University of Melbourne

Department of Mathematics and Statistics

Master Thesis

Profitable Strategies in Horse Race Betting Markets

Author:
Alexander Davis

Supervisor:
Dr. Owen Jones

October 18, 2013

Abstract:

In this paper we discuss a way to construct a strategy of betting that can generate abnormal returns in horse race betting markets. Using the theory of discrete choice models and Generalised Additive Models one can build a model to predict the probability of a given horse winning a given race. The paper further discusses the Kelly betting system as a way of exploiting estimates of winning probabilities so to generate abnormal returns.

1 Introduction

In finance and economics literature there is a commonly held notion that prices in a competitive market should accurately reflect all publicly available information about the worth of the commodities being transacted [29]. This is the notion of *weak-form efficiency*: it should be impossible to construct speculative strategies that consistently outperforms the market using only historical prices as inputs [46].

Betting markets have come into particular scrutiny, in this regard [42, 46, 16, 18]. For example, Johnson et al. were able to construct a model that achieved abnormal returns, through exploiting the discrepancy between the *track* probabilities (probability of winning implied by the horses' odds) and *model* probabilities, which are estimated via a statistical procedure [18]. To estimate the winning probabilities for horses, Johnson et al. used a discrete choice model known as McFadden's conditional logit model. In this model, it is supposed that the probability of horse j winning race i is dependent on a certain score function which is a linear combination of certain predictor variables. The predictor variables chosen were themselves statistics calculated from the historical prices (odds); thus, if one could use these model probabilities to construct a profitable betting strategy it would suggest that the market was not weak-form efficient. Indeed, this is what Johnson et al. were able to do (although they were hesitant about actually implementing such a betting strategy in real life, due to the operational effort required and the large risks involved [18]).

The motivation for this paper is to extend the model that was used in Johnson to obtain a more accurate approximation of the probabilities of horses' victory. The extension involves replacing the score function with a sum of smooth functions, which leads us to explore the theory of generalised additive models (GAMs). The modification of the model in this manner allows for a greater model fit, since the smooth functions have greater freedom to fit the data than the cruder score function used in Johnson (however, it does make the model less easy to interpret).

The thesis can roughly be divided into two different parts.

The fist deals with the problem of estimating the probabilities of a given horse winning a particular race. We initially discuss the discrete choice models that have been used to model individuals' economic behaviour, with particular focus on McFadden's conditional logit model. Then, we discuss the theory of generalised linear models (GLMs) before introducing the more sophisticated GAM. After this discussion we consider the problem of merging the theory of discrete choice models and GAMs by using vector generalised additive models (VGAMs).

The second part deals with the problem of choosing precisely how much one ought to bet on horse j in race i given that one has estimates of the probabilities of winning. The particular betting system that will be discussed is the well known Kelly betting system. We will discuss its motivation, its limitations and other possible betting systems one could employ.

2 Discrete Choice Models

As discussed before, one of our chief tasks in developing a betting strategy that yields positive profits is finding a satisfactory way of estimating the probabilities of a given horse j winning race i. This section introduces the set of models that are useful for estimating such probabilities.

To estimate these probabilities we shall consider a range of models for individual choice (where the set of choices is discrete) studied in the field of econometrics and mathematical psychology, called *discrete choice models* [28]. These models were originally conceived with the intent of modeling individuals' behavior when confronted with a set of discrete choices, and there are numerous applications found in the literature, in particular labor markets and transportation [28, 32, 2]. However, they are not an infrequent feature in the literature that concerns betting markets [18] (in particular, McFadden's conditional logit model, which we will introduce soon), where they are used either as a means of modeling bettor's behavior, or simply as a means of estimating actual probabilities.

2.1 The Multinomial Logit Model

Suppose that we wish to consider the behavior of a group of individuals who are presented with m choices, with probability P_j of making choice j, where $j = 1, ..., m-1$. Here we assume that the ordering of these choices (indicated by their respective indices) is inconsequential. We further suppose that we are able to express the probabilities in the following way:

$$\frac{P_j}{P_j + P_m} = F(\beta_j' x) \tag{1}$$

where $j = 1, ..., m-1$[1] and P_m is the probability of the individual making the mth choice, while β and x are vectors and F is the *logistic* cumulative distribution function, given by $F(x) = \exp(x)/(1+\exp(x))$ [28].

This yields, for $j = 1, ...m-1$, the following:

$$\frac{P_j}{P_m} = \frac{F(\beta_j' x)}{1 - F(\beta_j' x)} =: G(\beta_j' x)$$

By the law of total probability we then have that

$$\sum_{j=1}^{m-1} \frac{P_j}{P_m} = \frac{1 - P_m}{P_m} = \frac{1}{P_m} - 1$$

So, we may express P_m as:

$$P_m = \frac{1}{1 + \sum\limits_{j=1}^{m-1} G(\beta_j' x)}$$

[1] Even though there are m different choices, we restrict the value of j as a consequence of our representation of the probabilities in (1)

Consequently, using our definition of $G(\beta_j'x)$ we derive an expression for P_j as

$$P_j = \frac{G(\beta_j'x)}{1 + \sum\limits_{k=1}^{m-1} G(\beta_k'x)}$$

Now, since the cumulative distribution of F is logistic, then we have that

$$P_j = \frac{\exp(\beta_j'x)}{1 + \sum\limits_{k=1}^{m-1} \exp(\beta_k'x)} \tag{2}$$

Since in the case where F is logistic we can evaluate G as

$$\begin{aligned} G(\beta_j'x) &= \frac{F(\beta_j'x)}{1 - F(\beta_j'x)} \\ &= \frac{1}{1+\exp(-\beta_j'x)}\left(1 - \frac{1}{1+\exp(-\beta_j'x)}\right)^{-1} \\ &= \exp(\beta_j'x) \end{aligned}$$

Now, the model specified by equation (2) is called the *multinomial logit model* [28].

An example of a multinomial logit (MNL) model is that which is used by Schmidt and Strauss [43]. In this model they consider the case where individuals choose between m different occupations, where their respective choices are dependent on certain individual characteristics which they possess (such as age, gender, years of schooling), represented by vectors y_i for each individual i. The probability of an individual i choosing occupation j is then given by

$$P_{ij} = \frac{\exp(\alpha_j'y_i)}{\sum\limits_{t=1}^{m} \exp(\alpha_t'y_i)} \tag{3}$$

where the model must be normalised on account of the fact that by the total law of probability if we know $m-1$ many of the probabilities, we must know the remaining one. Such normalisation can be achieved simply by assigning, say, $\alpha_m' = 0$, in which case we yield the familiar form of the MNL model.

Say there exist k individual characteristics for the individuals concerned. Then, there exist $k(m-1)$ many parameters to be estimated in the model (these being the weights corresponding to each individual characteristic, for each different occupational choice). After these parameters are estimated, then given a new individual h with individual characteristics y_h , we may estimate the probability of that individual choosing occupation j, $\hat{P_{hj}}$, by substituting the vector $\hat{\alpha_j} = (\hat{\alpha_{j_1}}, ..., \hat{\alpha_{j_k}})$ and the vector y_h into equation (**??**), giving us

$$\hat{P_{hj}} = \frac{\exp(\hat{\alpha_j}' y_h)}{1 + \sum_{t=1}^{m-1} \exp(\hat{\alpha_t}' y_h)}$$

where we apply the normalisation of setting $\alpha_m = 0$ before estimating the vectors $\alpha_1, ..., \alpha_{m-1}$.

2.2 McFadden's Conditional Logit Model

Similar to the MNL model is *McFadden's conditional logit model*, named in recognition of the econometrician Daniel McFadden, who pioneered the use of this model in analysing discrete choice problems[2] [32]. McFadden was able to derive the conditional logit model using a utility theoretic argument [32], which we shall demonstrate later to derive a slightly more general model.

To explain McFadden's conditional logit model we will describe a particular application of it given in Boskin [2]. Similar to the the MNL model setup, we wish to model the behavior of a group of individuals who are presented with m choices (as in the Schmidt and Strauss case, these will be choices of occupation). However, unlike the MNL, in Boskin's model we wish to know how these individuals value the different characteristics of the particular occupations. Boskin focuses on three such characteristics: present value of potential earnings, training cost/net worth, and present value of time unemployed. By letting x_{ij} be the vector of values corresponding to the characteristics of occupation j as perceived by individual i, we then suppose that the probability of individual i choosing occupation j is given by

$$P_{ij} = \frac{\exp(\beta' x_{ij})}{\sum_{k=1}^{m} \exp(\beta' x_{ik})} \tag{4}$$

where we let β be the vector which weights the relative importance of the choice characteristics in determining if an individual will choose a given occupation; Boskin calls these the 'implicit prices' for the characteristics. Furthermore, an identification constraint must be imposed, which can be achieved by simply letting $\beta = 1$, for instance. So, for a model with k many choice characteristics, our implicit prices vector, β, must have dimension k, but with only $k - 1$ elements that vary.

After estimating the vector β for the model one can estimate the probability $\hat{P_{il}}$ that a new occupation l (i.e. one that was not used in the estimation procedure) will be chosen by a particular individual i, given that we know the vector x_{ij} of l's occupation characteristics as perceived by individual i. Supposing that this is the only new occupation under consideration, the estimate of the probability can be given by

[2] In fact, McFadden was to share the Nobel Prize in Economic Sciences in 2000 with James Heckman, for his work on discrete choice problems [17].

$$\hat{P}_{il} = \frac{\exp(\hat{\beta}' x_{il})}{\sum_{k=1}^{m+1} \exp(\hat{\beta}' x_{ik})}$$

Boskin actually considered four such models; after separating the total amount of job seekers into four different categories (white male, black male, white female, black female), different vectors β_t were estimated corresponding to these categories. As a cautionary note, in comparing the different estimates of these vectors so to ascertain the different 'implicit prices' these groups attached to the choice characteristics, one can only compare relative - not absolute- values of the coefficients, due to the scaling procedure.

It should be noted that, in fact, the Multinomial Logit model and McFadden's Conditional Logit model are algebraically equivalent. The following is a demonstration of this equivalence, given in Maddala [28]. For what follows we will drop the subscript j, as it unnecessarily complicates the notation.

Consider the MNL model $P_i/P_1 = \exp((\beta_i - \beta_1)x)$ where $x = (z_1, ..., z_n)$ and $\beta_i = (0, ..., \alpha, ..., 0)$. From these choices of x and β_i we yield the conditional logit model

$$\frac{P_i}{P_1} = \exp(\alpha(z_i - z_1))$$

We can also demonstrate this equivalence by considering the conditional logit model $P_i/P_1 = \exp((\alpha(z_i - z_1))$ where $\alpha = (\beta_1, ..., \beta_m)$ and $z_i = (0, ..., x, ...0)$. These choices of α and z imply that our model is simply the MNL model given by

$$\frac{P_i}{P_1} = \exp((\beta_i - \beta_1)x)$$

2.3 The Combined Model and its Derivation

We can combine the models of Schmidt and Strauss (3), and Boskin (4), to produce the following model of the probability of individual $i \in \{1, ..., n\}$ making choice $j \in \{1, ..., m\}$:

$$P_{ij} = \frac{\exp(\beta' x_{ij} + \alpha_j' z_i)}{\sum_{k=1}^{m} \exp(\beta' x_{ik} + \alpha_k' z_i)} \tag{5}$$

where x_{ij} is the known vector of choice characteristics of choice j as perceived by person i, z_i is the known vector of individual characteristics of person i, β is the vector of 'implicit prices' for the choice characteristics, and α is the vector of parameters which weight the particular choice characteristics for each choice j.

A derivation of this model from utility theory is given by Maddala [28]. In this derivation we suppose that the individual is economically rational; that is, we suppose that they strive to maximize their utility subject to constraints, such as their limited financial resources. Since individuals, however, cannot necessarily calculate with complete accuracy the level of utility of the choices they are presented with, we suppose that their utility is random.

Let's suppose that an individual is presented with m choices. We assume that this individual receives Y_j^* utility from making choice j, where

$$Y_j^* = V(x_j) + \epsilon_j \tag{6}$$

V being a function that maps the vector of attributes of the jth choice, x_j, to R, while ϵ_j is a random quantity that represents both variations in consumer tastes and errors in the consumer's perception of the utility of choice j (note z will re-enter this model later via a substitution).

One does not observe the utilities $Y_1^*, Y_2^*, ...Y_m^*$ (called latent variables), but rather the outcome of the individual's choice. In other words, we can define the following observable variable Y_j such that $Y_j = 1$ if $Y_j^* = \max(Y_1^*, ...Y_m^*)$, otherwise $Y_j^* = 0$. In other words, the observable variable Y_j indicates whether an individual has made choice j ($Y_j = 1$) or not ($Y_j = 0$).

If we suppose that the errors ϵ_j are i.i.d with the *Gumbel* distribution (type-I extreme value) [33] then $P(\epsilon_j < \epsilon) := F(\epsilon) = \exp(-\exp(-\epsilon))$, which has density $f(\epsilon) = \exp(-\epsilon - \exp(-\epsilon))$ (note that the errors are not symmetric).

Now, since $\{Y_j^* > Y_k^*\}, \forall k \neq j$ is equivalent to $\epsilon_k < \epsilon_j + V_j - V_k, \forall k \neq j$ (where V_j and V_k are shorthand for $V(x_j)$ and $V(x_k)$, respectively) we have that

$$\begin{aligned}
Pr(Y_j = 1) &= Pr(Y_j^* > Y_k^*), \forall k \neq j \\
&= Pr(\epsilon_k < \epsilon_j + V_j - V_k), \forall k \neq j \\
&= \int_{-\infty}^{\infty} \prod_{k \neq j} F(\epsilon_j + V_j - V_k) f(-\epsilon_j) d\epsilon_j \\
&= \int_{-\infty}^{\infty} \prod_{k \neq j} \exp(-\exp(-\epsilon_j - V_j + V_k)) \exp(-\epsilon_j - \exp(-\epsilon_i)) d\epsilon_j \\
&= \int_{-\infty}^{\infty} \exp\left[\epsilon_j - \exp(-\epsilon_j)\left(1 + \sum_{k \neq j} \frac{\exp(V_k)}{\exp(V_j)}\right)\right] d\epsilon_j \\
&= \int_{-\infty}^{\infty} \exp\left[\epsilon_j - \exp(-\epsilon_j)\left(\sum_{k=1}^{m} \frac{\exp(V_k)}{\exp(V_j)}\right)\right] d\epsilon_j
\end{aligned}$$

Letting $\lambda_j := \ln(\sum_{k=1}^m \frac{\exp(V_k)}{\exp(V_j)})$ and $\epsilon_j^* = \epsilon_j - \lambda_j$ the above integral becomes

$$\int_{-\infty}^{\infty} \exp(-\epsilon_j - \exp(-(\epsilon_j - \lambda_j))d\epsilon_j = \exp(-\lambda_j) \int_{-\infty}^{\infty} \exp(-\epsilon_j^* - \exp(-(\epsilon_j^*))d\epsilon_j^*$$
$$= \exp(-\lambda_j)$$
$$= \frac{\exp(V_j)}{\sum_{k=1}^{m} \exp(V_k)} \tag{7}$$

We can extend the concept of latent utility as expressed in equation (6) to include individuals $i \in \{1, ..., n\}$. As a consequence of this we will redefine $V's$ input, x_j, to be x_{ij}, the vector of values of characteristics of choice j as perceived by individual i. Moreover, we will specify the actual function, V, to be

$$V(x_{ij}) = \beta' x_{ij} + \alpha_j' z_i \tag{8}$$

so that the indirect utility obtained by individual i from making choice j is given by

$$Y_{ij}^* = \beta' x_{ij} + \alpha_j' z_i + \epsilon_{ij} \tag{9}$$

Moreover, we can define the observable variables Y_{ij} in a completely analogous way to that which was done before, so that $Y_{ij} = 1$ if individual i makes the jth choice, and is zero otherwise. Consequently, substituting (8) into expression (7) gives us

$$P_{ij} = \frac{\exp(\beta' x_{ij} + \alpha_j' z_i)}{\sum\limits_{k=1}^{m} \exp(\beta' x_{ik} + \alpha_k' z_i)} \tag{10}$$

which is the combined model, as desired.

2.4 Connection with the Luce Model

There is a noteworthy connection between McFadden's conditional logit model and the *Luce model.* The Luce model is a choice model derived from the Independence of Irrelevant Alternatives (IIA) principle, which states that the ratio of probabilities of making choices j and j' (also known as the odds ratio) do not change if the set of possible choices is expanded [28]. The combined model (which has the conditional logit and multinomial logit models as special cases) is itself a special case of the Luce model, as it is clear from our definition of the combined model that if we increase the total number of choices from m to m', we have that for choices j and l

$$\begin{aligned}
\frac{P_{ij}}{P_{il}} &= \frac{\exp(\beta' x_{ij} + \alpha'_j z_i)}{\sum\limits_{k=1}^{m} \exp(\beta' x_{ik} + \alpha'_k z_i)} \times \frac{\sum\limits_{k=1}^{m} \exp(\beta' x_{ik} + \alpha'_k z_i)}{\exp(\beta' x_{il} + \alpha'_l z_i)} \\
&= \frac{\exp(\beta' x_{ij} + \alpha'_j z_i)}{\exp(\beta' x_{il} + \alpha'_l z_i)} \\
&= \frac{\exp(\beta' x_{ij} + \alpha'_j z_i)}{\sum\limits_{k=1}^{m'} \exp(\beta' x_{ik} + \alpha'_k z_i)} \times \frac{\sum\limits_{k=1}^{m'} \exp(\beta' x_{ik} + \alpha'_k z_i)}{\exp(\beta' x_{il} + \alpha'_l z_i)}
\end{aligned}$$

The IIA principle has been criticized by Debreu [7], arguing that its existence leads to outcomes that seem counter-intuitive. Debreu considers the example of a person with an ear for classical music [37]. This person has to choose whether to add to their collection a recording of Claude Debussy's (only) String Quartet, or one of two different recordings of Ludwig van Beethoven's Eighth Symphony. We further suppose that the Beethoven recordings provide the same listening experience (so, they are essential identical), and that the classical music lover is indifferent between the music of Debussy and Beethoven.

Denote the Beethoven recordings by B_1 and B_2, and the Debussy recording by D. Moreover, denote by $P(T|T,U)$ the probability that option T is chosen given a choice between T and U, and $P(T|T,U,V)$ likewise for when this choice is expanded to include V. Then, according to the model, we have that

$$P(B_1|B_1,B_2) = P(D|D,B_1) = P(D|D,B_2) = \frac{1}{2} \tag{11}$$

which, from the IIA principle, implies that

$$P(D|B_1,B_2,D) = \frac{1}{3}$$

which is readily proved in the following way.

