$

because $\mathbb{E}(n_1) = \mathbb{E}(\sum_i X_k) = n\theta$. But this is not equal to θ in general and hence $\hat{\theta}_{\mathrm{MAP}}$ is biased. Sufficient conditions for consistency[3] are bias *and* variance that converges to zero with n.

First, it is easy to show the bias of the MAP goes to zero as n grows

$$\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_{\mathrm{MAP}} - \theta) = \lim_{n\to\infty} \frac{\theta + (\alpha - 1)/n}{1 + (\alpha + \beta - 2)/n} - \theta = 0.$$

[2] To be precise, bias is a frequentist computation based on conditioning on a specific value of the random parameter (i.e. like we always do in a frequentist setting). So I write $\mathbb{E}(\hat{\theta}_{\mathrm{MAP}})$ but a Bayesian would insist on writing $\mathbb{E}(\hat{\theta}_{\mathrm{MAP}}|\theta)$ and you find some authors write $\mathbb{E}_\theta(\hat{\theta}_{\mathrm{MAP}})$, all in attempts to make it clear that we're *not* taking expectation w.r.t. the joint density of (X, θ), but just X for a fixed value of θ.

[3] The statement in the class notes, requiring finite sample unbiasesedness, was unnecessarily restrictive; see Casella & Berger, Theorem 10.1.3.

For the variance,

$$\begin{aligned}\mathbb{V}(\hat{\theta}_{\text{MAP}}) &= \mathbb{V}\left(\frac{n_1+\alpha-1}{n+\alpha+\beta-2}\right)\\ &= \frac{\mathbb{V}(n_1)}{(n+\alpha+\beta-2)^2}\\ &= \frac{n\theta(1-\theta)}{(n+\alpha+\beta-2)^2}.\end{aligned}$$

The derivative of the numerator w.r.t. n is constant and the derivative of the denominator is linear in n, and hence $\lim_{n\to\infty}\mathbb{V}(\hat{\theta}_{\text{MAP}}) = 0$ as well. Hence, the MAP is consistent.

3. *Bayes for Gaussian random variables.* The posterior is

$$\begin{aligned}p(\theta|X_1,\ldots,X_n) &\propto p(X_1,\ldots,X_n|\theta)p(\theta)\\ &= \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2/\sigma^2\right)\times\\ &\quad \frac{1}{(2\pi)^{1/2}b}\exp\left(-\frac{1}{2}(\theta-a)^2/b^2\right)\end{aligned}$$

Now, dropping further constants and using the notation provided in the answer, we write

$$\begin{aligned}p(\theta|X_1,\ldots,X_n) &\propto \exp\left(-\frac{1}{2}\left[\frac{1}{n\mathsf{se}^2}\sum_{i=1}^{n}(X_i-\theta)^2+\frac{1}{b^2}(\theta-a)^2\right]\right)\\ &= \exp\left(-\frac{1}{2}\left[\frac{1}{n\mathsf{se}^2}\sum_{i=1}^{n}X_i-\frac{2}{n\mathsf{se}^2}\theta\sum_{i=1}^{n}X_i+\frac{n}{n\mathsf{se}^2}\theta^2+\frac{1}{b^2}\theta^2-\frac{2}{b^2}\theta a+\frac{1}{b^2}a^2\right]\right)\\ &\propto \exp\left(-\frac{1}{2}\left[-\frac{2}{\mathsf{se}^2}\theta\bar{X}+\frac{1}{\mathsf{se}^2}\theta^2+\frac{1}{b^2}\theta^2-\frac{2}{b^2}\theta a\right]\right)\\ &= \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\mathsf{se}^2}+\frac{1}{b^2}\right)\theta^2-2\left(\frac{\bar{X}}{\mathsf{se}^2}+\frac{a}{b^2}\right)\theta\right]\right)\end{aligned}$$

where we've continued to drop constants and collect in terms of a polynomial in θ. Completing the square says that a polynomial of form Ax^2+Bx+C can be converted to one in the form of $A(x-H)^2+K$, if you choose $H=-B/(2A)$. So, here, "H" is

$$-\frac{-2\left(\frac{\bar{X}}{\mathsf{se}^2}+\frac{a}{b^2}\right)}{2\left(\frac{1}{\mathsf{se}^2}+\frac{1}{b^2}\right)} = w\bar{X}+(1-w)a = \bar{\theta}.$$

Again, as we can freely add and lose constants that don't dependend on θ, we have

$$p(\theta|X_1, \ldots, X_n) \propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau^2}(\theta - \bar{\theta})^2\right]\right).$$

Seeing that this is the kernel of a Gaussian distribution, we then know that it must be that the posterior of θ is $N(\bar{\theta}, \tau^2)$.

The crucial observation is that $0 \leq w \leq 1$ and so the posterior mean is a convex combination of the data mean $\bar{X}$ and prior mean a; the greater precision of the prior, the close the posterior mean is to a, the more data (or smaller σ) the closer the posterior mean is to $\bar{X}$.

