Informal Notes for Mathematics for Machine Learning
Imperial College London

Lecture 8: Bayesian Inference

Lecturer: Matthew Wicker

1 Learning Objectives

In previous lectures we looked at a simple but powerful learning framework: maximum likelihood estimation. We also briefly observed how a probabilistic perspective can motivate key learning concepts such as regularization which we will dive into in more detail in future lectures. Now that we have had a primer on probability and multivariate probability, we are ready to introduce Bayesian inference as a robust mathematical framework that can help us reason about machine learning models and their outputs.

2 Bayes Theorem

In our lecture slides, we start by showing how sometimes just thinking about likelihoods can be misleading. We gave the example of having a bird described as "white with an orange beak" which seems like it describes seagulls far better than pigeons (because it does). However, when you consider the prior (to observing any data) probability of a randomly selected bird being a pigeon, it greatly outweighs that of seagulls. So much so, that according to Bayes theorem, stated formally below, we should still believe that the bird is more likely to be a pigeon than a seagull. This highlights a key principle of Bayes: new evidence should not determine belief in a vacuum, it should update our prior beliefs. In addition, in any prediction we make we should express our uncertainty. We can express our uncertainties about these and other quantities by taking a Bayesian approach to learning. Recall that Bayes theorem gives us the equation $P(A|B) = P(B|A)P(A)/P(B)$. To see exactly how this applies to the learning we have been doing, let us substitute the quantities A and B for familiar quantities we have seen before in this course:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

The distribution $P(\theta|\mathcal{D})$ is termed the *posterior* distribution as it quantifies the probability of a parameter θ *after* we have observed our dataset. While we will discuss in detail the usefulness of this distribution below, let us take a second to highlight the philosophical differences between Bayesian learning and frequentist learning. In general, the perspective of frequentist inference is that the observations we have are noisily drawn from a process (that we model) with a fixed, true parameter and the task of inference is to make our best guess at the fixed parameter that produced

the data. On the other hand, the Bayesian perspective takes the data as certain events that are non-probabilistic and models our probabilistic beliefs, captured in a random variable over our parameter space, about this data.

2.1 Motivating Bayes Theorem

Let us return to our bias coin estimation procedure, this time in the context of placing an $\alpha > 0$ pound bet on the outcome of the next coin flip. Recall that we model the result of a coin as a Bernoulli random variable:

$$p(X = x) = \begin{cases} \theta & \text{if } x = 1 \text{ (heads)} \\ (1 - \theta) & \text{if } x = 0 \text{ (tails)} \end{cases}$$

And in Lecture 5, we observed that the MLE (a frequentist quantity) is given by $\theta^{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$. Now, assume that the data we have observed is $\mathcal{D} = \{1, 1, 1, 1\}$. It is clear, then, that $\theta^{\text{MLE}} = 1$. Thus, according to our best-fit of the data, we can select α (the amount of our wager) to be as high as we want with absolutely no risk, because we have estimated the probability of tails to be 0.[1] To many, this seems like an absurd conclusion that is the result of an overly simplistic mathematical model. I think very few would wager any considerable sum of money on the next coin flip being heads. This is due to our prior knowledge that designing a two-sided coin with probability 1 of landing on heads would be physically very challenging.[2] Before seeing any data, we would all make the rational assumption that the coin's probability of landing on heads is 0.5. Yet, this does not appear anywhere in our MLE model. In the Bayesian case, this quantity is defined as $p(\theta)$, the prior probability of observing heads before we have any data.

2.2 Selecting the prior distribution

But how do we select our prior probability? This question is a hotly debated topic. In general there are two perspectives. First is the subjective Bayesian perspective that the prior is an opportunity for us to encode our prior knowledge into our model either through experience with the world or possibly from the results of previous experiments/data. On the other hand, the objective Bayesian perspective is that priors should have as little effect on the inference as possible and exist primarily as a means by which to capture the uncertainty in our inference. Though both have their merits, we will begin our study of Bayesian inference with a more subjective Bayesian approach to selecting a prior distribution.

Firstly, let us make one simplifying assumption that is often done in Bayesian modelling which is removing the denominator also known as the model evidence. Though we will discuss this

[1] Note that a proper frequentist would reason about our estimated parameter using concentration inequalities or p-values, which we will cover later in the course, however, this sort of things is not traditionally done in machine learning settings and so this remains a valid hypothetical example.