Suppose that $P(D|B_1,B_2,D) = a$, where $a \in (0,1)$. Now, by the equations from (11) we must have that

$$\frac{P(B_1|B_1,B_2)}{P(D|D,B_1)} = \frac{1/2}{1/2} = 1$$

but by the IIA principle, this means that

$$1 = \frac{P(B_1|B_1,B_2,D)}{P(D|B_1,B_2,D)} = \frac{P(B_1|B_1,B_2,D)}{a}$$

which implies that $P(B_1|B_1, B_2, D) = a$. However, by symmetry, we must have that $P(B_2|B_1, B_2, D) = a$. Consequently, by the law of total probability we have that

$$\begin{aligned} 1 &= P(B_1|B_1, B_2, D) + P(B_2|B_1, B_2, D) + P(D|B_1, B_2, D) \\ &= 3a \end{aligned}$$

which implies that $a = 1/3$.

This result seems intuitively implausible, as both Beethoven recordings are essentially the same; that is, the music lover's decision seems to be, rather, a binary one between Beethoven or Debussy, in which case we would expect $P(D|B_1, B_2, D) = \frac{1}{2}$, contrary to the Luce model [3].

Since, however, our designs for McFadden's conditional logit model are simply to find the determinants of a horse's 'winning-ness', rather than model the actual behaviour of individuals, this objection can be put to one side.

2.5 Application to Betting Markets and Extension

McFadden's conditional logic model is particularly useful to our purposes of estimating the probability of a particular horse j winning race i. We can consider the choice (winning) characteristics, x_j, as including, for instance, the age of the horse, its recent performance and its closing odds. In fact, one could narrow the characteristics of concern to being simply statistics derived from historical prices (odds), as is done in Johnson et al [18], where such statistics could include, for instance, the horse's closing odds, the number of changes in its odds and the total magnitude of changes in its odds over the course of the betting market. Consequently, the parameter β simply weights the importance of these different characteristics in determining whether a horse with such characteristics wins a given race.

For estimating the parameter β of McFadden's conditional logit model one can use the well-known procedure of maximum likelihood estimation, a method which shall be discussed later in regards to estimating the parameters of generalised linear models. The estimate following this method, $\hat{\beta}$, possesses the desirable properties of consistency ($\hat{\beta} \xrightarrow{p} \beta$) and asymptotic normality, a result proved by McFadden [32].

However, within the context of this thesis we wish to extend this model by generalising the 'winning-ness score', $< \beta, x_j >:= \beta' x_j$, so that instead of simply being a linear combination of the characteristics given in the vector x_j, we have that this new score is represented as the sum of smooth functions of these characteristics

$$f_1(x_1) + f_2(x_2) + ...f_m(x_m)$$

where m is the number of different characteristics (note how we have replaced our vector β with smooth functions).

[3] When experiments to test these models are conducted, the described phenomenon is called a 'similarity effect' [37].

To achieve this we must introduce the concept of a Generalised Additive Model (GAM). Before we can do this, however, we must first discuss a much simpler model which is close to the GAM, called the Generalised Linear Model (GLM).

3 Generalized Linear Models

3.1 Generalized Linear Models and their Parameters

Before we investigate the theory behind Generalized Additive Models, it is important to discuss Generalized Linear Models (GLMs). If, for all $i \in \{1, 2, ...n\}$, we define $\mu_i = E(Y_i)$ where Y_i are independent and have a distribution from an exponential family, g to be a smooth, monotonic function (called a 'link' function), X_i to be the ith row of a model matrix, X, and β to be a vector of parameters that we wish to estimate, then a Generalized Linear Model is defined by the equation [47]:

$$g(\mu_i) = X_i\beta$$

Note that the well-known Linear Statistical Model is simply a special case of a GLM (if we choose Y_i to be normally distributed and g to be the identity function). The added complexity of the GLM as compared to the Linear Statistical Model (LSM) has the drawback of requiring iterative algorithms for model fitting, as well as only yielding approximate distributional results instead of exact ones.

Given that Y has a distribution from an exponential family of distributions, we may write its density in the following form:

$$f_\theta(y) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y, \phi)\}$$

where θ is the 'canonical parameter' of the distribution, ϕ is called the scale parameter or dispersion parameter[31], while a, b and c are arbitrary functions. If we let $l = \log(f_\theta(y))$ then by taking the partial derivative of l with respect to θ and using the well-known result that $E(\partial l/\partial\theta) = 0$ [47] we have that

$$0 = E\left(\frac{Y - b'(\theta)}{a(\phi)}\right)$$

which implies that

$$E(Y) = b'(\theta) \tag{12}$$

Moreover, taking the double partial derivative of l with respect to θ and using another well-known result that $E(\partial^2 l\ /\partial\theta^2) = -E((\partial l\ /\partial\theta)^2)$ [31], we have that

$$\frac{-b''(\theta)}{a(\phi)} = -E\left(\frac{Y - b'(\theta)}{a(\phi)}\right)^2$$

which, using the result from (12), is equivalent to

$$\begin{aligned}\frac{-b''(\theta)}{a(\phi)} &= -E\left(\frac{Y - E(Y)}{a(\phi)}\right)^2 \\ &= -\frac{Var(Y)}{a(\phi)^2}\end{aligned}$$

Multiplying through by $-a(\phi)^2$ gives us the result

$$Var(Y) = b''(\theta)a(\phi) \tag{13}$$

It should be stressed that equation (12) is important as it links the canonical parameters of the distribution of Y with the model parameters in the vector, β (since $E(Y) = \mu = X\beta$). With regards to equation (13), typically we will set $a(\phi) = \phi/\omega$ and so

$$Var(Y) = V(\mu)\phi \tag{14}$$

where $V(\mu) = b''(\theta)/\omega$ is called the *variance function*[31].

To illustrate the applications of GLM theory, we will consider an example given in Wood [47].

Consider an epidemic within a population. In the initial stages of an epidemic, one might suppose that the rate at which new cases occur increases exponentially with time. An appropriate model for such phenomena might be one of the following form:

$$\mu_i = c\exp(bt_i) \tag{15}$$

where μ_i is the expected number of new cases on day t_i, and c and b are unknown parameters.

Now, using a link function of the form $g = \log$, by taking the logarithm of both sides of equation 15 we have that

$$\begin{aligned}\log(\mu_i) &= \log(c) + bt_i \\ &= \beta_0 + \beta_1 t_i\end{aligned}$$

where $\beta_0 = \log(c)$ and $\beta_1 = b$.

To determine what distribution we believe our random variables Y_i come from, consider that the number of new individuals who are infected with a disease on a given day is a count. Hence, a random variable that satisfies this requirement is one that possesses a Poisson distribution, which has density $f_{\lambda_i}(x) = \exp(-\lambda_i)\lambda_i^x/x!$. Notice that a Poisson distribution is from an exponential family, since:

$$f_\lambda(x) = \exp(-x\log(\lambda) - \lambda - \log(x!))$$

So, our GLM has a linear predictor in $\beta_0 + \beta_1 t_i$, a log link and a Poisson distribution for the response variables, Y_i.

3.2 Model Fitting and the Iterative Re-weighted Least Squares Method

To fit the GLM we will employ the widely used method of maximum likelihood estimation, which is found in both Wood and McCullagh [47, 31]. After defining our model fitting objective, we will follow an argument in Wood [47] to demonstrate that this objective is equivalent to minimising a non-linear weighted least squares objective (which we shall also definite later), and then discuss the iterative method needed to find a solution to this problem.

Due to the independence of the response variables Y_i , after observing the sample $y_1, y_2, ..., y_n$ we have that the likelihood function of β is given by [47]

$$L(\beta) = \prod_{i=1}^{n} f_{\theta_i}(y_i) \tag{16}$$

So, the log likelihood function of β, which is simply the logarithm of the expression in (16), is given by

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \log(f_{\theta_i}(y_i)) \\ &= \sum_{i=1}^{n} ([y_i\theta_i - b_i(\theta_i)]/a_i(\phi) + c_i(y_i, \phi)) \end{aligned}$$

Notice that the parameters θ_i and functions a_i, b_i and c_i may differ depending on the value of i (since the distributions of the $Y_i's$ are assumed to be from the exponential family, not necessarily identically distributed), however we restrict the scaled parameter, ϕ, to be constant. If we make a further restriction, as alluded to previously, by letting $a_i(\phi) = \phi/\omega_i$, then we can write the log likelihood as

$$l(\beta) = \sum_{i=1}^{n} (\omega_i[y_i\theta_i - b_i(\theta_i)]/\phi + c_i(y, \phi)) \tag{17}$$

Now, in employing the maximum likelihood estimation method we wish to find the vector β such that the likelihood function is maximised (note that maximising the logarithm of the likelihood function is equivalent to maximising the logarithm of the likelihood function, since the logarithm function is monotonic). To achieve this end we take partial derivatives of expression (17) with respect to the elements β_j of the vector β, which yields:

$$\begin{aligned}\frac{\partial l(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j}(\omega_i[y_i\theta_i - b_i(\theta_i)]/\phi + c_i(y,\phi)) \\ &= \frac{1}{\phi}\sum_{i=1}^{n}\left(\omega_i[y_i\frac{\partial \theta_i}{\partial \beta_j} - \frac{\partial b_i(\theta_i)}{\partial \beta_j}]\right) \\ &= \frac{1}{\phi}\sum_{i=1}^{n}\left(\omega_i[y_i\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} - b_i'(\theta_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}]\right)\end{aligned}$$

where the last step follows from the chain rule for partial derivatives. If we now use equation (12) we have that

$$\frac{\partial \mu_i}{\partial \theta_i} = b_i''(\theta)$$

and so by substituting the reciprocal of this expression into our expression for $\partial l(\beta)/\partial \beta_j$ yields

$$\frac{\partial l(\beta)}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i - b_i'(\theta))}{b_i''(\theta_i)/\omega_i}\frac{\partial \mu_i}{\partial \beta_j}$$

Furthermore, by substituting our expressions (12) and (14) into the above equation and letting $\partial l(\beta)/\partial \beta_j = 0$, we have the following result which holds for all j:

$$0 = \frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i - \mu_i)}{V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j} \tag{18}$$

Now, consider the following expression:

$$S = \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{V(\mu_i)} \tag{19}$$

Our expression in (19) is the statement of a non-linear weighted least squares objective: If we suppose that the weights $V(\mu_i)$ are independent of β and known, then one may take partial derivatives of (19) (for each β_j) in order to find the value of β which minimises the weighted squared difference between the observations y_i and the means μ_i. However, it is clear that if we follow such a procedure we derive, for each j, the equation (18), by the chain rule. So, our original model fitting problem is reduced to a non-linear weighted least squares problem.

In order to calculate the vector $\hat{\beta}$ that solves (18), where we consider $V(\mu_i)$ to be constant, we can employ an iterative method, known as *Iteratively Re-weighted Least Squares.*

We firstly let the vector $\hat{\beta^{[k]}}$ be the estimate of β after the kth iteration. Next we define $\zeta^{[k]}$ and $\mu^{[k]}$ to be vectors after the kth iteration with ith elements $\zeta_i^{[k]} = X_i\hat{\beta^{[k]}}$ and $\mu_i^{[k]} = g^{-1}(\zeta_i^{[k]})$, respectively, and g^{-1} is understood to be the inverse of the link function, g.

We iterate the following steps to form a convergent sequence of parameter estimates $\hat{\beta^{[k]}}$, given an initial guess of $\hat{\beta^{[0]}}$.

1. Calculate $V(\mu_i^{[k]}$, given the value for $\hat{\beta^{[k]}}$.

2. Minimise our expression in (19) with respect to β given the calculated value for $V(\mu_i^{[k]}$. Then, update the value of our β estimate by setting $\hat{\beta^{[k+1]}}$ equal to value of β which minimises S in (19).

3. Increment k by 1.

In order to calculate the second step in this iteration we can use another iterative method, known as *iterative linear least squares* [47]. Suppose that we have a model where

$$E(y) = f(\beta)$$

Where f is a non-linear vector valued function. To find an estimate of β that minimises the expression

$$S_2 = \sum_{i=1}^{n}(y_i - f_i(\beta))^2$$

we can use a Taylor series approximation of S_2, supposing that f_i behaves 'nicely'. If we guess that the best fit for our parameters is $\hat{\beta^{[p]}}$, then we proceed by taking a Taylor approximation of S_2 around the vector $\hat{\beta^{[p]}}$, giving us that

$$S_2 \simeq S_2^{[p]} = ||y - f(\hat{\beta^{[p]}}) + J^{[p]}\hat{\beta^{[p]}} - J^{[p]}\beta||^2 \tag{20}$$

where we define J to be the Jacobian matrix with entries $J_{ij}^{[p]} = \partial f_i/\partial\beta_j$, evaluated at $\hat{\beta^{[p]}}$. We can rewrite our expression in (20) to be

$$S_2^{[p]} = ||z^{[p]} - J^{[p]}\beta||^2$$

where we define $z^{[p]}$ to be the so-called pseudodata vector

$$z^{[p]} = y - f(\hat{\beta^{[p]}}) + J^{[p]}\hat{\beta^{[p]}}$$

Thus we have converted a non-linear least squares problem into a linear least squares problem, and through iteration we can construct a convergent sequence of estimates $\hat{\beta^{[p]}}$.

It so happens that this iterative algorithm (that is, the algorithm for finding $\beta^{\hat{[k]}}$ so to minimise S) is somewhat inefficient due to the iteration-within-an-iteration situation. As Wood points out, the iteration method employed in step 2 is a wasteful endeavour before the terms $V(\mu_i^{[k]})$ start to converge [47]. Consequently, Wood suggests that one ought to consider modifying step 2 so that instead of iterating to convergence for every k, we simply make one iteration.

One can express this algorithm in convenient matrix form, by using the following approach. Let $V_{[k]}$ be a matrix such that its diagonal entries are given by the terms $V(\mu_i^{[k]})$, so we can now express (19) as

$$S = ||\sqrt{V_{[k]}^{-1}}[y - \mu(\beta)]||^2$$

Using a Taylor expansion around $\beta^{\hat{[k]}}$ to replace μ we can approximate S by

$$||\sqrt{V_{[k]}^{-1}}[y - \mu^{[k]} - J_{\mu}((\beta) - \beta^{\hat{[k]}}||^2$$

where $J^{\beta^{\hat{[k]}}}$ is the Jacobian matrix with elements $J_{ij}^{\beta^{\hat{[k]}}} = \partial\mu_i/\partial\beta_j$, evaluated at $\beta^{\hat{[k]}}$. From this we can express the terms $J_{ij}^{\beta^{\hat{[k]}}}$ by

$$J_{ij}^{\beta^{\hat{[k]}}} = \partial\mu_i\beta_j|_{\beta^{\hat{[k]}}} = \frac{X_{ij}}{g'(\mu_i^{[k]})}$$

since we have that

$$g(\mu_i) = X_i\beta$$

which implies that

$$g'(\mu_i)\frac{\partial\mu_i}{\partial\beta_j} = X_{ij}$$

We can further express the approximation of S in a more compact form. If we let the diagonal elements of the matrix G equal to $g[(\mu_i^{[k]})$, so that we have $J^{\beta^{\hat{[k]}}} = G^{-1}X$, allowing us to re-write the previous approximation to S by

$$\begin{aligned} S &\simeq ||\sqrt{V_{[k]}^{-1}}G^{-1}[G(y - \mu^{[k]}) + X\beta^{\hat{[k]}} - X\beta||^2 \\ &= ||\sqrt{W^{[k]}}[z^{[k]} - X\beta||^2 \end{aligned}$$

where we use the fact that the pseudodata $z_i^{[k]}$ is equal to $g'(\mu^{[k]})(y_i - \mu_i^{[k]}) + \zeta_i^{[k]}$ and define the weight matrix, W, to have diagonal elements

$$W_{ii}^{[k]} = \frac{1}{v(\mu_i^{[k]})g'(\mu_i^{[k]})^2}$$

Consequently, the IRLS algorithm can be re-expressed in the following way:

1. Compute the pseudodata vector $z^{[k]}$ and the weights matrix $W^{[k]}$ given current values for $\mu^{[k]}$ and $\zeta^{[k]}$.
2. Minimise the expression $||\sqrt{W^{[k]}}[z^{[k]} - X\beta||^2$ with respect to β so to obtain $\beta^{\widehat{[k+1]}}$, which then allows us to update the vectors $\zeta^{[k+1]}$ and $\mu^{[k+1]}$.
3. Increment k by 1.

It also turns out that if we let W and z denote the limits of the convergent sequences $W^{[k]}$ and $z^{[k]}$, respectively, then we have that

$$S = -\frac{1}{2\phi}||\sqrt{W}(z - X\beta)||^2$$

is a quadratic approximation for the log likelihood of the model around the point $\hat{\beta}$ [47].

Now, it must be remarked that the parameter estimate for β can be obtained without ϕ necessarily being known. However, one may wish to estimate ϕ nonetheless; this can be achieved by using the so-called *deviance* of the model.

The deviance can be thought of as an analogue to the expectation of the residual sum of squares that features in linear statistical models [15], and is defined as

$$D(\beta) = 2[l(\hat{\beta}_s) - l(\beta)]\phi$$

where $l(\beta)$ is the likelihood corresponding to the vector β, whereas $l(\hat{\beta}_s)$ is the maximized likelihood obtained for the saturated model; that is, the model which has a parameter for each data point. Necessarily, $l(\hat{\beta}_s)$ must be greatest amongst *all* possible values of β. An estimator for ϕ is given by

$$\hat{\phi_D} = \frac{D(\beta)}{n - p}$$

where n is the number of observations and p is the amount of identifiable parameters in the model [47].

4 Generalised Additive Models: An Introduction

We define a *generalised additive model* for response variables $y_1, ..., y_i, ..., y_n$ to be a model of the form

$$g(\mu_i) = X_i^* \theta + f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + f_3(x_{4i}) + ... \tag{21}$$

where g is a (known) 'link' function that is monotonic and twice differentiable, and $\mu_i = E(Y_i)$, with Y_i having distribution from an exponential family. Moreover, X_i^* is the ith row of a model matrix for any strictly parametric model components, θ, while f_j's are smooth functions of the covariates, x_j [47].

One can have the the smooth functions effectively "absorb" the linear component $X_i^* \theta$ in the model, so that our expression in (39) is given by

$$g(\mu_i) = h_1(x_{1i}) + h_2(x_{2i}, x_{3i}) + h_3(x_{4i}) + ... \tag{22}$$

Note that the GAM is an extension of the generalised linear model previously discussed, where the smooth functions act as an extension of the linear predictors in the GLM. As is noted in Hastie et al, these smooth functions (also known as smooths) earn their 'smooth' title from their ability to estimate the relationship between $g(E(Y_i))$ and the covariates, with these estimates being less variable than $g(E(Y_i))$ itself [**?**]. The smooths do not adhere to a particular rigid form, but are rather nonparametric in nature.

An immediately observed advantage of the GAM is that it can provide a superior model fit than the cruder GLM. As a consequence of this greater generality, the GAM can provide a more satisfactory model of real world phenomena than a simple GLM, as is mentioned in Wood [47].