4. *Iterated Expectation & Variance.*

$$\begin{aligned}
\mathbb{E}_Y\left(\mathbb{E}_{X|Y}(X|Y)\right) &= \int \mathbb{E}(X|Y)p_Y(y)dy \\
&= \int \left(\int x p_{X|Y}(x|y)dx\right) p_Y(y)dy \\
&= \int\int x p_{X,Y}(x,y)dxdy \\
&= \int x \left(\int p_{X,Y}(x,y)dy\right) dx \\
&= \int x p_X(x)dx \\
&= \mathbb{E}(X)
\end{aligned}$$

The result for variance is easier to see in the other direction

$$\begin{aligned}
\mathbb{V}_X(X) &= \mathbb{E}_X\left[(X - \mathbb{E}(X))^2\right] \\
&= \mathbb{E}_{XY}\left[(X - \mathbb{E}(X))^2\right] \\
&= \mathbb{E}_{XY}\left[(X - \mathbb{E}(X|Y) + \mathbb{E}(X|Y) - \mathbb{E}(X))^2\right] \\
&= \mathbb{E}_{XY}\left[(X - \mathbb{E}(X|Y))^2\right] + \mathbb{E}_{XY}\left[(\mathbb{E}(X|Y) - \mathbb{E}(X))^2\right] \\
&\quad + 2\mathbb{E}_{XY}\left[(X - \mathbb{E}(X|Y)(\mathbb{E}(X|Y) - \mathbb{E}(X))\right], \qquad (1)
\end{aligned}$$

where replacing $\mathbb{E}_X$ with $\mathbb{E}_{XY}$ is an application of the law of total probability ("sum rule"):

$$\mathbb{E}_X(f(X)) = \int f(x)p_X(x)dx = \int f(x)\int p_{XY}(x,y)dxdy = \mathbb{E}_{XY}(f(X)).$$

For the first term of Eqn. (1), the definition of conditional probability allows us to replace $\mathbb{E}_{XY}$ with $\mathbb{E}_Y\mathbb{E}_{X|Y}$,

$$\begin{aligned}
\mathbb{E}_{XY}\left[(X - \mathbb{E}(X|Y))^2\right] &= \mathbb{E}_Y\left[\mathbb{E}_{X|Y}(X - \mathbb{E}(X|Y))^2\right] \\
&= \mathbb{E}_Y\left[\mathbb{V}(X|Y)\right]
\end{aligned}$$

For the second term of Eqn. (1), apply the result for iterated expectation and note that $\mathbb{E}(X|Y)$ does not depend on X,

$$\begin{aligned}
\mathbb{E}_{XY}\left[(\mathbb{E}(X|Y)-\mathbb{E}(X))^2\right] &= \mathbb{E}_{XY}\left[(\mathbb{E}(X|Y)-\mathbb{E}(\mathbb{E}(X|Y)))^2\right] \\
&= \mathbb{E}_{Y}\left[(\mathbb{E}(X|Y)-\mathbb{E}(\mathbb{E}(X|Y)))^2\right] \\
&= \mathbb{V}_{Y}\left[\mathbb{E}(X|Y)\right]
\end{aligned}$$

Now, note that the last term of Eqn. (1),

$$\begin{aligned}
\mathbb{E}_{XY}\left[(X-\mathbb{E}(X|Y)(\mathbb{E}(X|Y)-\mathbb{E}(X))\right] &= \mathbb{E}_{XY}\left[(X-\mathbb{E}(X|Y)\mathbb{E}(X|Y)\right] \\
&\quad - \mathbb{E}_{XY}\left[(X-\mathbb{E}(X|Y))\mathbb{E}(X))\right]
\end{aligned}$$

is zero, seen by taking each summand in turn:

$$\begin{aligned}
\mathbb{E}_{XY}\left[(X-\mathbb{E}(X|Y)\mathbb{E}(X|Y)\right] &= \mathbb{E}_{Y}\left[\mathbb{E}_{X|Y}\left[(X-\mathbb{E}(X|Y))\,\mathbb{E}(X|Y)\right]\right] \\
&= \mathbb{E}_{Y}\left[\big(\mathbb{E}_{X|Y}[X|Y]-\mathbb{E}(X|Y)\big)\,\mathbb{E}(X|Y)\right] \\
&= \mathbb{E}_{Y}\left[(0)\,\mathbb{E}(X|Y)\right] = 0
\end{aligned}$$

For the second term

$$\begin{aligned}
\mathbb{E}_{XY}\left[(X-\mathbb{E}(X|Y))\mathbb{E}(X))\right] &= \mathbb{E}_{Y}\left[\mathbb{E}_{X|Y}\left[(X-\mathbb{E}(X|Y))\,\mathbb{E}(X))\right]\right] \\
&= \mathbb{E}_{Y}\left[\big(\mathbb{E}_{X|Y}[X|Y]-\mathbb{E}(X|Y)\big)\,\mathbb{E}(X))\right] \\
&= \mathbb{E}_{Y}\left[(0)\,\mathbb{E}(X))\right] = 0.
\end{aligned}$$