[2] In fact, there is some debate as to weather or not a biased coin could be manufactured in the first place, https://www.stat.berkeley.edu/ nolan/Papers/dice.pdf.

quantity later, it will be analytically convenient for us to remove it here. This gives us the following formula:

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

Now, consider the fact that in order to compute our posterior probability, we will need to take the product of two densities. Thus, it may be analytically convenient to select a prior distribution whose density not only allows us to express our prior knowledge, but also one that allows us to easily take the product of our prior and likelihood distributions to arrive at our posterior.

2.2.1 Conjugate Prior for Bernoulli Variables

A conjugate prior, is a prior distribution whose form is such that when multiplied by the likelihood yields a distribution in the same form. Sometimes it is also used to simply describe a prior whose product with the likelihood gives us a closed-form posterior distribution. Recall that for a Bernoulli distribution, parameter θ the likelihood of observation x:

$$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

Under the i.i.d. assumption, then, the probability of our dataset will a product of powers of θ and $(1-\theta)$ with the powers growing as we observe more data. For example, the liklihood of a parameter θ given the observations in our initial example would be:

$$p(\mathcal{D}|\theta) = \theta^4(1-\theta)^0$$

Thus, the maximizing assignment of the likelihood (the maximum likelihood estimate) would be $\theta = 1$, as we discussed. Given this is our likelihood, it is intuitive that we would want to also express our prior as a product of powers. In fact, we have seen such a distribution, the Beta distribution whose unnormalized pdf is:

$$p_{\text{beta}}(x) = \text{const.} \cdot x^{\alpha-1}(1-x)^{\beta-1}$$

Importantly, then, if we must understand how the pdf of the Beta distribution changes for different selections of α and β. When do these values represent our beliefs about the coin before we see any data? This is a critical question for the subjective Bayesian to answer. One potentially sensible choice is to select $\alpha = 2$ and $\beta = 2$ which has mode/mean at 0.5 and puts prior density away from the unlikely cases of $\theta = 1$ and $\theta = 0$. Now, if this is taken to be our prior, let us observe how we arrive at a posterior. Recall:

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

First, lets plug in our beta prior density and then values:

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= p(\mathcal{D}|\theta)\theta^1(1-\theta)^1 \end{aligned}$$

Now, let us write out the form of our likelihood and then add in our dataset from before:

$$
\begin{aligned}
p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)\theta^1(1-\theta)^1 \\
&= \left(\prod_{x^{(i)}\in\mathcal{D}} (\theta)^{x^{(i)}}(1-\theta)^{x^{(i)}}\right)\theta^1(1-\theta)^1 \\
&= \left((\theta)^{\sum_{x^{(i)}\in\mathcal{D}} x^{(i)}}(1-\theta)^{n-\sum_{x^{(i)}\in\mathcal{D}} x^{(i)}}\right)\theta^1(1-\theta)^1 \\
&= (\theta)^{1+\sum_{x^{(i)}\in\mathcal{D}} x^{(i)}}(1-\theta)^{1+\sum_{x^{(i)}\in\mathcal{D}} x^{(i)}} \\
&= (\theta)^{1+4}(1-\theta)^{1+0}
\end{aligned}
$$

So, at the end, we have that our posterior is simply a beta distribution with parameter 5 and 1. Lets now plot (Figure 1) this and make some observations. Firstly, notice that now our posterior mean is 0.83 rather than 1. Moreover, we have some clear uncertainty about this. For example, we can see that we assign roughly 10% chance that the true probability is less than 0.6. Its unclear at first glance how to reason about these uncertainties in the frequentist case.