However, there are drawbacks to using the GAM: one must choose some appropriate way of representing these smooth functions, while the degree of smoothness of these functions must also be chosen. Consequently, a certain degree of subjectivity has been introduced into the simpler model GLM. The following sections aims to provide some help towards finding satisfactory solutions to these two requirements of the model.

5 Smooth Functions

In this section we concern ourselves with finding a satisfactory way of representing the smooth functions in our GAM. The general method that is used for representing these smooth functions is by expressing them as linear combinations of *basis functions.* In which case, in the model fitting procedure our task will become that of estimating the respective coefficients of the selected basis functions. As might be guessed, the question of how to choose appropriate basis functions will be central to the discussion that follows.

Taking the lead from Wood [47] we will simplify matters somewhat by considering a less sophisticated model than the GAM so to make our discussion of smooths as clear as possible. Suppose that we have a set of response variables $y_1, ..., y_n$ where the ith response can be represented in the following way

$$y_i = f(x_i) + \epsilon_i \tag{23}$$

and we further suppose that the error terms ϵ_i are i.i.d. normal random variables with mean 0 and variance σ^2. Suppose f has a basis representation in terms of basis functions $b_1, ..., b_m$. Then, we have that

$$f(x) = \sum_{j=1}^{m} \beta_j b_j(x) \tag{24}$$

We will consider two different possible types of basis functions given this scenario: the polynomial basis, and the cubic spline basis.

A *polynomial basis* is a collection of functions of the form $b_1(x) = 1$, $b_2(x) = x$, ..., $b_m(x) = x^{m-1}$. This seems to be a natural starting point for considering a suitable basis for f; from Taylor's theorem we know that continuous functions can be well approximated *locally* by linear combinations of polynomials [20]. So we might suppose that for a given smooth, an appropriate choice of basis might be a polynomial basis. However, such bases tend to perform not as well globally [47].

An alternative to the polynomial basis that proves to be slightly more appealing is a cubic spline basis. A cubic spline is a continuous function (and continuous in first and second derivatives) that is constructed from joining cubic functions together at chosen *knots* (which are usually located at the data points, or equidistant to each other). As will be proved soon, such splines are, in a sense, optimally smooth. Such cubic splines can be represented as a linear combination of other functions; this is what we call a *cubic spline basis.* So, if we suppose that our smooth function is a cubic spline, we can express it as a linear combination of the elements of such a basis.

Remarkably, one can represent the functions in a cubic spline basis in terms of a polynomial in two variables, z and x, where each basis equals this polynomial with the z value set to a particular knot value [13]. An example of such a polynomial representation of the basis functions for a cubic spline is the following given in Wood [47]. If we let the set $K = \{x_i^*, ..., x_{m-2}^*\}$ denote the knot locations, then a cubic spline basis is defined by $b_1(x) = 1$, $b_2(x) = x$ and $b_{i+2}(x) = R(x, x_i^*)$, where

$$R(x,y) = \frac{[(z-\frac{1}{2})^2 - \frac{1}{4}][(x-\frac{1}{2})^2 - \frac{1}{4}]}{4} - \frac{(|z-x| - \frac{1}{2})^4 - \frac{1}{2}(|z-x| - \frac{1}{2})^4 - \frac{7}{240}}{24}$$

Note that for either the polynomial or cubic spline representations for the smooth, we can express the model in (23) as a linear model

$$y = X\beta + \epsilon \tag{25}$$

where the model matrix X has columns equal to the number of basis functions, y is a vector of responses and ϵ is a vector of residuals. Thus, for the polynomial basis case we have that the ith row of the model matrix is given by the $1 \times m$ vector

$$X_i = [1, x_i, ..., x_i^{m-1}]$$

whereas for the cubic spline basis case it is given by

$$X_i = [1, x_i, R(x_i, x_1^*), ..., R(x_i, x_{m-2}^*)]$$

The parameters for these linear models can then be estimated through standard least squares regression [**?**]. Thus, for the case of the cubic spline basis representation, we call the resulting basis functions *regression splines* [47]. At this point it is appropriate to discuss why the cubic spline is considered a desirable smooth, before we discuss an important extension of regression splines.

5.1 Properties of Smoothing Splines

Given a set of points (x_i, y_i) where $i = 1, ..., n$ and $x_i \leqslant x_{i+1}$, we say that the *natural cubic spline* [47] interpolating these points is defined to be a continuous function $g : [x_1, x_n] \to \Re$ where over each interval $[x_i, x_{i+1}]$ the function is a cubic polynomial, $g(x_i) = y_i$, with continuous first and second derivatives and $g''(x_1) = g''(x_n) = 0$. This definition enables us to prove the following important result:

Theorem 1. *Suppose that f is a continuous function on $[x_1, x_n]$ which interpolates the set of points (x_i, y_i) and possesses an absolutely continuous first derivative. Then,*

$$J(f) = \int_{x_1}^{x_n} f''(x)^2 dx \geqslant \int_{x_1}^{x_n} g''(x)^2 dx$$

where $J(f)$ is a measure of "smoothness" of the interpolating function.

In other words, the natural cubic spline $g(x)$ is the smoothest of all such functions. The following is a proof given by Green and Silverman. [12]

Proof. Firstly, we suppose that $f \neq g$ is an interpolating function of the points (x_i, y_i). Let us define $h(x) = f(x) - g(x)$. Now

$$\begin{aligned}\int_{x_1}^{x_n} f''(x)^2 dx &= \int_{x_1}^{x_n} (h''(x) + g''(x))^2 dx \\ &= \int_{x_1}^{x_n} h''(x)^2 dx + \int_{x_1}^{x_n} 2h''(x)g''(x) dx + \int_{x_1}^{x_n} g''(x)^2 dx \qquad (26)\end{aligned}$$

Now, consider the term $\int_{x_1}^{x_n} h''(x)g''(x)dx$. For the following calculation of this integral, note that since g is cubic over any open interval (x_i, x_{i+1}) its second derivative must be constant. By the method of integration by parts we have that

$$\begin{aligned}\int_{x_1}^{x_n} h''(x)g''(x)dx &= [h'(x)g''(x)]_{x_1}^{x_n} - \int_{x_1}^{x_n} h'(x)g'''(x)dx \\ &= -\int_{x_1}^{x_n} h'(x)g'''(x)dx \qquad \text{since} \qquad g''(x_1) = g''(x_n) = 0 \\ &= -\sum_{i=1}^{n-1} g'''(x_i^+) \int_{x_i}^{x_{i+1}} h'(x)dx \\ &= -\sum_{i=1}^{n-1} g'''(x_i^+)[h(x_{i+1}) - h(x_i)] \\ &= 0 \qquad \text{since} \qquad h(x_i) = f(x_i) - g(x_i) = 0, \forall i\end{aligned}$$

Thus, the middle term of equation (26) vanishes, and we have the following result

$$\int_{x_1}^{x_n} f''(x)^2 dx = \int_{x_1}^{x_n} h''(x)^2 dx + \int_{x_1}^{x_n} g''(x)^2 dx \geqslant \int_{x_1}^{x_n} g''(x)^2 dx$$

Now, equality holds if and only if $h''(x) = 0$ over the open interval (x_1, x_n). Now, given that $h''(x) = 0$ for $x \in (x_1, x_n)$ we must have that $h'(x) = c$, where c is some constant. But, since $h(x_1) = h(x_n) = 0$, we must have that $c = 0$, and so $h(x) = 0, \forall x \in [x_1, x_n]$. Hence, our equality must hold if and only if $h(x) = 0$ over the closed interval $[x_1, x_n]$. □

The above case of a natural cubic spline is quite restrictive; it may be that instead of interpolating the points (x_i, y_i) we would rather have the values $g(x_i)$ of the function g to be unconstrained. We can modify our objective so that g, though not interpolating the specific set of points, is still a sufficiently close fit to the data, while still smooth. One way of achieving this is to estimate the values $g(x_i)$ with the objective of minimising the following expression:

$$\sum_{i=1}^{n} [y_i - g(x_i)]^2 + \lambda \int_{x_1}^{x_n} g''(x)^2 dx$$

where the integral term can be considered as being a roughness penalty, while λ is a term which weights the relative importance of model fitting against model smoothness (later, we will see how one can choose an appropriate λ for this task). The solution to this problem, g, is called a *smoothing spline* [39]. From this definition we have a most remarkable result.

Theorem 2. *Amongst all functions f, the smoothing spline g minimises the expression*

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_{x_1}^{x_n} f''(x)^2 dx \tag{27}$$

Proof. A proof for this result goes as follows. Suppose that there exists another function, h, such that h minimizes expression (27). Then, we can use a cubic spline, g, such that it interpolates the points $(x_i, h(x_i))$. Consequently, g and h have the same sum of squared errors term in expression (27); but, since g is a cubic spline, by our previous result we must have that $\lambda \int g''(x)^2 dx \leq \lambda \int h''(x)^2 dx$, hence g indeed minimizes (27). □

It should be stressed that for the objective of minimising expression (27) at no point do we need to specify a basis for the spline; finding values of the points $g(x_i)$ that minimise (27) is comprehensive enough to determine the resulting smoothing spline. This can easily be observed in the case of the interpolating spline: due to the $4(n-1)$ constraints imposed by continuity of (first and second) derivatives and the function at specified knot points, along with the vanishing second derivatives at the end points, one can fit the $n-1$ cubic polynomials that comprise the cubic spline (as four constraints completely determines the coefficients of a cubic). In the case of minimising expression (27) the only difference is that the values that g takes at the knot values (which are the data points in this case) must be estimated, rather than being set equal to the responses y_i.

5.2 Penalised Regression Splines

The drawback of the approach previously outlined is that since the number of parameters to be estimated (the terms $g(x_i)$) equals the number of data points, the model is somewhat computationally inefficient. This inefficiency happens to not be so problematic in the case where the smooth is univariate (i.e. when f is a function of just one covariate), however for the case of multivariate smooths the computations can become quite expensive [47].

The task of minimising expression (27) in the proof can be modified somewhat so that instead of estimating the free terms $g(x_i)$, we suppose our cubic spline can be expressed in terms of a cubic spline basis such as in our previously mentioned example. Such an objective results in what is called a *penalised regression spline* [**?**]. Now, the expression we wish to minimise becomes

$$\sum_{i=1}^{n}[y_i - X_i\beta]^2 + \lambda \int_{x_1}^{x_n} f''(x)dx \tag{28}$$

where the model matrix X_i is a row of the cubic spline basis functions evaluated at the datum x_i.

It happens that we can express the penalty term as

$$\int f''(x)^2 dx = \beta^T S \beta$$

where S is some matrix of known coefficients. This can be shown since if we twice differentiate the basis functions we have that

$$f''(x) = \beta^T [b_1''(x), ..., b_m''(x)]^T$$

and since a scalar is its own transpose, this implies that

$$\begin{aligned} \int f''(x)^2 dx &= \int \beta^T [b_1''(x), ..., b_m''(x)]^T [b_1''(x), ..., b_m''(x)] \beta \\ &= \beta^T S \beta \end{aligned}$$

where $S = \int [b_1''(x), ..., b_m''(x)]^T [b_1''(x), ..., b_m''(x)] dx$

Thus, the regression spline problem is to minimise the expression

$$\sum_{i=1}^{n} [y_i - X_i \beta]^2 + \lambda \beta^T S \beta \tag{29}$$

This conveniently leads to a straightforward way to estimating the coefficients of the basis functions. Denoting expression (29) by $S(\beta)$ we have that

$$\begin{aligned} S(\beta) &= (y - X\beta)^T (y - X\beta) + \lambda \beta^T S \beta \\ &= y^T y - 2\beta^T X^T y + \beta^T (X^T X + \lambda S) \beta \end{aligned} \tag{30}$$

Differentiating 30 with respect to β and setting to zero yields the equation

$$(X^T X + \lambda S) \hat{\beta} = X^T y$$

the solution for which is given by

$$\hat{\beta} = (X^T X + \lambda S)^{-1} X^T y \tag{31}$$

The influence matrix (also known as the 'hat' matrix) is defined to be the matrix A such that $Ay = \hat{\mu}$. Since $\hat{\mu} = X\hat{\beta}$, we must have that

$$A = X(X^T X + \lambda S)^{-1} X^T y$$

However, in solving for $\hat{\beta}$ we would not use the expression given in (31) as it is computationally inefficient to calculate the matrix inverse in the expression. In any case, since we are concerned with the GAM rather than this much simpler model, we do not need to trouble ourselves with estimation in this scenario.

5.3 A Further Comment on Smoothness

Before we proceed further, a certain matter must be addressed regarding the use of a roughness penalty for smoothing. That is, it must be made clear that the use of a roughness penalty $\lambda \int f''(x)^2 dx$ is not the only way to control the level of smoothness: one could also consider comparing models that have different knot choices. Such an approach has its pitfalls, however, since a model with k-1 many evenly spaced knots will not be typically nested within a model that has k many knots evenly spaced knots, consequently making methods of model selection such as backwards selection unfeasible [47]. Furthermore, comparing models that have different numbers of knots, but where the knot locations are fixed, provides an unsatisfactory form of model selection since such a method usually results in uneven distances between knots, something which itself can lead to poor model fitting (especially since it happens that the fit of penalised regression splines is often very dependent on the choice of knot locations) [47].

Having settled on the roughness penalty approach as a means of ensuring our smooth functions are indeed smooth, the key task is to determine an appropriate way to choose the parameter λ which weights the tradeoff between model fitting and smoothness. Note that λ will be chosen between two extremes; when $\lambda = 0$ the penalty disappears and our spline will interpolate the points as in the theorem (this results in over fitting); when $\lambda = \infty$ in order to minimise the expression (28) we must have that $f''(x) = 0$ for all x, and so our smooth function must be linear. We will discuss later how one is to choose an appropriate λ to balance the need for model fitting and smoothness.

6 Choice of Spline Basis

In our previous discussion of the advantage of using a cubic spline basis as a representation for the smooths in our GAM, we were mute about the different possible choices of such bases. In this section we wish to address this question by exploring different types of cubic spline bases and commenting on their respective advantages and disadvantages.

6.1 A Knot Parameterisation Method

Firstly, suppose we wish to define a cubic spline basis for a smooth $f(x)$ where we specify the knots $x_1, ..., x_k$. Moreover, we will further suppose that f has continuous first and second derivatives, while setting the second derivatives equal to zero at the boundary knots x_1 and x_k. The important feature in this model is how we define the parameters to be estimated: we let $\beta_j = f(x_j)$ and $\delta_j = f''(x_j)$. For notational convenience we let $h_j := x_{j+1} - x_j$ before we define the following basis functions:

$$\begin{aligned} a_j^-(x) &= (x_{j+1} - x)/h_j & a_j^+(x) &= (x - x_j)/h_j \\ c_j^-(x) &= [(x_{j+1} - x)^3/h_j - h_j(x_{j+1} - x)]/6 & c_j^+(x) &= [(x - x_j)^3/h_j - h_j(x - x_j)]/6 \end{aligned}$$

Now, we define the smooth itself to be

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \tag{32}$$

where $x \in [x_j, x_{j+1}]$. Our expression in (32) is simply the cubic that lies between the knots x_j and x_{j+1}.

It turns out that one can represent the smooth completely in terms of the parameters β through the following argument found in Wood [47].

Given a knot x_m (assume it is not a boundary knot, however) consider the two cubic functions that join together at the knot x_m, $f(x_l)$ and $f(x_r)$, where x_l and x_r are points of f to the left and right of the knot x_m, respectively (i.e. $x_l \in [x_{m-1}, x_m]$ and $x_r \in [x_m, x_{m+1}]$). Then, by the continuity of the derivative of f we can set the derivatives of the two cubics equal to each other so we yield

$$-\frac{\beta_j}{h_j} + \beta\beta_{j+1}h_j + \delta_{j+1}\frac{h_j}{6} + \delta_{j+1}\frac{3h_j}{6} - \delta_{j+1}\frac{h_j}{6} = -\frac{\beta_{j+1}}{h_{j+1}} + \frac{\beta_{j+2}}{h_{j+1}} - \delta_{j+1}\frac{3h_j}{6} + \delta_{j+1}\frac{h_{j+1}}{6} - \delta_{j+2}\frac{h_{j+1}}{6}$$

which is equivalent to

$$\beta_j\frac{1}{h_j} - \beta_{j+1}\left(\frac{1}{h_j} + \frac{1}{h_{j+1}}\right) + \beta_{j+2}\frac{1}{h_{j+1}} = \delta_j\frac{h_j}{6} + \delta_{j+1}\left(\frac{h_j}{3} + \frac{h_{j+1}}{3}\right) + \delta_{j+2}\frac{h_{j+1}}{6}$$

We can repeat this for all the joining knots as we index over $j = 1, ..., k-2$ and use the border conditions that $\delta_1 = \delta_k = 0$ (since $f''(x_1) = f''(x_k) = 0$), giving us a system of equations which can be expressed as the following matrix equation

$$B\delta^{-} = D\beta \tag{33}$$

where we let $\delta^{-} = (\delta_2, ..., \delta_{k-1})^T$ while the matrices B and D have, for $i = 1, ..., k-2$, the entries

$$D_{i,i} = \frac{1}{h_i} \qquad D_{i,i+1} = -\frac{1}{h_i} - \frac{1}{h_{i+1}}$$

$$D_{i,i+2} = -\frac{1}{h_{i+1}} \qquad B_{i,i} = \frac{h_i + h_{i+1}}{3}$$

and for $i = 1, ..., k-3$ the entries

$$B_{i,i+1} = \frac{h_{i+1}}{6} \qquad B_{i+1,i} = \frac{h_{i+1}}{6}$$

From equation (33) we can define $F^{-} = B^{-1}D$. Then, defining F to be the matrix

$$\begin{bmatrix} 0 \\ F^{-} \\ 0 \end{bmatrix}$$

where 0 is simply a row of zeros, we arrive at the equation $\delta = F\beta$, and so we may express f as

$$\begin{aligned} f(x) &= [a_j^{-}(x)\beta_j + a_j^{+}(x)\beta_{j+1} + c_j^{-}(x)F_j\beta + c_j^{+}(x)F_{j+1}\beta]\mathbf{1}_{x\in[x_j,x_{j+1}]} \\ &= \sum_{i=1}^{n} b_i(x)\mathbf{1}_{x\in[x_j,x_{j+1}]} \end{aligned}$$

where $\mathbf{1}_{x\in[x_j,x_{j+1}]}$ is the indicator function, F_j denotes the jth row of F and the $b_i(x)$ are basis functions which result from the new expression of f purely in terms of the coefficients of β.

It also happens to be the case, from Wood [47], that

$$\int_{x_1}^{x_k} f''(x)^2 dx = \beta^T D^T B^{-1} D\beta$$

Thus, one can readily incorporate such splines into a GAM framework with $S = D^T B^{-1} D$ for the roughness penalty.