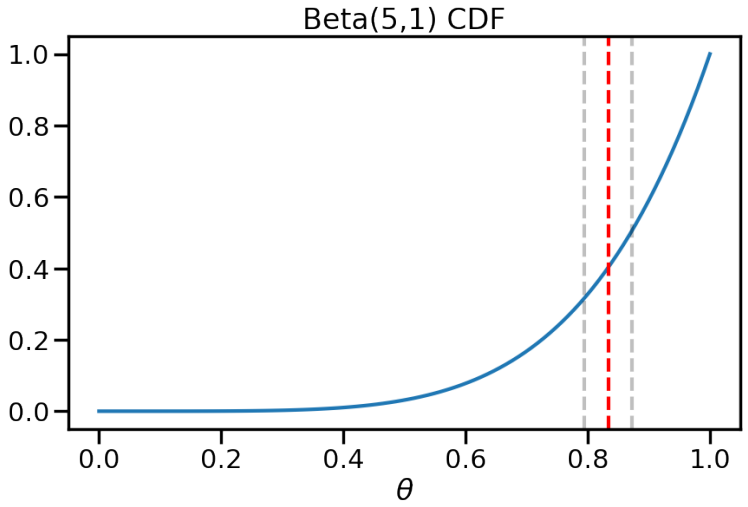

Figure 1: Our beta posterior distribution

Of course, we can observe the mean and variance in this particular case because we have plotted it, but it is useful to have a more general result. Let us now consider computing the mean and variance of our posterior distribution. Notice how we have the posterior but proportionally, $p(\theta|\mathcal{D}) \propto (\theta)^5(1-\theta)^1$. One needs to compute the normalizing constant for this distribution. In the above case, we can perform the process known as *equating coefficients* which is where we know (e.g., via conjugacy) that our posterior is proportional to a known probability density. In the above, we observe that the posterior above is the beta distribution with $\alpha = 6$ and $\beta = 2$. Thus, the normalizing constant is known in general to be: $\left(\int_x (x)^{\alpha-1}(1-x)^{\beta-1}dx\right)^{-1}$, which is $\left(\int_x (x)^{\alpha-1}(1-x)^{\beta-1}dx\right)^{-1} = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$. We also know the closed form mean and variance of the

distribution:

$$\mathbb{E}_\theta[p(\theta;\alpha,\beta)] = \frac{\alpha}{\alpha+\beta}$$

$$\text{Var}_\theta[p(\theta;\alpha,\beta)] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

2.3 Equating Coefficient for Inference in Conjugate Models

In the above example, we implicitly used a technique called equating coefficients. This approach to posterior inference is particularly powerful when dealing with probabilistic models with conjugate priors. The central process is the following:

- Set up your inference by selecting a conjugate prior and likelihood.
- Expand your likelihood and prior with their probability density functions.
- Simplify and reduce as much as possible, and strategically rearrange to make the next step easier.
- Given that we know the form of the posterior, solve for the coefficients that give the parameters of the posterior.

Above, because we knew that the Beta distribution was the conjugate prior to the Bernoulli distribution, in the end we saw that our posterior was just a Beta distribution. This allowed us to compute the closed form normalizing constant, mean, mode, and variance which, as we will see, are all very useful for inference.

Please note that in the lecture slides we have also gone over an electrical communication example using a normal distribution. It is a good idea to look through this set up to get another example of the equating coefficients method.

2.4 Posterior Predictive Distribution

This section was discussed in Lecture 9. Above, we have seen and discussed prior distributions for our Bayesian models. But in machine learning, we are not just interested in computing a posterior distribution. We are interested in *using* the posterior distribution to make predictions about unseen objects and events. To uncover how to do this in a Bayesian framework let's express probabilistically the quantity we are interested in. First, let us assume a supervised learning scenario. Next, we know we have some unseen feature vector x. The probability we are interested in is, $p(x^*|\mathcal{D})$ or in words, *the probability distribution of a feature vector x given the dataset we have previously observed*. Recall that we have access to $p(\theta|\mathcal{D})$ and so by integrating over θ we can compute what we want. In particular:

$$p(x^*|\mathcal{D},\theta) = \int_\theta p(x^*|\theta)p(\theta|\mathcal{D})$$

Not only does this equation guide us in making our predictions (as we will see in future lectures), but it also allows us to evaluate our model given a held out set of data. For example, we may consider computing the (log-)likelihood of seeing the data-points $\{0, 1, 1\}$ under our posterior distribution. Let us start with just computing the likelihood of seeing a point x^*. Using the i.i.d. assumption, we have that:

$$\begin{aligned} p(x^*|\mathcal{D}) &\propto \int_\theta p(x^*|\theta)p(\theta|\mathcal{D}) \\ &= \int_\theta p(x^*|\theta)p(\theta|\mathcal{D}) \end{aligned}$$

A Lecture 8: Bayesian Inference

Question 1 (Electrical Communication)**.** Consider the electrical communication example from lectures, where we had a Gaussian distribution on the source voltage, i.e. $p(S = s) = \mathcal{N}(s; 0, 1)$. This time, we make multiple observations, i.e. $V_n|s \overset{\text{iid}}{\sim} \mathcal{N}(s, \sigma^2)$.