The main advantage of using this kind of representation is that we can interpret the parameters in terms of their relationship to the smooth function. Its main disadvantage is that we must choose the knot locations.

6.2 B-Splines and P-Splines

An alternative to the previous spline basis choice is that of the B-spline. In this case, the basis is comprised of functions that are only non-zero over a certain region of the space. This region of space that each basis function is non-zero over is set to be the intervals between $m+3$ adjacent knots, where $m+1$ is the order of the basis [4], which means for a cubic spline we have that $m = 2$.

In order to construct a B-spline basis with k parameters for a given smooth, we must choose knots $x_1, ..., x_{k+m+1}$. Moreover, due to the restrictions placed on the region for which the splines are non-zero, we choose the knots so that the spline is to be evaluated only between knots x_{m+2} and x_k.

By defining the following basis functions recursively (where we keep m general)

$$B_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m+2} - x_{i+1}} B_{i+1}^{m-1}(x) \;\; i = 1, ..., k$$

with

$$B_i^{-1}(x) = \mathbf{1}_{x \in [x_i, x_{i+1}]}$$

The importance of B-splines is their penalised extension, known as P-splines. In the P-spline case a roughness penalty is introduced where instead of taking an integral of the squared double derivative of the smooth function, the penalty operates directly on the parameters β_i in the form of a difference penalty. The penalty, denoted by P, is given by

$$P = \sum_{i=1}^{k-1} (\beta_{i+1} - \beta_i)^2 = \beta_1^2 - 2\beta_1\beta_2 + 2\beta_2^2 - 2\beta_2\beta_3 + ... + \beta_k^2$$

The idea behind this difference penalty is that smoothness can be achieved by ensuring that the difference in scale between consecutive basis functions is controlled by penalising excessive differences in their scalar coefficients (being the parameters β_i). Conveniently, P has the following matrix representation

$$P = \beta^T \begin{bmatrix} 1 & -1 & 0 & \dots \\ -1 & 2 & -1 & \dots \\ 0 & -1 & 2 & \dots \end{bmatrix} \beta$$

The advantage of the P-splines is their simplicity and flexibility. P-splines' major disadvantages are that their justification is found wanting in the case where the knots are not evenly spaced apart (basis functions far apart have their β coefficients penalised to

[4] We allow m to be general, since the B-Spline approach works for smooths that are not necessarily cubic splines.

the same extent as basis functions which are close, which seems unreasonable) while the penalties are less easy to interpret with respect to the properties of the smooth function to be fit than others, such as $\int f''(x)^2 dx$.

6.3 Thin-Plate Splines

Let us consider a slightly modified problem where we suppose that our response y_i is not a function of a scalar, but rather a d-dimensional *vector* $\mathbf{x_i}$, given by the expression

$$y_i = g(\mathbf{x_i}) + \epsilon_i$$

where g is smooth and ϵ_i is an error term.

Now, define the functional J_{md} which represents a roughness penalty to be

$$J_{md}(f) = \int \cdots \int_{R^d} \sum_{\nu_1+\ldots+\nu_d=m} \frac{m!}{\nu_1!\ldots\nu_d!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1}\ldots\partial x_d^{\nu_d}} \right)^2 dx_1 \ldots dx_d \tag{34}$$

Using this expression, we can consider the problem of estimating the function $\hat{f}$ that minimises

$$||\mathbf{y} - \mathbf{f}||^2 + J_{md}(\mathbf{f}) \tag{35}$$

where $\mathbf{f} = (f(\mathbf{x_1}, ..., f(\mathbf{x_n}))^T$, λ is the smoothing parameter (which we will discuss how to choose later) and m is chosen in advance. In fact, m is chosen so that $2m > d$ (in fact, one usually chooses m so that $2m > d + 1$) since otherwise too few derivatives are taken to capture the smoothness of the function f.

Given the fitting objective in (35) it so happens that the estimated function $f(\hat{\mathbf{x}})$ has the representation

$$f(\hat{\mathbf{x}}) = \sum_{i=1}^{n} \delta_i \zeta_{md}(||\mathbf{x} - \mathbf{x_i}||) + \sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x}) \tag{36}$$

where the parameters δ_i and α_i must be estimated [47] [5].

We further have that functions $\phi_i(\mathbf{x})$ are defined to be linearly independent polynomials that span the space of polynomials in R^d. These functions are important as they constrain the parameter values of the vector δ through the equation

$$T^T(\delta) = 0$$

[5] The expression we have in (36) is sometimes taken to be the definition of the Thin-Plate spline [12]

where T is the matrix with entries $T_{ij} = \phi_j(x_i)$.

The functions ζ_{md} are defined by the rather ungainly formulae

$$\zeta_{md}(r) = \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d} \log(r) \qquad \text{(for d even)}$$

$$\zeta_{md}(r) = \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} \log(r) \qquad \text{(for d odd)}$$

where Γ is the well-known gamma function.

One can now change the form of the the fitting objective given in (35) by simply defining the matrix $\mathbf{E}$ with entries $E_{ij} = \zeta_{md}(||\mathbf{x_i} - \mathbf{x_j}||)$ so that we now wish to minimise the expression

$$||\mathbf{y} - E\delta \; T\alpha||^T + \lambda\delta^T E\delta$$

with respect to α and δ subject to $T\delta = 0$.

The strength of the thin-plate spline is its ability to address the shortcomings of the other bases described. Specifically, it removes subjectivity in knot choices, and enables one to measure the performance of different choices of bases, while being more useful beyond a univariate context. The disadvantage of such a choice of basis, unsurprisingly, is the computational cost involved, due to there being as many parameters to estimate as data and the computational cost of model estimation is- with only one exception- proportional to the cube of the number of parameters [47]. The need to address this disadvantage motivates the concept of a thin plate regression spline, which we won't discuss here.

7 Representing GAMs as Penalised GLMs

Once we have chosen appropriate basis functions to represent our smooth functions, we must construct a suitable way of representing these functions, as well as the linear part of the model, before we begin estimating the model parameters. Following the method employed in Wood, we seek to represent our GAM as a penalised-GLM, thus allowing us to utilise the GLM previously covered in the estimating procedure [47].

7.1 The Additive Model: A Motivating Example

However, before we delve into the GAM, we will consider an example of an analogous (yet less sophisticated) model. An *additive model* with two univariate smooths is defined so that the responses $y_1, .., y_i, ..., y_n$ have the following relationship with covariates x and z

$$y_i = f_1(x_i) + f_2(z_i) + \epsilon_i \tag{37}$$

with ϵ_i normally distributed with mean 0 and standard deviation σ^2 [47].

The key observation to be made from this model is that there is an identification problem as the functions f_1 and f_2 are only estimable within an additive constant. This is easy to observe since if we define $h_1 := f_1 + \delta$ and $h_2 := f_2 - \delta$ where $\delta \neq 0$ then:

$$h_1(x) + h_2(z) = f_1(x) + f_2(z)$$

meaning that the functions are not uniquely defined unless one imposes a further constraint on the model.

If we suppose that the functions $f_1(x)$ and $f_2(z)$ can be represented as linear combinations of basis functions, with the first basis function constant, then

$$f_1(x) = \nu_1 + \sum_{j=2}^{m_1-1} \nu_j b_j(x)$$

and

$$f_2(z) = \gamma_1 + \sum_{j=2}^{m_2-1} \gamma_j c_j(z)$$

By setting the intercept term for the function $f_2(z)$ equal to zero (i.e. $\gamma_1 = 0$) we can remove the identification problem. Following this added constraint we can express (37) as a linear model

$$y_i = X_i\beta + \epsilon_i$$

where $X_i = [1, b_1(x_i), ..., b_{m-1}(x_i), c_2(z_i), ..., c_{m_2-1}(z_i)]$ and $\beta = [\nu_1, \nu_2, ..., \nu_{m_1-1}, \gamma_2, ..., \gamma_{m_2-1}]^T$

If the basis functions chosen for each smooth are cubic, then one can set up a penalised regression spline by adding roughness penalties $\int f_1''(x)^2 dx$ and $\int f_2''(x)^2 dx$ to least squares score $||y - X\beta||^2$, thus meaning that our model fitting objective is to estimate the parameter β

$$\sum_{i=1}^{n} [y_i - X_i\beta]^2 + \lambda_1 \beta^T S_1 + \lambda_2 \beta^T S_2 \beta \tag{38}$$

where we have that

$$\int f_1''(x)^2 dx = \beta^T S_1 \beta$$

and

$$\int f_2''(x)^2 dx = \beta^T S_2 \beta$$

In the GAM (a generalisation of the additive model) we will demonstrate that one can establish a model fitting objective that is analogous to that described in (38). However, in this more general setting instead of being represented as a linear model, the GAM can be manipulated into the form of a GLM.

7.2 GAM as a GLM

As explained in an earlier section, a GAM model is a model of the form

$$g(\mu_i) = X_i^* \theta + f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + f_3(x_{4i}) + ... \tag{39}$$

where g is a (known) 'link' function that is monotonic and twice differentiable, and $\mu_i = E(Y_i)$, with Y_i having distribution from an exponential family. Moreover, X_i^* is the ith row of a model matrix for any strictly parametric model components, θ, while f_j's are smooth functions of the covariates, x_j (note that x_j could be a vector, such as $x_2 := (x_{2i}, x_{3i})$ defined as the input for the smooth f_2 in our example in (39)) [47].

For tractability and notational convenience (which will be apparent later) we will just consider the GAM model with smoothers that have scalar inputs, which allows us to express our model in the slightly simpler form

$$g(\mu_i) = X_i^* \theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + ... + f_k(x_{ki}) \tag{40}$$

where k is the number of covariates x_j.

As discussed previously, the introduction of the smooth functions f_j forces us to to define basis functions for representing these smooths, and to choose corresponding roughness penalties. If we focus on the first task, let us suppose that for each j there exist q_j basis functions for f_j, which are given by $b_{j1}, ..., b_{jq_j}$, so that we have

$$f_j(x_{ji}) = \sum_{m=1}^{q_j} \beta_{jm} b_{jm}(x_{ji}) \tag{41}$$

where β_{jm} are coefficients of the basis functions b_{jm} that must be estimated.

With this choice of basis for the jth smooth, we can now construct a model matrix corresponding to this smooth, $\tilde{X}_j$, which has the form

$$\begin{bmatrix} b_{j1}(x_{j1}) & \dots & b_{jq_j}(x_{j1}) \\ \vdots & \dots & \vdots \\ b_{j1}(x_{jn}) & \dots & b_{jq_j}(x_{jn}) \end{bmatrix}$$

so that the following expression holds:

$$\mathbf{f}_j = \tilde{X}_j \tilde{\beta}_j \tag{42}$$

where we define $\mathbf{f}_j$ to be the vector with ith entries $f_j(x_{ji})$, and $\tilde{\beta}_j$ to be the vector $[\beta_{j1}, ..., \beta_{jq_j}]$

Before we set up the complete model for *all* smooths, we must deal with the possible case that our model in (39) is not identifiable. One way of addressing this problem is by constraining the values of the vector $\mathbf{f}_j$ so that the sum of its elements equals 0. Consequently, we have that

$$\mathbf{1}^T \tilde{X}_j \tilde{\beta}_j = 0 \tag{43}$$

where $\mathbf{1}$ denotes an $n \times 1$ vector of 1's.

To incorporate this constraint into the model we can use a certain method, demonstrated in Wood [47]. For a given $m \times p$ matrix C and $p \times 1$ parameter vector α ($\mathbf{1}^T \tilde{X}_j$ and $\tilde{\beta}_j$ in our case, respectively), suppose that

$$C\alpha = 0$$

We call C a *constraint matrix* [47].

If we desire to rewrite our model in terms of $p - m$ unconstrained parameters then we can perform what is called a QR decomposition of the matrix C^T, which proceeds in the following way. We can re-express the transpose of the constraint matrix C as

$$C^T = U \begin{pmatrix} P \\ 0 \end{pmatrix}$$

where P is an $m \times m$ upper triangular matrix, and U is a $p \times p$ orthogonal matrix. This decomposition is possible for *any* real matrix, not just constraint matrices [47]. U can be expressed as $U = (D : Z)$, where Z is a $p \times (p-m)$ matrix. This expression is particularly useful, as one can observe that if we let

$$\alpha = Z\gamma$$

where γ is any $p - m$ dimensional vector, then by the orthogonality of the matrices D and Z we have that

$$C\alpha = \begin{pmatrix} P^T & 0 \end{pmatrix} \begin{pmatrix} D^T \\ Z^T \end{pmatrix} Z\gamma = \begin{pmatrix} P^T & 0 \end{pmatrix} \begin{pmatrix} 0 \\ I_{p-m} \end{pmatrix} \gamma$$

where it can be easily observed that this product equals zero, no matter the value of γ.

Our task is to find a matrix Z such that it has $q_j - 1$ orthogonal columns and satisfies

$$\mathbf{1}^T \tilde{X}_j Z = 0$$

From this, one can reparameterise $q_j - 1$ of the coefficients of the basis of functions corresponding with the jth smooth function so that we have a new $(q_j - 1) \times 1$ vector β_j such that

$$\tilde{\beta}_j = Z\beta_j$$

and so by the property of Z, β_j automatically satisfies the constraint given in (43). One can define a new model matrix corresponding to the jth smooth function

$$X_j := \tilde{X}_j Z$$

so that the constraint is incorporated into the model. It must be remarked at this juncture that the matrix Z does not need to be found, explicitly; such a matrix is formed as a by-product of the QR decomposition previously stated.

As a result of this new parameterisation of the vector to be estimated,β_j, and the new model matrix X_j corresponding to the jth smooth, we now have that our GAM model can be represented as

$$g(\mu_i) = X_i \beta \tag{44}$$

for the ith response, where we define $X = [X^*|X_1|X_2|...]$ and $\beta^T = [\theta^T, \beta_1^T, \beta_2^T, ...]$, X^* corresponding to the purely parametric model components represented in θ.

The model stated in expression (44) is a GLM, as promised. Consequently, we can use our previous work regarding GLMs to fit this model, which we will be our endeavour in the next section.

8 Fitting the GAM

Since the model stated in the expression (44) of the previous section is a GLM, we can construct its likelihood function, $l(\beta)$. From here, one could fit the model using maximum likelihood estimation outlined in the GLM section. However, there is a significant danger of over fitting if each of the smooths have several basis functions for their representation. Consequently, we introduce roughness penalties into the model estimation procedure, just as was done in the additive model case.

To represent the roughness penalty corresponding to the jth smooth we can use a quadratic form such as $\beta_j^T \tilde{S}_j \beta_j$, $\tilde{S}_j$ being a matrix of known coefficients. We can then re-write this in terms of the full coefficient vector, β, by letting S_j be the matrix $\tilde{S}_j$ with entries set to zero so that the following expression holds

$$\beta^T \tilde{S}_j \beta = \beta^T S_j \beta$$

With such quadratic forms set for each of the respective smooths one can define a penalised likelihood function

$$l_p(\beta) = l(\beta) - \frac{1}{2} \sum_j \lambda_j \beta^T S_j \beta \tag{45}$$

Now, it is convenient notationally to define the matrix $S = \sum_j \lambda_j \beta^T S_j \beta$. Then, if we maximise the expression (45) by taking its partial derivative with respect to the parameters β_j and setting them to zero, we yield for each j

$$\begin{aligned} 0 &= \frac{\partial l_p}{\partial \beta_j} \\ &= \frac{\partial l}{\partial \beta_j} - [SB]_j \\ &= \frac{1}{\phi} \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} - [SB]_j \end{aligned}$$

where we denote the jth row of the matrix SB by $[SB]_j$.

However, the last expression can easily be seen to be the equation that is obtained when we seek to minimise the penalised non-linear least squares problem given by

$$S_p = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{Var(Y_i)} + \beta^T S \beta$$

where we suppose that the terms $Var(Y_i)$ are known.

This can be found to be approximate to

$$||\sqrt{W^{[k]}}(z^{[k]} - X\beta)||^2 + \beta^T S \beta \tag{46}$$

using arguments analogous to those used in the case of the GLM, where $W^{[k]}$ is a weight matrix and $z^{[k]}$ is a pseudodata vector (see GLMs section).

Thus we have justification for what is called the *Penalised-Iteratively Re-weighted Least Squares*, which has the following algorithm analogous to that of the IRLS

1. Compute $z^{[k]}$ and $W^{[k]}$ given the current value for $\hat{\beta^{[k]}}$.

2. Minimise our expression in (46) with respect to β to find $\hat{\beta^{[k+1]}}$.

3. Increment k by 1.

9 Estimating the Smoothing Parameter

In this section we focus on the task of choosing the smoothing parameter, λ, for a given smooth function. As Green and Silverman explain [12], there are two opposing philosophical views about how one considers the problem of choosing the smoothing parameter. One point of view considers the choice of λ to be at the convenience of the statistician; one can adjust the smoothing parameter to 'explore' the different possibilities of the model, while settling on a sensible λ if one desires to make a prediction from the data. From this approach, the choice of λ is subjective.

However, there is a contrary philosophy; that there must be a definite method from which the smoothing parameter is determined. From this point of view, λ is chosen by the data, given an appropriate method for calculating it. Green and Silverman caution against calling this an 'objective', since there is no objective way of selecting the method which then determines the smoothing parameter [12]. However, we can discuss what are the possible appropriate methods of calculating λ without necessarily settling on a 'best' method; indeed, throughout this section this is precisely what we intend to do.

To calculate λ we have to consider two different scenarios: when the scale parameter ϕ is known, and when it is unknown. We will give an exposition of the methods used under these different scenarios, under the initial assumption that g is the identity map (later, we will consider what happens in the more general case where g is permitted to be other than the identity).

9.1 Un-Biased Risk Estimator (UBRE)

In this scenario we consider the problem of estimating λ when ϕ is known. This scenario arises when the random variable, Y_i, is binomial or poisson distributed (in which case, ϕ is equal to 1), but is not the case, for instance, when Y_i has a normal or gamma distribution [47].

INCOMPLETE MUST – RE – STATE – MODEL

One possible method to select the parameter under this scheme is to minimize an estimate of the expected mean squared error (MSE), also known as the un-biased risk estimator: this is the *un-biased risk estimator method*[6]. Now, the expected MSE is given by the following expression, where M denotes the mean squared error and A is the influence matrix [47].

$$\begin{aligned} E(M) &= \frac{E(||\mu - X\hat{\beta}||^2)}{n} \\ &= \frac{E(||y - Ay||^2)}{n} - \sigma^2 + \frac{2tr(A)\sigma^2}{n} \end{aligned}$$

The last equality is not obvious and holds by the following argument

$$
\begin{aligned}
||\mu - X\hat{\beta}||^2 &= ||\mu - Ay||^2 \\
&= ||\mu + \epsilon - Ay - \epsilon||^2 \\
&= ||y - Ay - \epsilon||^2 \\
&= ||y - Ay||^2 + (\epsilon^T\epsilon) - 2\epsilon^T(y - Ay) \\
&= ||y - Ay||^2 + (\epsilon^T\epsilon) - 2\epsilon^T(\mu + \epsilon) + 2\epsilon^T A(\mu + \epsilon) \\
&= ||y - Ay||^2 - (\epsilon^T\epsilon) - 2\epsilon^T\mu + 2\epsilon^T A\mu + 2\epsilon^T A\epsilon
\end{aligned}
$$

Whence, after taking expectations and dividing by n, and using the facts that $E(\epsilon^T\epsilon) = E(\sum_i \epsilon_i^2) = n\sigma^2$, $E(\epsilon^T) = 0$, and the trick that

$$
\begin{aligned}
E(\epsilon^T A\epsilon) &= E(tr(\epsilon^T A\epsilon)) && \text{(the trace of a scalar equals itself)} \\
&= E(tr(A\epsilon\epsilon^T)) \\
&= tr(AE(\epsilon\epsilon^T)) \\
&= tr(AI\sigma^2) \\
&= tr(A)\sigma^2
\end{aligned}
$$

we yield the desired result. Since we wish to estimate this expression, we can use the following as an estimate:

$$\nu_u(A) = \frac{||y - Ay||^2}{n} - \sigma^2 + \frac{2tr(A)\sigma^2}{n}$$

This is called the Un-Biased Risk Estimator (UBRE), as it is indeed an unbiased estimator of the“risk function” when this function is defined to be the MSE. This estimator is also known as Mallow’s Cp [30].

Thus, our objective, given that σ is known, is to estimate the smoothing parameter(s) in order to minimize the UBRE (note that the smoothing parameter appears in A, the influence matrix)

9.2 Cross Validation

The above approach, while useful for when σ is known, but is not well suited to the problem when σ is unknown (in which case, there is a tendency to over-smooth). Instead of using an estimate of the MSE in order to find the optimal λ, we try to minimize the mean square prediction error; that is, given the fitted model, we minimize the mean square error in predicting a new observation. We have that the expected mean square prediction error, denoted by P, can be computed in the following way (where we suppose that $y^* = X^*\beta + \epsilon^*$).

$$\begin{aligned}
P &= \frac{1}{n}E||y^* - X^*\hat{\beta}||^2 \\
&= \frac{1}{n}E||y^* - \mu + \mu - X^*\hat{\beta}||^2 \\
&= \frac{1}{n}E(||y^* - \mu||^2 + 2(y^* - \mu)^T(\mu - X^*\hat{\beta}) + ||\mu - X^*\hat{\beta}||^2) \\
&= \frac{1}{n}(E||y^* - \mu||^2 + 2E(y^* - \mu)^T(\mu - X^*\hat{\beta}) + E||\mu - X^*\hat{\beta}||^2) \\
&= \frac{1}{n}(E||y^* - \mu||^2 + E||\mu - X^*\hat{\beta}||^2) \qquad \text{(by independence of } y^* \text{ and } \hat{\beta}) \\
&= \sigma^2 + E(M)
\end{aligned}$$

One way to estimate P is by using a method called cross validation. In this setup we aim to minimize what is called the *ordinary cross validation score* (denoted by ν_o) [47], which is given by the following equation:

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu_i}^{[-i]})^2$$

where $\hat{\mu_i}^{[-i]}$ denotes the prediction of $E(y_i)$ using the model which is fitted to all data, except the datum y_i. It can be shown that the expectation of the ordinary cross validation score is approximately equal to P, by the following argument:

$$\begin{aligned}
\nu_o &= \frac{1}{n}\sum_{i=1}^{n}(\mu_i + \epsilon_i - \hat{\mu_i}^{[-i]})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[(\mu_i - \hat{\mu_i}^{[-i]})^2 + 2\epsilon_i(\mu_i - \hat{\mu_i}^{[-i]}) + \epsilon_i^2]
\end{aligned}$$

But, since we have that ϵ_i and $\mu_i - \hat{\mu_i}^{[-i]}$ are independent, and $E(\epsilon_i) = 0$, the expectation of $\epsilon_i(\mu_i - \hat{\mu_i}^{[-i]})$ vanishes, and so we have that

$$\begin{aligned}
E(\nu_o) &= E(\frac{1}{n}\sum_{i=1}^{n}[(\mu_i - \hat{\mu_i}^{[-i]})^2 + \epsilon_i^2]) \\
&= \frac{1}{n}E(\sum_{i=1}^{n}(\mu_i - \hat{\mu_i}^{[-i]})^2) + \sigma^2 \\
&\approx E(M) + \sigma^2
\end{aligned}$$

with the last approximation becoming closer as n increases, since the model fitted to the data other than y_i, $\hat{\mu}^{[-i]}$, becomes closer to the model fitted to all the data, $\hat{\mu}$, since any individual point eventually has a negligible impact on the fitted model.

It should be noted, however, that ordinary cross validation is a desirable method in and of itself, even without the justification arising from the above argument. This is because the method avoids the problem that comes from other methods that judge the ability of the model to predict data that they were fitted to; the problem being that such methods usually result in choosing a more complicated model than is necessary.

Now, since the above method requires fitting the model n many times, this procedure becomes computationally inefficient when n is large. However, there is a way to reduce the total number of model fits to 1, using a neat algebraic trick. That is, if we add zero represented by the term $(\hat{\mu}_i^{[-i]} - \hat{\mu}_i^{[-i]})^2$ to our cross validation score, we get the following expression:

$$\nu_o = \frac{1}{n}\sum_{j=1}^{n}(y_j^* - \hat{\mu}_j^{[-i]})^2$$

where we define $y^* = y - \bar{y}^{[-i]} + \bar{\mu}^{[-i]}$ and $\bar{y}^{[-i]}$ is a vector with ith element being y_i and zero everywhere else, while $\bar{\mu}^{[-i]}$ is a vector with ith entry $\hat{\mu}^{[-i]}$ and zero everywhere else.

Through minimising the cross validation score we fit the model, $\hat{\mu}^{[-i]}$, which not only enables us to find $\hat{\mu}_i^{[-i]}$, but also yields an influence matrix, A. This influence matrix, however, is the same as that obtained when fitting the model to all the data, since the new form of the cross validation score has the same structure as that when fitting all the data. Hence, using A_i to denote the ith row/column of the influence matrix (noting that the influence matrix is symmetric) we have that:

$$\begin{aligned}\hat{\mu}_i^{[-i]} = A_i y^* &= A_i y - A_{ii} y_i + A_{ii}\hat{\mu}_i^{[-i]} \\ &= \hat{\mu}_i - A_{ii} y_i + A_{ii}\hat{\mu}_i^{[-i]}\end{aligned}$$

Now, subtracting y_i from both sides of this expression and then rearranging it results in:

$$y_i - \hat{\mu}_i^{[-i]} = \frac{y_i - \hat{\mu}_i}{1 - A_{ii}} \tag{47}$$

which now allows us to re-express the ordinary cross validation in terms of the model fitted to all the data, and the influence matrix, through the following equation:

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{1 - A_{ii}}\right)^2 \tag{48}$$

which now requires only one fitting step (of the entire sample) in order to calculate it. While it may appear that such a formula should be computationally intensive, given the requirements of performing matrix inversions in order to find the terms A_{ii}, it happens that these terms can actually be found in linear time [12]. Moreover, using what is called the Reinsch algorithm one can find the terms μ_i in linear time [12], which implies that one can compute the ordinary cross validation score itself in linear time.

So, our objective is to minimise the OCV score with respect to the smoothing parameter, λ (recalling that the parameter λ affects ν_0 through the terms A_{ii}). It should be noted, however, that the λ which minimises the ordinary cross validation score may possibly not be unique. [12].

9.3 Generalised Cross Validation

There is, however, a problem with the method of ordinary cross validation. When we consider the fitting objective of the additive model, which is to find the vector β, given smoothing parameters λ_i, that minimise the expression:

$$||y - X\beta||^2 + \sum_{i=1}^{m} \lambda_i \beta^T S_i \beta$$

we would expect that all inferences made about β would be the same if we instead wished to minimise the expression:

$$||Qy - QX\beta||^2 + \sum_{i=1}^{m} \lambda_i \beta^T S_i \beta$$

where Q is an arbitrary orthogonal matrix (with appropriate dimension). Unfortunately, and perhaps surprisingly, the OCV score is *not* necessarily the same for both the original and rotated expressions [47] (this lack of rotational invariance is due to the way the diagonal terms of the hat-matrix, A, are affected by the orthogonal matrix, Q). Since this is undesirable, an alternative to OCV should be sought after. [6]

Since we know that the choice of Q affects the terms A_{ii}, one can consider calculating the OCV score given a "good" choice of Q. It is argued in Wood [47] that a "good" rotation would be one that made the terms A_{ii} close to each other, since if they were further apart the calculation of the OCV score would give an excessive degree of weight to some points at the expense of others, and so be determined by a relatively small amount of points from the total sample. It turns out that there exists an orthogonal matrix, Q, such that one can make the terms A_{ii} equal, as the following argument by Wood demonstrates: [47]

The influence matrix of the rotated problem is given by

$$A_Q = QAQ^T$$

where Q^T denotes the matrix transpose of Q. However, since the matrix A is symmetric, we can write $A = BB^T$ so that

$$A_Q = QBB^TQ^T$$

[6] It is also mentioned in Green [12] that the original motivation for an alternative to the OCV method was due to computational reasons, as it was deemed undesirable to be required to calculate the terms A_{ii}. However, since it has been found that one can compute the OCV score in linear time, in the words of Hastie et al. [15], "the original motivation for the generalized cross validation is no longer valid."

Now, if we were to choose Q so to ensure that each row of QB had the same length, then we have that each element down the diagonal of the influence matrix, A_Q, would be of equal magnitude. Moreover, since

$$\begin{aligned} tr(A_Q) &= tr(QAQ^T) \\ &= tr(AQ^TQ) \\ &= tr(A) \end{aligned}$$

each element down the diagonal of the influence matrix, A_Q, must be equal to

$$\frac{tr(A)}{n} \tag{49}$$

However, we do not know, as yet, if such a Q exists. Its existence, however, can be proved using a special kind of orthogonal matrix, known as a *Givens rotation.* A Givens rotation is an orthogonal matrix that, when multiplied from the left with another matrix (M) of an appropriate dimension, results in rotating only two of M's rows while leaving the others unchanged. One can, for any matrix M and any two rows from M, find a Givens rotation that rotates those two rows. [47]

An example of a Givens rotation is the following 3×3 matrix

$$\begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

This particular Givens rotation acts on other $3 \times k$ matrices by simply rotating its first two rows, while keeping the third row unchanged.

Now, speaking more generally the lengths of the two rotated rows of the matrix M vary continuously in θ. Moreover, when θ equals 90 degrees, the *lengths* of the two rows being rotated will have swapped. Hence, there must exist some angle $\hat{\theta}$ such that the lengths of the two rows of under rotation will be equal.

Using the above fact, we can prove the following theorem.

Theorem 3. *For an arbitrary (finite dimensional) matrix B, one can find an orthogonal matrix Q such that the lengths of the rows of the matrix QB are all equal.*

Proof. Firstly, consider the two rows of B that have the smallest and largest lengths, given by $L_{min}(B)$ and $L_{max}(B)$, respectively. Then, apply a Givens rotation, G_1, to B with the value of θ_1 necessary so that post multiplication we have that the lengths of these two rows are equal. Now, since this shared length is between $L_{min}(B)$ and $L_{max}(B)$, we must have that the interval between the smallest and largest lengths of the new matrix, G_1B, is no bigger than the interval between $L_{min}(B)$ and $L_{max}(B)$. If this interval fails to decrease (and the rows are not all the same), then we can repeat this step (again selecting two rows, a Givens rotation G_2 and a theta θ_2 to equalise the lengths of the rows). Note that since B has finitely many rows, after some finitely many repetitions of this step we must have that the distance between smallest and largest lengths is *strictly less* than before. Call the product of these Given matrices H. Note that since the product of orthogonal matrices is also orthogonal, H must itself be orthogonal (since Givens matrices are themselves orthogonal).

Now, let us define the function $g(B)$ to be

$$g(B) := L_{max}(B) - L_{min}(B)$$

and further define the function $f(B)$ to be

$$f(B) = \inf_{U \in O} \frac{g(UB)}{g(B)} \tag{50}$$

where O is the set of orthogonal matrices that are permitted to multiply the matrix B. We can quickly observe that $f(B) < 1$, since by our previous argument we have that $g(HB) < g(B)$ and $H \in O$.

Since O is complete in the Euclidean metric [21] it must be that there exists a $U_B \in O$ such that $f(B) = g(U_BB)/g(B)$ (i.e.$f(B)$ achieves its minimum at $U = U_B$). Now, observe that

$$\begin{aligned}
f(U_BB) \times f(B) &= \inf_{U \in O} \frac{g(UU_BB)}{g(U_BB)} \times \inf_{U \in O} \frac{g(UB)}{g(B)} \\
&= \inf_{U \in O} \frac{g(UU_BB)}{g(U_BB)} \times \frac{g(U_BB)}{g(B)} && \text{(by the property of } U_B\text{)} \\
&= \inf_{U \in O} \frac{g(UU_BB)}{g(B)} \\
&= \inf_{\{V : V = UU_B, U \in O\}} \frac{g(VB)}{g(B)} && (51) \\
&\geq f(B)
\end{aligned}$$

We have an inequality since the choice of orthogonal matrix in the infimum is constrained in (51). However, it is clear that the inequality is actually an equality, since we can use the fact that the identity matrix I is itself orthogonal, so that

$$\begin{aligned}
\inf_{U \in O} \frac{g(UU_BB)}{g(B)} &\leq \frac{g(U_BB)}{g(B)} \\
&= f(B)
\end{aligned}$$

by the property of U_B.

So, we have that

$$f(U_B B) \times f(B) = f(B)$$

which implies that $f(B) = 1$ or 0. However, since we know that $f(B) < 1$, this implies that $f(B) = 0$.

Thus, we have that $L_{min}(U_B B) = L_{max}(U_B B)$, and so we have equalised the lengths of the rows of the matrix. Hence, we can simply set $Q = U_B$, and so we are done. □

Substituting the expression (49) into the OCV score given by equation (48) gives us what is called the *Generalised Cross Validation* (CGV)[47] score:

$$\nu_g = n \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{n - tr(A)} \right)^2 = \frac{n||y - \hat{\mu}||^2}{[n - tr(A)]^2} \tag{52}$$

where $||y - \hat{\mu}||^2$ is simply the sum of squared residuals in norm notation.

As is mentioned in Wood [47], it is actually unnecessary to perform the rotation Q in order to calculate the GCV score; indeed, this is clear since we only require that Q exists, as our resulting score is independent of the actual rotation matrix that would be, theoretically, used to equalize the norms of the rows of the rotated influence matrix.

Furthermore, it is clear that our expression in (52) is rotationally invariant, since neither the prediction error term given by $||y - \hat{\mu}||$, nor the trace of A are affected by multiplication by orthogonal matrices. Since the GCV score is simply the OCV score calculated on the rotated version of the model, we must have that the GCV score is as accurate an estimate of the prediction error as the OCV.

Some further discussion can enlighten us as to the differences between the OCV and GCV methods. As mentioned in Green, the diagonal elements of the influence matrix, A_{ii}, are called *leverage values*, since they affect how much the prediction value $\hat{\mu}_i$ is affected by the datum y_i [12]. If we substitute the expression (47) from our derivation of the OCV score into the GCV score we find that

$$\begin{aligned} \nu_g &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1 - A_{ii}}{1 - tr(A)/n} \right)^2 (y_i - \hat{\mu}_i^{[-i]})^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{(1 - A_{ii})^2}{1 - tr(A)/n} \right)^2 \left(\frac{y_i - \hat{\mu}_i^{[-i]}}{1 - A_{ii}} \right)^2 \end{aligned}$$

This closely resembles the OCV score asymptotically (as $n \to \infty$ we have that $\hat{\mu}_i^{[-i]} \to \hat{\mu}$), except that where there are data with large leverage values, the residuals of the model with such points deleted are then weighted down.

9.4 Smoothing Parameter Estimation in the General Case

So far we have explored the ways in which we can estimate the parameter λ in the special case where the link function, g, is the identity function. However, if we want to expand our scope of analysis to include cases where the link function is non-trivial, we must make adjustments to our previous methods.

Firstly, the fitting objective for the GAM model happens to be equivalent to minimizing the following expression with respect to β [47].

$$D(\beta) + \sum_{j=1}^{m} \lambda_j \beta^T S_j \beta$$

where $D(\beta)$ is the model *deviance* (recall from our discussion on GLMs). which can be thought of as an analogue to the expectation of the residual sum of squares that features in linear statistical models [15], and is defined as

$$D(\beta) = 2[l(\hat{\beta}_s) - l(\beta)]\phi$$

where $l(\beta)$ is the likelihood corresponding to the vector β, whereas $l(\hat{\beta}_s)$ is the maximized likelihood obtained for the saturated model; that is, the model which has a parameter for each data point. Necessarily, $l(\hat{\beta}_s)$ must be greatest amongst *all* possible values of β.

Furthermore, it is known that the deviance, given λ, can be approximated by

$$||\sqrt{W}(z - X\beta)||^2 + \sum_{j=1}^{m} \lambda_j \beta^T S_j \beta$$

where the terms $\sqrt{W}$ and z are the limits of the sequences of weights matrices and pseudo-data vectors, respectively, from the IRLS method previously mentioned in our discussion on GLMs [47]. [7]

Consequently, we can now modify the previously discussed methods of smoothing parameter estimation, by replacing the sum of squared residuals term with the approximation of the deviance, $||\sqrt{W}(z-X\beta)||^2$, in the previous arguments used to derive the UBRE, OCV and GCV scores. Hence, in our new scores we simply substitute $||\sqrt{W}(z - X\hat{\beta})||^2$ for the sum of squared residuals.

So, for example, the generalised cross validation score can now be expressed as

$$\nu_g^w = \frac{n||\sqrt{W}(z - X\hat{\beta})||^2}{[n - tr(A)]^2}$$

[7] This approximation follows by demonstrating that the first two derivatives of the deviance and $||\sqrt{W}(z - X\beta)||^2$ are equal [47]

However, since the approximation of the deviance was conducted *given* the value of λ, this GCV score only holds locally for the λ used to calculate the terms from the IRLS algorithm. Although, as is mentioned in Wood [47], a GCV score that is globally applicable is given by

$$\nu_g^w = \frac{nD(\hat{\beta})}{[n - tr(A)]^2}$$

where we have, again, the fact that $||\sqrt{W}(z - X\hat{\beta})||^2$ and $D(\hat{\beta})$ are approximately equal.