a. Write down Bayes' rule to find the posterior for $p(s|v_1, v_2, \dots v_N)$.

b. By completing the square, find the density of the posterior $p(s|v_1, v_2, \dots v_N)$.

c. Show that the likelihood function (which is a function of s!) can be rewritten as

$$p(v_1, v_2, \dots, v_N|s) = c \cdot p(\bar{v}|s) = c \cdot \mathcal{N}\left(\bar{v}; s, \frac{\sigma^2}{N}\right), \tag{1}$$

where $\bar{v} = \frac{1}{N}\sum_{n=1}^{N} v_n$.

d. Find the joint distribution $p(\bar{v}, s)$.

e. Use the Gaussian conditioning rule to find $p(s|\bar{v})$.

f. Reflect on which method you find easier.

Question 2 (Electrical Communication Errors)**.** Consider the electrical communication example from lectures, where $p(v|s) = \mathcal{N}(v; s, \sigma^2)$ and a Bernoulli S, i.e. $p(S = s) = p^s(1-p)^{1-s}$. Assume that the noise distribution in the model is the same as that of the data generating process. Now consider a *decision rule*, where we guess the transmitted signal using the rule $\hat{S} = \text{argmax}_s\, p(s|v)$.

Now consider the true frequency of S to follow $\pi(s) = \mathcal{B}(0.6)$. Calculate the probability of making an error $\mathbb{P}\left(\hat{S} = S\right)$, as a function of our prior probability p.

Hint: Remember the difference between the data generating distribution ($\mathbb{P}/\pi$), and our model (P/p). When calculating probabilities w.r.t. the data generating distribution, $\hat{S}$ is a function of S.

Question 1 (Electrical Communication)**.** Consider the electrical communication example from lectures, where we had a Gaussian distribution on the source voltage, i.e. $p(S = s) = \mathcal{N}(s; 0, 1)$. This time, we make multiple observations, i.e. $V_n|s \overset{\text{iid}}{\sim} \mathcal{N}(s, \sigma^2)$.

a. Write down Bayes' rule to find the posterior for $p(s|v_1, v_2, \dots v_N)$.

b. By completing the square, find the density of the posterior $p(s|v_1, v_2, \dots v_N)$.

c. Show that the likelihood function (which is a function of s!) can be rewritten as

$$p(v_1, v_2, \dots, v_N|s) = c \cdot p(\bar{v}|s) = c \cdot \mathcal{N}\left(\bar{v}; s, \frac{\sigma^2}{N}\right), \tag{1}$$

where $\bar{v} = \frac{1}{N}\sum_{n=1}^{N} v_n$.

d. Find the joint distribution $p(\bar{v}, s)$.

e. Use the Gaussian conditioning rule to find $p(s|\bar{v})$.

f. Reflect on which method you find easier.

$P(S=s) = \mathcal{N}(S; 0, 1)$. $P(V_n|s) \sim \mathcal{N}(S, \sigma^2)$

a Bayes: $P(B|A)$: $\frac{P(A|B)P(B)}{P(A)}$ → $P(S|V_n) = \frac{P(V_n|s)P(S)}{P(V_n)}$ (likelihood, prior, posterior)

$P(V_n|S) = P(\{V_i\}_{i:1\dots n}|S)$, V_i's are iid ∴ $P(V_1, V_2, V_3 \dots) = P(V_1)P(V_2)\dots P(V_n) = \hat{\prod_{i=1}} P(V_i)$

$P(S|V_1, \dots, V_n)$: $\frac{P(V_1, \dots, V_n|S)P(S)}{P(V_1 \dots V_n)} = \frac{\hat{\prod_{i=1}}(V_i|S)P(S)}{P(V_1 \dots V_n)}$

b. Normal density × Normal density: normal density

$$\hat{\prod_{i=1}}(V_i|S) = \hat{\prod_{i=1}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(V_i - S)^2}{\sigma^2}}$$

$P(S)$: $N(0,1)$ → $\frac{1}{\sqrt{2\pi \times 1^2}} e^{-\frac{1}{2}\frac{(S-0)^2}{1^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{S^2}{2}}$

$$P(S|V_i) \propto P(\{V_i\}|S)P(S) = \hat{\prod_{i=1}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(V_i - S)^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}S^2}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\wedge} e^{-\frac{1}{2\sigma^2}\sum_i (V_i - S)^2} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{S^2\sigma^2}{\sigma^2}}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\wedge} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\sum_i \left[(V_i - S)^2\right] + S^2\sigma^2}$$

$$\propto e^{-\frac{1}{2\sigma^2}\left[\sum V_i^2 - 2S\sum V_i + nS^2 + S^2\sigma^2\right]}$$

$$= c^{-\frac{1}{2\sigma^2}\left((n+\sigma^2)S^2 - (2\sum V_i)S + \sum[V_i]^2\right)}$$