Even though the use of the model deviance in the GCV score seems appropriate given its status as an analogue to the residual sum of squared errors, it so happens that the approximation is, in practice, quite poor [47]. The solution proffered in Wood is to consider the approximation as being justified up to an additive constant, so that

$$D(\hat{\beta}) + k = ||\sqrt{W}(z - X\hat{\beta})||^2$$

where k is some constant (this constant, k, would be required to be estimated from an initial model fit). It is mentioned, however, that even with this adjustment there is not a marked improvement in the calculation of the GCV score [47].

On the other hand, if we consider the UBRE score modified by the circumstances of our more general model (which we denote by ν_g^p), we have for a given λ that

$$\nu_u^w = \frac{1}{n}||\sqrt{W}(z - X\hat{\beta})||^2 - \sigma^2 + \frac{2}{n}tr(A)\sigma^2$$

Furthermore, we can extend this score to be globally applicable, by again re-using the approximate equality of $||\sqrt{W}(z - X\hat{\beta})||^2$ and the deviance, giving us

$$\nu_u = \frac{1}{n}D(\beta) - \sigma^2 + \frac{2}{n}tr(A)\sigma^2$$

Conveniently, it does not matter if our approximation is wrong by an additive constant when we come to minimize this UBRE score (since any vertical translation does not change the position of the minimum).

A further possibility for modifying the GCV and UBRE scores is to replace the deviance term with the Pearson statistic, which is given by the expression

$$X^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})} \qquad \text{(recall that } V(\mu) = Var(Y)/\phi\text{)}$$

in which case we have that the 'Pearson' scores are given by

$$\nu_g^p = \frac{n \sum_{i=1}^{n} V(\hat{\mu_i})^{-1}(y_i - \hat{\mu_i})^2}{[n - tr(A)]^2}$$

and

$$\nu_u^p = \frac{1}{n} \sum_{i=1}^{n} V(\hat{\mu_i})^{-1}(y_i - \hat{\mu_i})^2 - \sigma^2 + \frac{2}{n} tr(A)\sigma^2$$

for the GCV and UBRE methods, respectively. However, even these alternatives have shortcomings. As a consequence of the dependence of the variance function V on $\hat{\mu_i}$, the Pearson statistics fails to closely approximate the deviance. Moreover, there is a tendency for these scores to over smooth. [47]

10 Degrees of Freedom

In the linear statistical model (LSM), the degrees of freedom tells us the number of parameters that are allowed to vary in the model. We have that the degrees of freedom is given by the trace of A, the influence matrix [47]. Analogous to the LSM, in the GAM we have that the *effective degrees of freedom* is defined to be $tr(A)$. Unlike the LSM counterpart, the degrees of freedom for the GAM model need not be integer valued, which makes our definition less straightforward to interpret. However, we can get a sense of what the effective degrees of freedom is by considering what occurs when we vary the smoothing parameters λ_i between 0 and ∞. In the case where for all i we have $\lambda_i = 0$ the model has the greatest flexibility, and so the effective degrees of freedom will be at a maximum (which is the number of coefficients in the vector, β, minus the number of constraints). Conversely, as the λ_i approaches ∞ we must have that the model's flexibility is reduced, and so the effective degrees of freedom will also be reduced.

Since we recognise that the effective degrees of freedom for a model is affected by the values of the roughness penalties, given that in the model we may have several smooth functions to estimate one may wish to focus simply on the the degrees of freedom corresponding to a particular smooth function (as each function will most likely have different values for their respective roughness penalties). We can even extend this approach and consider the degrees of freedom corresponding to each $\hat{\beta}_j$ in the model.

Noting that the influence matrix, A, is given by

$$A = X\left(X^TX + S\right)^{-1}X^T$$

(where $S = \sum_{j=1} \lambda_j S_j$), we can define

$$P = \left(X^TX + S\right)^{-1}X^T$$

so that tr(A) = tr(XP)

Now, we can further define P_i^0 to be the matrix P with the all the elements changed to zero, except with the ith row unchanged. In which case, if we consider the product $P_i^0 y$, this matrix has entry $\hat{\beta}_i$ corresponding to the element$[P_i^0 y]_{(i,i)}$, with zero everywhere else. Moreover, we have that

$$\begin{aligned}\sum_{i=1}^{p} tr(XP_i^0) &= tr(\sum_{i=1}^{p} XP_i^0)\\ &= tr(XP)\\ &= tr(A)\end{aligned}$$

where p is the number of rows of the matrix P. Hence, we can consider $tr(XP_i^0)$ as being the degrees of freedom corresponding with the ith parameter.

We can also prove a further result. Computing the matrix product XP_i^0 yields the matrix

$$\begin{matrix} x_{1i}[(X^TX+S)^{-1}X^T]_{(i,1)} & \dots & x_{1i}[(X^TX+S)^{-1}X^T]_{(i,n)} \\ \vdots & x_{ii}[(X^TX+S)^{-1}X^T]_{(i,i)} & \vdots \\ x_{ni}[(X^TX+S)^{-1}X^T]_{(i,1)} & \dots & x_{ni}[(X^TX+S)^{-1}X^T]_{(i,n)} \end{matrix}$$

where n is the number of rows of the matrix X. From this computation we have that

$$\begin{aligned} tr(XP_i^0) &= \sum_{k=1}^{n} x_{ki}[(X^TX+S)^{-1}X^T]_{(i,k)} \\ &= [PX]_{(i,i)} \end{aligned}$$

Consequently, we have that the effective degrees of freedom corresponding to the parameter $\hat{\beta}_i$ is given by entry (i,i) of the matrix

$$F := PX = \Big(X^TX+S\Big)^{-1}X^TX \tag{53}$$

One may use this matrix as a way of contrasting the estimated parameters of the un-penalised model against those of the penalised model. The estimated parameters of the un-penalised model, denoted by $\tilde{\beta}$, are given by

$$\tilde{\beta} = \Big(X^TX\Big)^{-1}X^Ty \tag{54}$$

The parameters for the penalised model, however, are given by

$$\begin{aligned} \hat{\beta} &= \Big(X^TX+S\Big)^{-1}X^Ty \\ &= \Big(X^TX+S\Big)^{-1}X^T\Big(X^TX\Big)^{-1}XX^Ty \\ &= F\tilde{\beta} \end{aligned}$$

with the last equality holding by equations 53 and 54.

The matrix F is now seen to be a map from the unpenalised parameter estimates to the penalised parameter estimates, with the diagonal entries themselves indicating the degree to which the penalised parameter estimates $\hat{\beta}_i$ change in response to a unit change in the un-penalised parameter estimates $\tilde{\beta}_i$ since

$$\frac{\partial\hat{\beta}_i}{\partial\tilde{\beta}_i} = F_{(i,i)}$$

This provides some intuition behind how these entries correspond with our notion of the effective degrees of freedom associated with parameter $\hat{\beta}_i$. Each un-penalised parameter is associated with one degree of freedom, but with the imposition of a roughness penalty the degrees of freedom is scaled down by a factor $F_{(i,i)}$. [8]

[8] Wood warns us, though, to consider that this interpretation may be found wanting since there is no constraint forcing $F_{(i,i)} > 0$, however it does seem to satisfy this inequality for typical choices of bases and penalties [47] .

11 A Simple Application with VGAMs

We now wish to utilise the theory we have previously discussed in the context of the problem of producing abnormal returns in a horse race betting market. Our application will be purely focussed on the first task of finding estimates for the probabilities of victory for given horses in a particular race. The data we use is somewhat contrived: it consists of one hundred races where in each race there is exactly ten horses (not necessarily the same horses appear in every race, however) [9]. For each horse j in race i there are certain characteristics associated with them. These include:

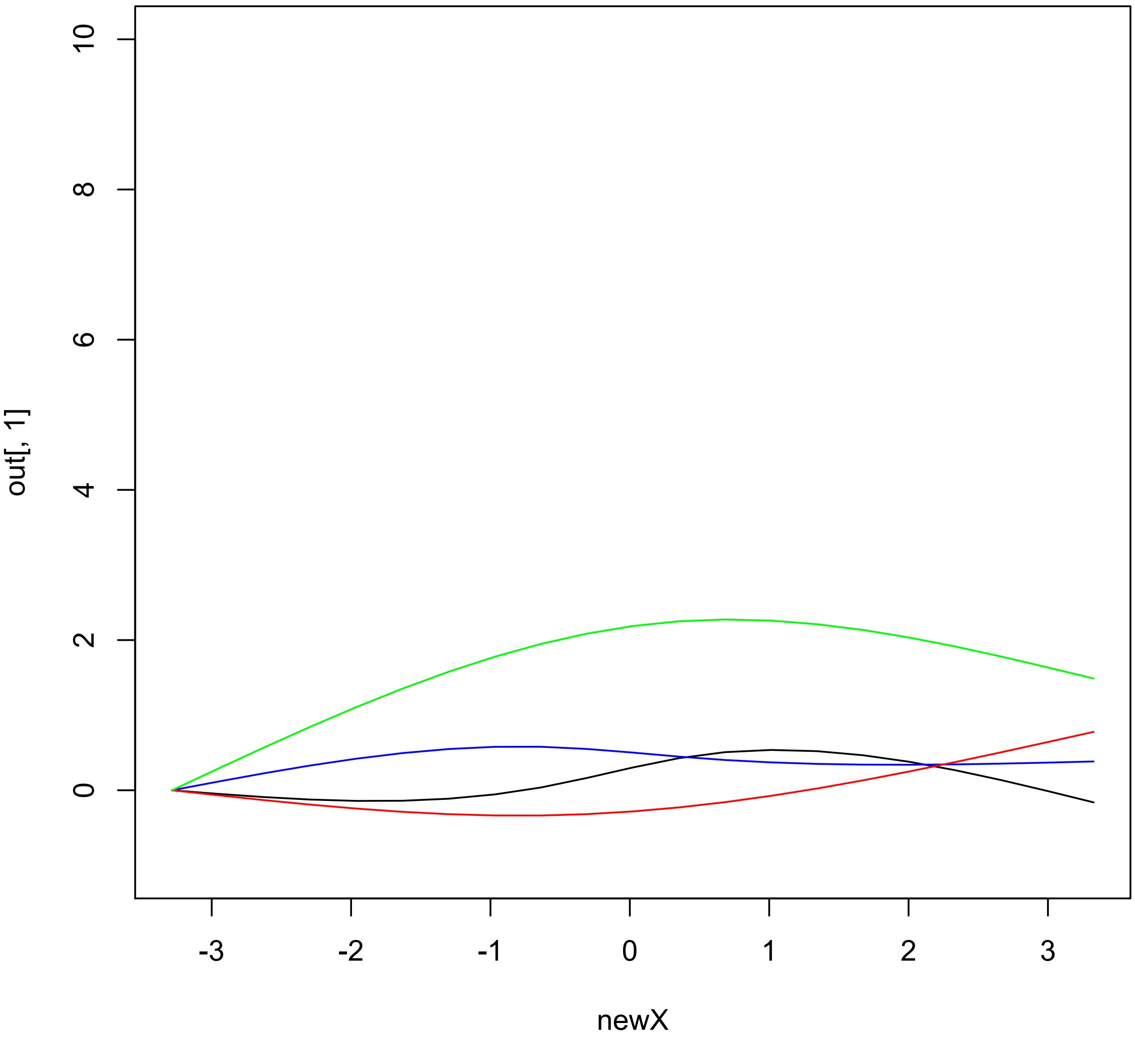
- Their closing odds (i.e. the odds set at the start of the race) (a)
- The days since their last first placing in a race (b)
- The logarithm of their track probability, adjusted for the bookmaker's margin (c)
- Some unknown variable (this particular variable's origin was not made clear to the reader! (d)

Moreover, we have knowledge of their actual placing in each race (note that there are no ties, so there is always a winner). We use what is called a *vector generalised additive model* (VGAM) [**?**] to combine our discrete choice model and the GAM, so that we can estimate the probabilities of a certain horse j winning race i. In the VGAM with $m + 1$ many horses and k many predictors $x_1, \dots, x_k$ of a horses 'winning-ness' we have that for a race with horses $j = 1, ..., m$ that the probability of horse j winning (denote this by $Y = j$) is modeled by.

$$\log \frac{P(Y = j)}{Pr(Y = m + 1)} = \beta_1 (x_{j1} - x_{m+11}) + \dots \beta_k (x_{jk} - x_{m+1k})$$

An extension to this model allows the coefficients β_t to become actual smooth functions, however. Applying this model to the data provided we find that the determinant (a) is significant (as one would hope), and we can yield a representation of it as a smooth function

[9] The source of the data is UK racetrack odds, provided by Johnnie Johnson


10
8
6
4
2
0
out[, 1]
-3
-2
-1
0
1
2
3
newX

12 The Kelly Betting System

12.1 Betting Systems

Once we have obtained estimates of the probabilities of horse j winning race i we require a betting system to take advantage of this information, if we wish to demonstrate that one can generate profits from observed inefficiencies in the market. By a betting system, one means a prescription for how much a gambler ought to bet on each successive round (race) given knowledge of the probabilities and odds corresponding to the different possible outcomes (winning horses). One particular betting system that is commonly employed is the Kelly betting system, named after J.L Kelly Jr [19]. This system is an appropriate gambling strategy for games where the expected value of playing is positive, whereby the gambler's goal is to maximise the expected value of the logarithm of their wealth (supposing they have knowledge of the respective payoffs accruing from each outcome, as well as their respective probabilities). But, before describing the Kelly betting system's origins and desirable properties, it is instructive to briefly consider some alternative betting systems that could be pursued.

Naively, a gambler may wish to set as their objective the maximization of their expected wealth. In the context of a sequence of gambles with positive expected value, however, this implies that the gambler must bet the entirety of their wealth on each successive gamble. Consequently, as the number of successive bets approaches infinity, the probability of the gambler reaching ruin (zero wealth) approaches one, almost surely (supposing that the successive outcomes are independent and probability of winning is not equal to one, of course!) [3]. As this is undesirable for the gambler who wishes to play arbitrarily many rounds, it is clear that some strategy that commits the gambler to secure some fraction of their wealth is preferable. For instance, if the gambler were to commit a sufficiently small fixed fraction of their income towards every round of betting, by the law of large numbers their wealth would increase towards infinity as the number of bets approaches infinity [3]. However, it may be that their wealth approaches infinity very slowly using the fixed fraction approach, and so a betting system with better understood (and hopefully, optimal) asymptotic properties is perhaps more desirable to the gambler. It is this last point that, as we will see, provides the central motivation for the Kelly betting system.

On the other hand, one may suppose that the gambler possesses a utility function, $U(W)$, which takes as its input the gambler's wealth, W. Given this function, the objective of the gambler is to maximize their expected utility of wagering. This approach leads to a particular betting system for the gambler; often, the gambler is assumed to possess (to some degree) a concave utility function, which implies a level of risk aversion, and so their betting system will necessarily possess the desirable property of avoiding wagering their entire wealth on risky bets [11]. But, given that our brief is to simply demonstrate the existence of a profitable betting strategy, it is unnecessary that we should trouble ourselves with using the utility theory method, and instead opt for a betting system that has a clear monetary objective. [10]

[10] There is, however, an important connection between the expected utility approach and the Kelly betting system, which is discussed in 8.6.

12.2 The Kelly Criterion and its Motivation

Given the previous discussion, a reasonable goal that we may wish for our betting system is that it optimizes the rate of growth of wealth. To illustrate the advantages of the Kelly Betting System in this regard, we shall consider a (much simpler) special case [11] where the gambler is presented with an infinite sequence of independent, identically distributed gambles (say, coin tosses) that each return a payoff (if the gambler wins) that is equal to the amount of money staked, and are biased in the gambler's favour.

To put it more formally, suppose that the gambler has an original wealth of W_0 and considers betting on the outcome of a sequence of independent, identically distributed random variables, $X_1, ..., X_n, ...$ such that $\forall n,\ Pr(X_n = 1) = p > 1/2$, and $Pr(X_n = -1) = 1 - p$. Suppose that they choose to stake b_n dollars on the nth bet, and that if they win the bet ($X_n = 1$), they receive b_n, whereas if they lose the bet ($X_n = -1$), they lose b_n. Hence, if we denote the gambler's wealth after the nth gamble by W_n, we have, by recursion, that:

$$W_n = W_{n-1} + b_n X_n = W_0 + \sum_{i=1}^{n} b_i X_i$$

Now, given that the sequence of gambles, $X_1, ...X_n$, are identical, one might suppose that any betting system that optimises the rate of growth of wealth involves betting a certain fixed fraction of one's wealth on each successive bet [12]. Consequently, under this assumption, we may express the gambler's wealth after n trials as:

$$W_n = (1+l)^{S_n}(1-l)^{n-S_n} W_0$$

where S_n is the random variable indicating the number of successes after the nth gamble, and $l \in (0,1)$ is the fixed fraction of wealth bet on each successive gamble. [13]

Now, consider the relation:

$$\frac{W_n}{W_0} = \exp\left(n \log\left(\frac{W_n}{W_0}\right)^{\frac{1}{n}}\right)$$

The term inside the exponential is n times what is called the exponential growth rate of wealth after the nth bet [45]. We shall denote this rate by the function $G_n(l)$, so that:

[11] Indeed, this is the example explored in Kelly's original paper [19], but the arguments herein follow closely those in Thorp [44] [45].

[12] We, as yet, do not have rigorous grounds for this supposition. However, we will soon demonstrate that this is indeed justified.

[13] The need for this is apparent, due to the fact that if $l = 1$ we have that $\lim_n W_n = 0$, almost surely, whereas if $l = 0$ we have that $W_n = W_0$, for all n, and so the gambler effectively forfeits any possible gains presented by the opportunity to gamble.

$$\begin{aligned} G_n(l) &= \log\left(\tfrac{W_n}{W_o}\right)^{\frac{1}{n}} \\ &= \frac{S_n}{n}\log(1+l) + \frac{(n-S_n)}{n}\log(1-l) \end{aligned}$$

But, by the strong law of large numbers,[14] as n approaches infinity we have that $G_n(l)$ approaches:

$$\begin{aligned} p\log(1+l) + (1-p)\log(1-l) &= E[G_n(l)] \\ &= E[\log(W_n^{\frac{1}{n}})] - E[\log(W_0^{\frac{1}{n}})] \end{aligned}$$

And, since W_0 is constant, for any given n the objective of maximising the above expression is the same as maximising

$$E[\log(W_n)] \tag{55}$$

Hence, if we wish to maximise the asymptotic exponential growth rate of wealth, we ought choose l so to maximise expression 55; this is what is known as the Kelly Criterion. The existence of a maximum can be demonstrated by elementary calculus in this special case, by maximising $g(l) := E(G_n(l))$ (one should notice that n does not appear in the expectation of $G_n(l)$). By letting $g'(l) = 0$ we have that:

$$0 = \frac{p}{1+l} - \frac{1-p}{1-l} \; = \frac{p-(1-p)-l}{(1+l)(1-l)}$$

which has solution $l^* = 2p-1$, which we can furthermore prove to be a global maximum, since:

$$g''(l) = -\frac{p}{(1+l)^2} - \frac{1-p}{(1-l)^2} < 0, \;\; l \in (0,1)$$

Then, using the fact that $g'(0) = 2p-1 > 0$ and $\lim_{l\to 1^-} = -\infty$, and appealing to the intermediate value theorem we find that the value l^* which maximises $g(l)$ (and hence $E[\log(W_n)]$) is unique. As a side note, the value that g(l) attains at its maximum is denoted by R, and is called the optimal growth rate[15] [19]. Given our previous work it can easily be computed as:

$$R = p\log(2p) + (1-p)\log(2(1-l)) = p\log(p) + (1-p)\log(1-p) + \log(2)$$

[14] That is, $\lim_{n\to\infty} \frac{S_n}{n} = p$, almost surely

[15] This term was originally called the 'information rate' by Claude Shannon, whose work motivated Kelly's paper [19]

We have shown that l^* maximises $E[\log(W_n)]$ within the class of fixed fraction betting systems, but it is also the case that l^* achieves this within the class of all betting systems. To demonstrate this we will state and proof a stronger result, found in Thorp [44].

If we consider a slightly more general gambling game to the one so far described, where the gambler is confronted with a sequence of independent random variables which return Q_n per 1 dollar bet on the nth trial, and let b_i denote the amount of the gambler's wealth wagered on the ith trial, then since $W_n = \prod_{i=1}^n W_i/W_{i-1}$ and $W_i = W_{i-1} + b_iQ_i$ we have that:

$$\begin{aligned}
E[\log(W_n)] &= E\left[\log\left(\prod_{i=1}^{n}\frac{W_i}{W_{i-1}}\right)\right] \\
&= E\left[\sum_{i=1}^{n}\log\left(\frac{W_i}{W_{i-1}}\right)\right] \\
&= \sum_{i=1}^{n} E\left[\log\left(1+\frac{b_iQ_i}{W_{i-1}}\right)\right]
\end{aligned}$$

Letting $L_i = b_i/W_{i-1}$, since b_i and W_{i-1} are determined by the first $i-1$ trials and Q_i is the outcome of the ith trial, which is independent of all previous trials, we must have that L_i is independent of Q_i. Consequently, we can choose L_i to maximize the expression $E[\log(1 + L_iQ_i)] = E[E[\log(1 + L_iQ_i)|L_i]]$ (of course, noting that $L_i \in [0, 1]$).

Theorem 4. *In the aforementioned game, if one can always find, for each i, an $l_i \in (0, 1)$ such that $E[\log(1 + l_iQ_i)] > 0$, then for each i there exists an l^* such that the expression $E[\log(1 + L_iQ_i)]$ has a unique maximum at $L_i = l_i^*$, almosts surely.*

Note that, if we consider a special case of this claim where Q_i has the same distribution as X_i from our original gambling game (i.e. where $\forall i, Pr(Q_i = 1) = p > 1/2$ and $Pr(Q_i = -1) = 1 - p$) then using our previous arguments about maximizing $g(l)$, we can set for each i, $l_i^* = 2p - 1$, and so we yield the Kelly Betting System.

Proof. For the purpose of ignoring trivialities we limit ourselves to only consider the case for when for each i, Q_i is nonzero, almost surely. Then, we proceed by acknowledging that since the expression $E[\log(1 + l_iQ_i)]$ is defined, we must have that the domain of this function (of l_i) lies within an interval of the form $[0, a_i)$ or $[0, a_i]$, where we let $a_i = \min(1, \sup\{l_i : l_iQ_i > -1\})$. We can compute the second derivate of this function (with respect to l_i), which is given by $-E[Q_i^2/(1 + l_iQ_i)^2]$. Since it is negative on either of the possible domains, we must have that any maximum (if it exists) is unique. If the function is defined at a_i, then $E[\log(1 + l_iQ_i)]$ lies within a compact interval, and so by the continuity of this function we can apply the Bolzano-Weierstrass maximum theorem [35] to imply that our function obtains a maximum over this interval. On the other hand, if the function is not defined at a_i, we nonetheless must have a maximum since $\lim_{l_i \to a_i^-} E[\log(1 + l_iQ_i)] = -\infty$ (since the second derivative is negative).

Moreover, by the independence of L_i and Q_i we can consider these random variables as being functions $L_i(s_1)$ and $Q_i(s_2)$ on a product measure space $S_1 \times S_2$. Consequently, we have that

$$\begin{aligned}E\log(1+L_iQ_i) &= \int_{S_1}\int_{S_2}\log(1+L_i(s_1)Q_i(s_2)) \\ &= \int_{S_1} E\log(1+L_i(s_1)Q_i(s_2)) \\ &\leq E\log(1+l^*Q_i) \qquad \text{(by the concavity of the logarithm function)}\end{aligned}$$

For equality, we require that $E\log(1+L_iQ_i) = E\log(1+l_i^*Q_i)$, almost everywhere, which is equivalent to the condition that $L_iQ_i = l_i^*Q_i$ almost surely. But, due to independence, this implies that either $L_i = l_i^*$ or $Q_i = 0$, a.s. However, since $Q_i \neq 0$, a.s, we must have that $L_i = l_i^*$, a.s. □

Thus, having proved the theorem, we have that the Kelly betting system maximises $E[\log(W_n)]$ within the class of all possible betting systems, fixed fractional or otherwise.

Further to this, we can prove three close results that describe the relationship between the function $g(l)$ and the asymptotic behavior of W_n (these proofs are found in [44]). In this theorem we will revert back to our original example introduced at the beginning of this section.

Theorem 5. *(i): if $g(l) > 0$, then $\lim_{n\to\infty} W_n = \infty$, almost surely*
(ii): if $g(l) < 0$ then $\lim_{n\to\infty} W_n = 0$, almost surely
(iii): if $g(l) = 0$ then $\limsup_{n\to\infty} W_n = \infty$, while $\liminf_{n\to\infty} W_n = 0$, almost surely

Proof. (i): consider from our previous work that:

$$g(l) := E[G_n(l)] = \lim_{n\to\infty} \log\left(\tfrac{W_n}{W_o}\right)^{\frac{1}{n}} \qquad \text{almost surely}$$

So, supposing that $g(l) > 0$, we can find an $N(\omega) \in \mathbb{N}$ for a given $\omega \in A \subset \Omega$ (where Ω is the space of all infinitely long sequences of Bernoulli trials and A has probability measure 1) so that $\forall n \geq N(\omega)$ we have:

$$\log\left(\tfrac{W_n}{W_o}\right)^{\frac{1}{n}} \geq \frac{g(l)}{2} > 0 \tag{56}$$

which implies that

$$W_n \geq W_o \exp\left(\frac{ng(l)}{2}\right)$$

which gives us the result that if $g(l) > 0$, then $\lim_{n\to\infty} W_n = \infty$, almost surely.

(ii): The second result can be obtained in almost the same way: if we suppose that $g(l) < 0$ we can similarly choose an $N(\omega) \in \mathbb{N}$ and reverse the inequalities in 56 to imply that $\forall n \geq N(\omega)$ we have:

$$W_n \leq W_o \exp\left(\frac{ng(l)}{2}\right)$$

and so we yield the result that if $g(l) < 0$ then $\lim_{n\to\infty} W_n = 0$.

Before we consider part (iii) of our theorem, we will firstly state and prove a useful lemma.

Lemma 6. If we let S_n be the sum of n i.i.d. Bernouilli random variables with probability of success p, and let $M > 0$, then $S_n \geq np + M$ occurs infinitely often with probability 1.

Proof. Firstly, we will require the Law of Iterated Logarithm [4], which in the case of Bernouilli trials states that

$$\limsup_{n\to\infty} \frac{S_n - np}{\sigma\sqrt{2n\log(\log(n))}} = 1 \tag{57}$$

almost surely

Suppose that the lemma is false. Then, there exists an M such that $S_n < np + M$, eventually, with non-zero probability. However, by (57) this implies that

$$\limsup_{n\to\infty} \frac{S_n - np}{\sigma\sqrt{2n\log(\log(n))}} \leq 0 \qquad \text{eventually}$$

with non-zero probability, which is a contradiction. □

(iii): In this case, we have to restrict ourselves to the consideration of $\liminf W_n$ and $\limsup W_n$. For the latter case, note that for an infinite sequence of Bernoulli random variables, the number of successes after the nth trial, S_n, can be arbitrarily far away from the mean np as n tends towards infinity. Now, given an $M \in \mathbb{N}$, if we suppose that $S_n \geq np + M$ then we have that:

$$\begin{aligned}
\log\left(\frac{W_n}{W_o}\right)^{\frac{1}{n}} &\geq \frac{np+M}{n}\log(1+l) + \frac{n-(np+M)}{n}\log(1-l) \\
&= g(l) + \frac{M}{n}\log\left(\frac{1+l}{1-l}\right) \\
&= \frac{M}{n}\log\left(\frac{1+l}{1-l}\right)
\end{aligned}$$

So, by the monotonicity of the logarithm function, we have that:

$$W_n \geq W_o\left(\frac{1+l}{1-l}\right)^M \tag{58}$$

Now using our lemma that $S_n \geq np + M$ occurs infinitely often with probability 1, we can choose M to be arbitrarily large so that our result from (58) yields:

$$\limsup_{n\to\infty} W_n = \infty$$

In a similar way we can prove that $\liminf_{n\to\infty} W_n = 0$. □

These results are important regarding the concern of gambler's ruin: the situation where the gambler's wealth is reduced to zero [16]. Since it is easily checked that $g(l^*) > 0$ (recall that we are supposing that $p > 1/2$), it follows that Kelly betting ensures that the gambler faces ruin with probability zero. Furthermore, since $g(l) > 0$ for $l \leq l^*$, the gambler (supposing they are employing a fractional strategy) will avoid ruin if they choose to bet below the Kelly optimal fraction, l^*. The gambler may also avoid ruin if they bet slightly above the Kelly optimal fraction, so long as $g(l) > 0$; however, this results in a greater risk (in the form of larger bets) at the expense of a lower growth rate of wealth, and as such is unattractive to most gamblers (in finance parlance, these riskier strategies are said to be dominated in geometric risk-return by Kelly betting [26]).

As might be reasonably guessed, there are indeed other desirable results that follow from pursuing the Kelly Betting System. Continuing with this example, it can be proved that if there exists another fractional betting system with fraction l where, if we let $W_n(l^*)$, $W_n(l)$ denote the wealth attained by the nth bet by pursuing the Kelly Betting System and its alternative, respectively, then $\lim_{n\to\infty} W_n(l^*)/W_n(l) = \infty$, almost surely.

Again, using an argument from [44], since we know that $g(l^*) > g(l)$ if $l \neq l^*$ (by the uniqueness of l^* which we have previously established) we have that

$$\begin{aligned}\log\left(\tfrac{W_n(l^*)}{W_n(l)}\right)^{\frac{1}{n}} &= \log\left(\tfrac{W_n(l^*)}{W_o}\right)^{\frac{1}{n}} - \log\left(\tfrac{W_n(l)}{W_o}\right)^{\frac{1}{n}} \\ &= \frac{S_n}{n}\log\left(\frac{1+l^*}{1+l}\right) + \frac{n-S_n}{n}\log\left(\frac{1-l^*}{1-l}\right) \qquad (59)\end{aligned}$$

However, expression 59 approaches $g(l^*) - g(l) > 0$ by the strong law of large numbers. Then using theorem 5 (i), we have that $\lim_{n\to\infty} W_n(l^*)/W_n(l) = \infty$. It should be mentioned that this result generalizes to consider alternative betting strategies that are not necessarily fixed fractional; indeed, if a betting system Φ is "essentially different" [17] to the Kelly Betting System, then our result will still hold.

Another desirable property of the Kelly Betting System is that it minimises the expected time (that is, the number of bets) for the gambler's wealth W_n to reach some predetermined value, C, amongst all other possible strategies, as C approaches infinity.

[16] Strictly speaking, gambler's ruin occurs when at some n we have that $W_n = 0$. Clearly, in this setup gambler's ruin is impossible. Thus, for our purposes we will adjust this definition to be that gambler's ruin occurs if W_n converges in probability to 0, as is done in Thorp [44]

[17] A strategy, Φ, is essentially different to a strategy Φ^* if $E[\log(W_n(\Phi^*)) - \log(W_n(\Phi))]$ grows faster as n approaches infinity than the standard deviation of $\log(W_n(\Phi^*)) - \log(W_n(\Phi))$[45][3].

The proof of this result for even the relatively simple special case that we are considering is lengthy [44], and so is omitted [18]. This result, however, provides yet another important justification for employing the Kelly Betting System.

Of note is how this betting system is remarkably simple, since at each step the gambler need only know their present wealth and probability of the forthcoming trial in order to determine how much they are willing to bet; such a strategy is called a *Markov betting system*, [9] [19] since the sequence of random variables $(W_i)_{i\in\mathbb{N}}$ forms a Markov process.

12.3 A More General Setting

Now, while we have provided a thorough exposition of some of the important properties of the Kelly Betting System through the special case of independent, identically distributed trials where the gambler only need decide on whether to bet, nor not to bet, it is important to note that these results illustrated hold in a more general setting [3]. Breiman extended Kelly's work by considering a sequence of identically distributed gambles where there are multiple (but fintely many) choices to bet on at each stage, where the odds (i.e. respective payoffs) on each choice are not necessarily equal, and where the respective outcomes are not necessarily mutually exclusive [3].

In this more general setup, Breiman defined a "favorable game" to be one such that there exists a fixed fraction strategy Δ so that, if one bets according to this strategy, then $\lim_{n\to\infty} W_n(\Delta) = \infty$, almost surely [3]. Referring back to our original example in subsection 8.2, it is clear that this special case is indeed a favorable game (since the Kelly system satisfies the necessary criteria). From here, Breiman considered two particular criteria for which to judge the desirability of particular betting strategy: The minimal time requirement, which through fixing a value C judges which betting system minimizes the expected time (trials) taken to reach or exceed this value, and; the magnitude condition, which fixes the number of trials n and considers the betting system which obtains, in expectation, the greatest wealth by the nth trial. Breiman then demonstrated that the strategy Δ^* which achieved both of these objectives (asymptotically) was indeed the Kelly Betting System.

The vector that represents the optimal allocation of wealth to the different bets at time n is, in this general case, not unique, however the value of $E\log(W_n)$ is still the same for the different optimal vectors. Moreover, uniqueness can be guaranteed by simply modifying the problem so that each outcome is mutually exclusive. It is important to note that in this modification, the optimal fractions corresponding to each outcome A_j are indeed equal to $Pr(A_j)$, irrespective of the odds (i.e, payoffs) associated with these outcomes. This ensures that Kelly betting system diversifies the wager, rather than making the bettor invest all their resources into a single outcome that yields the greatest expected value (this concurs with our example in subsection 8.2, even though the wager is not between mutually exclusive events in that case).

[18] The proof involves showing that the expected number of trials to reach the goal of C can be approximated by $\log C/g(l)$. This is minimized when $g(l)$ is maximized, which occurs when $l = l^*$

[19] Actually, it is even simpler, since the proportion the gambler wagers is constant throughout all bets; this strategy is called a *proportional stationary Markov betting system.* [9]

A simple illustration of why the odds for respective outcomes do not influence the optimal proportion in Kelly betting (under mutually exclusive events) is to consider a slight adjustment of the example in 8.2, where instead of the gambler deciding whether to bet or not bet, their decision is to determine how much to wager on the event of a head ($\{H\}$) and how much on the event of a tail ($\{T\}$), where a one dollar bet yields a and b dollars, respectively. We furthermore constrain our choices of a and b so that $pa > 1$ and $(1-p)b > 1$, ensuring that each wager is biased in the bettors favour. Then, if we let the random variable Y_t denote the event that the coin is a Head ($Y_t = 1$, with probability p) or Tail ($Y_t = 0$, with probability $1-p$), we have that:

$$
\begin{aligned}
E[\log(W_t)|W_{t-1}] &= E[\log(aY_tW_{t-1}l + b(1-Y_t)W_{t-1}(1-l))] \\
&= p\log(aW_{t-1}l) + (1-p)\log(bW_{t-1}(1-l))
\end{aligned}
$$

After taking the derivative of this expression with respect to l and letting this derivative equal zero, we have that

$$
0 = \frac{p}{l} - \frac{1-p}{1-l}
$$

whence, we find a solution $l^* = p$, which is clearly independent of our choice of odds a and b. This differs to the result reached in our original example in 8.2, since in that case the gambler's choices were not mutually exclusive (they were effectively between the events {H} and {H,T}).

While we have so far discussed the generalisations of certain asymptotic properties, there are also desirable non-asymptotic properties of the Kelly betting system. One in particular worth mentioning is that which is discussed in Bell and Cover [1]. In their paper, Bell and Cover consider the case of two gamblers who compete with each other to accumulate the largest amount of wealth after a single trial by simultaneously wagering on the same game after given the initial choice of substituting their starting wealth (normalized to 1) for some other random variable with mean 1. The solution for both players, since it is symmetric, is for each of them to substitute their initial wealth for a uniformly distributed random variable, and then proceed with making a Kelly bet. [1]

To further explain this result we will consider a special case, where the game is a Bernouili random variable with the payoff of winning equal to the amount wagered. We may suppose that two players, A and B, are competing to accumulate the most wealth after the outcome of this Bernouilli random variable (note that they are not betting against each other here, but rather "the house", so to speak). They both have initial wealth of 1, but they then switch their initial wealth for another random variable $V_A, V_B \geq 0$, respectively, with the constrain that $E[V_A] = E[V_B] = 1$. We shall further assume that the random variables V_A, V_B and X are independent (that is, they do not observe each other's wealth after exercising the opportunity to switch, and switching does not affect the outcome of the trial that they intend to wager on). In this setup, we allow for the players to effectively "not" switch their wealth; that is, one could have $V_A = 1$, almost everywhere, for example. Denoting the Bernouilli random variable by X, if we further suppose that this random variable is biased in the bettors' favour, where we assume that they are betting on the event that $X = 1$, we have that

$$Pr(X = 1) = p > 1/2 \qquad Pr(X = 0) = 1 - p$$

Then we have that the wealth of gamblers A and B after the outcome of the trial, denoted by W_A and W_B, respectively, are given by

$$W_A = V_A[(1 - a(V_A)) + a(V_A)2X] \tag{60}$$

and

$$W_B = V_B[(1 - b(V_B)) + b(V_B)2X] \tag{61}$$

where $a(V_A)$ and $b(V_B)$ are the proportions of A's and B's wealth wagered (after the switch), respectively.

Now, we can state a theorem given in Ferguson [10], which is simply a special case of that which is proven in Bell and Cover [1].