Gaussian: $e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} = e^{\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)}$ — need to take out $(n+\sigma^2)$

→ $-\frac{1}{2\sigma^2}(n+\sigma^2)\left[S^2 - \frac{2S\sum V_i}{n+\sigma^2} + \frac{\sum(V_i^2)}{n+\sigma^2}\right)$

$\propto \left[S^2 - \frac{2\sum V_i}{n+\sigma^2}S + \frac{\sum V_i^2}{n+\sigma^2}\right)$

$2\mu: \frac{2\sum v_i}{1+\sigma^2} \to \mu: \frac{\sum v_i}{1+\sigma^2}$

Letting $\sigma_p^2 = \frac{\sigma^2}{1+\sigma^2}$ and $\mu_p: \frac{\sum v_i^2}{\sigma^2+N}$

$\frac{1}{2\sigma_p^2}: \frac{(1+\sigma^2)}{2\sigma^2} \to 2\sigma_p^2: \frac{2\sigma^2}{1+\sigma^2} \to \sigma_p^2 = \frac{\sigma^2}{1+\sigma^2}$

$\therefore p(s|v_i) = \mathcal{N}\left(\frac{\sum v_i^2}{\sigma^2+N}, \frac{\sigma^2}{1+\sigma^2}\right)$

c. Show that the likelihood function (which is a function of s!) can be rewritten as

$$p(v_1, v_2, \ldots, v_N|s) = c \cdot p(\bar{v}|s) = c \cdot \mathcal{N}\left(\bar{v}; s, \frac{\sigma^2}{N}\right),$$

where $\bar{v} = \frac{1}{N}\sum_{n=1}^{N} v_n$.

Show likelihood $\propto$ dist of mean

$p(v_i|s) \to \prod_{i=1}^{N} p(v_i|s) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(v_i-s)^2}{\sigma^2}} = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\sum v_i^2 - 2s\sum v_i + Ns^2}$

$\bar{v}: \frac{1}{N}\sum_{i=1}^{N} v_i$

$p\left(\frac{1}{N}\sum_{i=1}^{N} v_i\right) = \text{const} \cdot e^{\frac{N}{2\sigma^2}\left(\frac{1}{N}\sum_{i=1}^{N} v_i - s\right)^2} = e^{\frac{N}{2\sigma^2}\left(\left(\frac{1}{N}\sum v_i\right)^2 - \frac{2s}{N}\sum v_i + s^2\right)}$

$\frac{1}{2\sigma_s^2}: \frac{N}{2\sigma^2} \to \sigma_s^2 = \frac{\sigma^2}{N}$

$\mu: s$

d. Find the joint distribution $p(\bar{v}, s)$.

$p(\bar{v}, s): p(\bar{v}|s)p(s)$

$\bar{v}: \frac{1}{N}\sum_{n=1}^{N} v_n$

e. Use the Gaussian conditioning rule to find $p(s|\bar{v})$.

Gaussian Conditioning

$$p\begin{pmatrix}x\\y\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}x\\y\end{pmatrix}; \begin{pmatrix}m_x\\m_y\end{pmatrix}, \begin{pmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{pmatrix}\right)$$

We have:

$$p(x|y) = \mathcal{N}\left(x; m_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - m_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right)$$
$$p(x) = \mathcal{N}(x; m_x, \Sigma_{xx})$$

$p(s,\bar{v}) = \mathcal{N}\left(\begin{pmatrix}s\\ \bar{v}\end{pmatrix}; \begin{pmatrix}m_s\\ m_{\bar{v}}\end{pmatrix}, \begin{pmatrix}\sigma_s^2 & \sigma_{s\bar{v}}^2\\ \sigma_{\bar{v}s}^2 & \sigma_{\bar{v}}^2\end{pmatrix}\right)$

m_s : $s \sim N(0,1) \rightarrow m_s = 0$

$m_{\bar{v}}$: $\bar{v} = \frac{1}{N}\sum v_n$, $v \sim N(s, \sigma^2)$

$E[\bar{v}] = \frac{1}{N} E[\sum v_n] = \frac{1}{N} E[v_1 + v_2 \cdots] = \frac{1}{N}(E[v_1] + E[v_2] \cdots) = \frac{1}{N}(Ns) = s \rightarrow 0$

$\Sigma_{xx} = \sigma_s^2 = Var(s) = 1$

Σ_{xy} : $Cov(s,\bar{v}) = E\left[\left(\frac{1}{N}\sum v_i\right)s\right] - \underbrace{E[s]}_{=0} E\left[\frac{1}{N}\sum v_i\right]$ $\quad Cov(x,y) = E[xy] - E[x]E[y]$

$= Cov(\bar{v}, s)$

$= E\left[\left(\frac{1}{N}\sum s + \epsilon_i\right)s\right]$

$= E\left[\frac{1}{N}\sum s^2 + \epsilon_i s\right] = \frac{1}{N}\left(\sum s^2 + \sum \epsilon_i s\right) = \frac{1}{N}(Ns^2 + \sum \epsilon_i s)$

$= E[s^2] + E\left[\sum \epsilon_i s\right]$

$= E[s^2] + \underbrace{E\left[\sum \epsilon_i s\right]}_{\text{dependent}} \rightarrow E[\epsilon_i] E[s]$

$= E[s^2]$

$= E[(s - E[s])^2]$, $E[s] = 0$

$= Var(s)$

$= 1$

$\Sigma_{yy} = \sigma^2_{\bar{v}} = Var\left[\frac{1}{N}\sum v_i\right] = Var\left[\sum v_i\right] = Var\left(\frac{1}{N}\sum s + \epsilon_i\right)$

$= Var\left(\frac{1}{N}\sum s + \sum \epsilon_i\right) = Var\left(\frac{1}{N}(Ns + \sum \epsilon_i)\right) = Var\left(s + \frac{1}{N}\sum \epsilon_i\right)$

$= Var(s) + Var\left(\frac{1}{N}\sum \epsilon_i\right)$ $\quad \epsilon \sim N(0, \sigma^2)$

$= 1 + \frac{1}{N^2} N\sigma^2$

$1 + \frac{\sigma^2}{N}$

$p(s|\bar{v}) = N\left(s; \frac{\bar{v}}{1+\frac{\sigma^2}{N}}, 1 - \left(1 + \frac{\sigma^2}{N}\right)^{-1}\right) = N\left(s; \frac{\bar{v}}{1+\frac{\sigma^2}{N}}, 1 - \frac{1}{1+\frac{\sigma^2}{N}}\right)$

Question 2 (Electrical Communication Errors)**.** Consider the electrical communication example from lectures, where $p(v|s) = \mathcal{N}(v; s, \sigma^2)$ and a Bernoulli S, i.e. $p(S = s) = p^s(1-p)^{1-s}$. Assume that the noise distribution in the model is the same as that of the data generating process. Now consider a *decision rule*, where we guess the transmitted signal using the rule $\hat{S} = \operatorname{argmax}_s p(s|v)$.

Now consider the true frequency of S to follow $\pi(s) = \mathcal{B}(0.6)$. Calculate the probability of making an error $\mathbb{P}\left(\hat{S} = S\right)$, as a function of our prior probability p.

Hint: Remember the difference between the data generating distribution ($\mathbb{P}$/π), and our model (P/p). When calculating probabilities w.r.t. the data generating distribution, $\hat{S}$ is a function of S.

$p(V|S) : N(V; S, \sigma^2)$ $\quad p(S=s) = p^s(1-p)^{1-s}$

$V = S + \varepsilon$, ε

$\hat{S} : \underset{s}{\operatorname{argmax}}\ p(s|v)$ $\quad$ True S : $\pi(s) = B(0.6)$: $0.6^s(1-0.6)^{1-s} = 0.6^s(0.4)^{1-s}$

$P(1) = 0.6$, $P(0) = 0.4$

find $P(\hat{S} = S)$

$P(\hat{S}(v) = S) : P(\hat{S}(v) = 1 \mid S = 1)P(S = 1) + P(\hat{S}(v) = 1 \mid S = 0)P(S = 0)$

$= P(\quad)\pi + P(\quad)(1-\pi)$

$\hat{S} : \operatorname{argmax}\ p(v|s)p(s)$