Theorem 7. *In the aforementioned scenario, an optimal strategy for player A to pursue is to switch their original wealth to a random variable V_A which is uniformly distributed on the interval $(0, 2)$, and then choose to bet the proportion $a(V_A) = 2p-1$ of their wealth (i.e. the Kelly bet) on the outcome of $X = 1$.*

Note that by symmetry of the scenario we must have that this theorem also holds for player B too. Furthermore, the theorem does not exclude the possibility of other strategies achieving an optimal outcome for the player ("an" not "the" optimal strategy). To prove this result we must firstly prove the following lemma.

Lemma 8. If we suppose that U is a random variable with uniform distribution over the interval $(0, 2)$ and T is a random variable, independent to U, with any distribution so long as $T \geq 0$, a.s, then

$$Pr(U < T) \leq \frac{1}{2}E(T) \tag{62}$$

Proof. Note that U has the following distribution, where $\mathbf{1}$ is the indicator function (also known as the characteristic function):

$$F_U(u) = \frac{1}{2}u\mathbf{1}_{\{u\in(0,2)\}} + \mathbf{1}_{\{u\geq 2\}} \tag{63}$$

which has density given by

$$f_U(u) = \frac{1}{2}\mathbf{1}_{\{u\in(0,2)\}} \tag{64}$$

Denoting the joint probability density of T and U by $f_{(T,U)}(t, u)$ and the marginal density of T by $f_T(t)$, we have that

$$\begin{aligned}
Pr(U < T) &= \int_0^2 \int_u^\infty f_{(T,U)}(t,u)dtdu \\
&= \int_0^2 \int_u^\infty f_T(t)f_U(u)dtdu && \text{(by independence)} \\
&= \int_0^2 \int_u^\infty \frac{1}{2} f_T(t)dtdu && \text{(by 64)} \\
&= \int_0^\infty \int_0^{\min(t,2)} \frac{1}{2} f_T(t)dudt && \text{(since } T \geq 0\text{, almost everywhere)} \\
&= \int_0^\infty \frac{1}{2}\min(t,2) f_T(t)dt && \text{(by 64)} \\
&\leq \frac{1}{2}\int_0^\infty t f_T(t)dt \\
&= \frac{1}{2}E(T)
\end{aligned}$$

□

The following proof for our theorem, found in Ferguson [10], can now be given.

Proof. By symmetry, if a strategy is optimal to one player is must be optimal to the other (as player A wants to maximise $Pr(W_A > W_B)$, player B wishes to minimise it). So, any optimal strategy must require that $Pr(W_A \geq W_B) = 1/2$. Now, if we suppose that player A follows the strategy as stated in the theorem, it suffices to show that $Pr(W_A < W_B) \leq 1/2$. We have that

$$\begin{aligned}
Pr(W_A < W_B) &= Pr(V_A[(1-a(V_A)) + a(V_A)2X] < V_B[(1-b(V_B)) + b(V_B)2X]) \\
&= Pr\bigg(U < \frac{V_B[(1-b(V_B)) + b(V_B)2X]}{[(1-a(U)) + a(U)2X]}\bigg) \\
&= Pr\bigg(U < \frac{V_B[(1-b(V_B)) + b(V_B)2X]}{[(2(1-p) + 2(2p-1)X]}\bigg) \\
&\leq \frac{1}{2}E\bigg(\frac{V_B[(1-b(V_B)) + b(V_B)2X]}{[(2(1-p) + 2(2p-1)X]}\bigg) && \text{(by the lemma)} \\
&= \frac{1}{2}E\bigg(p\frac{V_B[(1+b(V_B)]}{2p} + (1-p)\frac{V_B[(1-b(V_B))]}{2(1-p)}\bigg) \\
&= \frac{1}{2}E(V_B) \\
&= \frac{1}{2} && \text{which gives us the desired result, and so proves the theorem.}
\end{aligned}$$

□

It must be emphasised that this result only holds if both players A and B are wagering on the *same* game- if they are i.i.d games this is not sufficient. An example of where the result will fail to hold if this condition is not met is also given in Ferguson [10]. Let

$p = 3/4$ and suppose that player A adopts the Kelly strategy, while Player B decides to not switch their wealth, and then aims to maximise their expected wealth (i.e. they bet their entire wealth).

Player A has the density of their wealth given by

$$f_{W_A}(w) = \frac{1}{2}\mathbf{1}_{\{w\in(0,1)\}} + \frac{1}{4}\mathbf{1}_{\{w\in(1,3)\}}$$

Player B has the probability mass function of their wealth given by

$$f_{W_B}(w) = \frac{1}{4}\mathbf{1}_{\{w=0\}} + \frac{3}{4}\mathbf{1}_{\{w=2\}}$$

Now, if player B has final wealth of 0, they lose almost surely. If player B has final wealth of 2, then they have a 3/4 probability of winning, due to independence. Hence, the probability of player B winning, is given by

$$Pr(W_B > W_A) = 0 \times \frac{1}{4} + \frac{3}{4} \times \frac{3}{4} = \frac{9}{16} > \frac{1}{2}$$

which implies that the strategy from the theorem is no longer optimal.

It is remarked in Bell and Cover that the randomisation of one's initial wealth (by switching to one's wealth for the uniform distribution) is a "purely game theoretic protection against competition" [1]. As such, Bell and Cover do not actually endorse the randomisation of one's wealth in practise, but instead simply advocate the Kelly betting system on its own.

12.4 Limitations of the Kelly Criterion

The most significant drawback with betting according to the Kelly Criterion is that, while wealth grows at the maximal rate, the bets increase (in absolute terms) in size. This results in great risk for the gambler as their wealth increases- something which may or may not be acceptable to the gambler, depending on their aversion/preference for risk [20]. This increase in risk is reflected by the fact that the variance in wealth increases exponentially as n approaches infinity, which one can observe in the simple example expounded in 8.2. Since

$$E\log(1 + lX_i) = p\log(1 + l) + (1 - p)\log(1 - l) = g(l)$$

we can deduce that

$$\begin{aligned}\text{Var}[\log(1 + lX_i)] &= E[\log(1 + lX_i)]^2 - [E\log(1 + lX_i)]^2 \\ &= p[\log(1 + l)]^2 + (1 - p)[\log(1 - l)]^2 - [p\log(1 + l) + (1 - p)\log(1 - l)]^2 \\ &= p(1 - p)[[\log(1 + l)]^2 + [\log(1 - l)]^2 - 2\log(1 + l)\log(1 - l)] \\ &= p(1 - p)\left[\log\left(\frac{1 + l}{1 - l}\right)\right]^2\end{aligned}$$

[20] Conversely, it has been argued that the Kelly Criterion may be too conservative; that is, some gamblers may actually have a greater preference for risk than that implied by the Kelly strategy [1].

which is clearly constant with respect to n; we shall denote this expression by s^2. Thus, since the X_i are independent and identically distributed, the variance of the logarithm of the bettor's wealth after the nth trial is given by:

$$\begin{aligned}\mathrm{Var}[\log(W_n)] &= \sum_{k=1}^{n} \mathrm{Var}[\log(1 + lX_i)] \\ &= ns^2\end{aligned}$$

Another limitation of the Kelly Criterion is the assumption that wealth is infinitely divisible. Systematic overbetting (relative to the Kelly Criterion) leads to a sub-optimal growth rate of wealth, while increasing risk [26]. This may indeed occur if the bettor wagers a slightly larger proportion of their wealth on successive bets as a consequence of rounding up (since we cannot infinitely divide money in the real world), or due to errors in the estimates of the probability distribution (if it is unknown) [27]. In fact, if the gambler exceeds the Kelly Bet by too much, they may encounter ruin with probability one, as the number of trials approaches infinity. If we consider our original example in section 8.2 and solve the equation below (where the right hand side is simply $g(l)$):

$$0 = p\log(1 + l) + (1 - p)\log(1 - l)$$

for l, numerically, we can find the solution $\hat{l}$ so that for $l > \hat{l}$ we have that $g(l) < 0$, which ensures that $W_n(l)$ approaches 0, almost surely (by our asymptotic results of W_n in 8.2) .

A further problem associated with Kelly betting is its disadvantage relative to constant betting (also known as flat betting), in the sense that the ratio of the expected return on investment (i.e. the expectation of the total wealth accumulated divided by the total wealth wagered) for Kelly betting against the expected return on constant betting approaches one half, as the bias in the gambler's favour approaches zero [8]. By constant betting we mean a strategy where an individual bets a fixed amount each round of betting (as opposed to a fixed fraction of one's wealth). So, in the context of our original example in 8.2, a flat bettor would, for all n trials, bet b dollars on the ith trial and receive bX_i dollars.

Ethier et al. were, in fact, able to demonstrate a slightly more general result [8], the arguments for which shall be briefly outlined. Firstly, they considered a sequence of independent and identically distributed random variables, as in 8.2, with the specification that $p = (1+\epsilon)/2$ where $\epsilon > 0$ (note that this implies that the Kelly bet is $l^* = \epsilon$), denoting these random variables $X_1(\epsilon), X_2(\epsilon)$.... Now, they considered the return on investment after the nth trial of the gambler who bets the proportion $l = \alpha\epsilon$, denoted by $R_n^\alpha(\epsilon)$, which is:

$$R_n^\alpha(\epsilon) = \frac{W_n - W_0}{\alpha\epsilon \sum_{k=0}^{n-1} W_k} \tag{65}$$

After proving that this converges in distribution as n approaches infinity to the random variable $R^{\alpha}(\epsilon) = (\alpha\epsilon\sum_{k=1}^{\infty}(W_0/W_k))^{-1}$, they then demonstrate a further convergence (in distribution) result as $\epsilon \to 0^{+}$, regarding the ratio of the return on investment for the "alpha" Kelly strategy and the expectation of the return on investment for the constant bettor (noting that the expected return on investment for the constant bettor is $E\sum_{i=1}^{n} bX_i/bn = EX_1$):

$$R^{\alpha}(\epsilon)/EX_1(\epsilon) \xrightarrow{d} \gamma\left(\frac{2}{\alpha} - 1, \frac{2}{\alpha}\right)$$

where $\gamma(a, b)$ is a Gamma random variable, with density on the interval $(0, \infty)$ given by $b^a x^{a-1} \exp(-bx)/\Gamma(a)$, and $\Gamma(a)$ is the integral $\int_0^{\infty} x^{a-1} \exp(-x)dx$.

From this result, they demonstrate that the class of random variables of the form R^{α}/EX_1 is uniformly integrable in ϵ, which implies that the limit of the expectation is equal to the expectation of the limit. The limit is the expectation of the Gamma random variable, which is $(2/\alpha - 1)/(2/\alpha) = 1 - \alpha/2$. Hence, letting $\alpha = 1$, we yield the result as claimed for the Kelly bettor:

$$\lim_{\epsilon \to 0^{+}} ER(\epsilon)/EX_1(\epsilon) = \frac{1}{2}$$

This result is not necessarily a significant shortcoming of Kelly betting since by following such a strategy one will never face ruin (except in the asymptotic sense, where it occurs with probability zero in any case), whereas the constant bettor may indeed experience ruin. Thus, by employing the Kelly betting strategy one may consider the lower return on investment as being the cost of insurance against the possibility of ruin [8].

Another issue with Kelly betting worth noting concerns the coin tossing game, as discussed in section 8.2, and is mentioned in MacLean et al. [26]. If the gambler stops betting after the $2n$th trial and the amount of successes equals the amount of failures, then the bettor will experience a fall in wealth, since in this case we would have $W_n = W_0(1-l^2)^n < W_0$. This is particularly relevant, just as in the previous limitation described, for games in which the probability of winning, p, is only slightly above one half.

One final, and considerable, limitation of Kelly betting is that many of its optimal properties only hold asymptotically; that is, it may take a significantly large amount of trials before the advantages from using the Kelly system eventuate. Browne [5] considers an investor faced with the problem of choosing what amount of their wealth should be allocated towards a risk-free (cash) asset, which has a fixed annual rate of return of 7%, versus a risky asset (stock) which has rates of return determined by a sequence of independent and identically distributed normal random variables of mean 15% and standard deviation of 30%. Browne then found the number of trials (days) before one could be at least 95% sure that the Kelly strategy had delivered a level of wealth at least 10% higher than simply investing all of the wealth in cash (and likewise for stock); these numbers were 157 and 10, 286 years, respectively.

12.5 Connection with Utility Theory

In the field of economics, it is supposed that the individual investor (or gambler) possess a utility function of their wealth and so will allocate their financial resources in such a way as to maximize their expectation of this utility function. It is possible that a given investor will be confronted with a sequence of single-period investment decisions, and as such will strive to maximize the expectation of their final utility (the utility of their wealth after the final period). A *myopic* strategy is one in which the investor treats every period in the sequence as if it were the last one, without looking further beyond that step [34]. One would not expect such behaviour to be necessarily optimal: optimising at each successive step is no guarantee that one will optimize over the entire sequence of investment decisions. However, there is a particular class of utility functions for which, interestingly, the myopic strategy is optimal. These are the logarithmic and power functions [14]:

$$U_\gamma(x) = x^\gamma/\gamma, \qquad \gamma \leq 1 \tag{66}$$

Notice that $\lim_{\gamma \to 0^+} U_\gamma(x) = \log(x)$. The fact that the optimal system of investing is myopic is particularly advantageous, since the decision maker need only have knowledge of the forthcoming gamble in order to behave optimally in the long-run [14].

Now, it is clear that if an individual has a utility function $U_0(x)$, where x denotes their wealth, then the objective for this individual is to maximise

$$E[\log(x)]$$

and so we yield the Kelly betting (investing) system for this individual, which is reasonable since the Kelly system is clearly myopic optimal.

The class of functions in 67 also has the property of *constant relative* risk aversion, defined to be $R(x) = -xU''_\gamma(x)/U'_\gamma(x)$. This is a standard measure of risk aversion in utility theory; it has the desirable property that for any two individuals with utility functions u, v , if their respective relative risk aversion values are $R_u(x), R_v(x)$ with $R_u(x)/x > R_v(x)/x, \forall x$, then individual u is more risk averse than individual v (more risk averse in the sense that if a gamble would not be taken by individual v, then it would also not be taken by individual u) [38]. From this fact we can observe how the individual with a utility function of $U(x) = \log(x)$ will take greater risks (in the form of larger bets) as their wealth increases, since their *absolute* risk aversion[26], defined to be $x^{-1}R(x)$, equals $1/x$ and so falls as their wealth, x, increases. So we see that the perspective from utility theory concurs with our previous understanding that the Kelly betting system is, indeed, a risky one.

Another utility theoretic concept can help us further analyse different betting strategies. A betting system A *stochastically dominates* another betting system B if, denoting the wealth attained by employing these systems as W_A and W_B, respectively, we have that for all levels of wealth x, $Pr(W_A > x) \geq Pr(W_B > x)$,with strict inequality holding for at least one such x[23]. This is an important concept in utility theory as it allows one to rank strategies: strategy A stochastically dominates strategy B if and only if all individuals with increasing utility functions will choose (by the criteria of maximising their expected utility) strategy A over B [23]. It turns out that the Kelly betting system is not stochastically dominated by any other strategy, giving it further justification from a utility theory perspective [1].

However, there is much controversy surrounding the interpretation of $\log(x)$ as a utility function, and much criticism directed at some of the assertions made about the Kelly betting system as it relates to economic theory [40, 41, 1, 22, 14, 36]. The suggestion that investors in the real world behave, or indeed *ought* to behave, as if they possessed log utility has been strongly opposed by economists [41, 14]. As is noted by Samuelson and Hakansson, the degree of risk aversion implied by the logarithmic utility is too low to justify it being an accurate reflection of investors' actual behavior; empirical findings suggest that investors in the real world have a level of risk aversion that is comparable to that implied by the function $U_{-3}(x)$ [14], from definition 67. As a consequence of this scepticism, economists have also mentioned their dissatisfaction [40] regarding the suggestion that the function $\log(x)$ is a reasonable solution to the famous St Petersburg Paradox, as is claimed in Bell. [1] [21]

Proponents of the Kelly approach attest to its prescriptive results; it can tell the bettor what, and how much, to invest in with greater clarity than simply the directive to maximize one's utility [22]. Indeed, according to Latane, given its optimal properties the objective of maximizing the expected logarithm of wealth ought to be a substitute for the economics profession's favored method of simply maximizing expected utility [22]. Although, this too has been the subject of great criticism [36, 41, 40], with Samuelson and Ophir stressing that the optimal asymptotic properties of Kelly betting do not automatically sanction it as being an objectively superior criteria for investment; that investors may (and, indeed, often) have goals that are not related to an infinite horizon. Aware of such criticism, Bell and Cover [1] justify their advocacy of the Kelly betting system, neither on the premise that it may be interpreted as a utility function, nor on Latane's reasoning, but on the basis that it represents an optimal course of action for serving the plausible goals of an investor to outperform other strategies asymptotically. [1]

12.6 Modifications and Alternatives

Given the previously mentioned shortcomings of the Kelly betting system, several modifications and alternatives to this system have been devised. They have their own limitations, however; usually their drawback is that of compromising the asymptotic performance of the bettor's wealth.

One particular modification is what is known as *fractional Kelly betting.* In this case, the bettor only wagers a scalar multiple $\alpha \in (0, 1)$ of the fraction of wealth wagered by the Kelly bettor (for example, our discussion of the "alpha" bettor in section 8.3). This strategy enables the bettor to avoid placing large bets that may be too risky from the

[21] This is the famous "paradox" from the 18th century where, hypothetically, a gambler is offered a bet with infinite expectation, but is unwilling to pay an arbitrarily large amount to play. The suggested explanation, given by Daniel Bernoulli, [1] is that the gambler does not seek to maximize their expected wealth, but rather their expected *utility* of wealth, where their utility is some concave function of wealth- for instance, the function $\log(x)$. The discussion around possible solutions to the paradox is an interesting topic in itself; a comprehensive study of this is found in Samuelson [40].

bettor's perspective, while still allowing for growth in wealth (albeit at a slower rate than a pure Kelly strategy, by our previous results regarding Kelly's asymptotic dominance over other strategies). However, simulations in Maclean et al [25] demonstrate that while Kelly betting and fractional Kelly betting where α is large tends to result in more chances of larger gains than conservative fractional Kelly betting, in the latter case we observe a much smaller probability of having a devastating collapse in wealth due to the result of a sequence of unfavorable outcomes. Moreover, the pure Kelly strategy fails to stochastically dominate fractional strategies [25], demonstrating that the pure Kelly strategy is not strictly superior to fractional betting strategies, from a utility theory perspective [23].

There is a further connection between the fractional Kelly betting system and utility theory. Consider a problem where an investor must choose between allocating their resources into either a risk-free asset (such as cash) that returns some fixed rate, or a risky asset (stock) which has random returns. Furthermore, suppose that this investor possesses a utility function of the form

$$U_\delta(x) = \delta x^\delta, \qquad \gamma < 0 \tag{67}$$

Then, letting $\alpha = 1/(1-\delta)$ we have that the optimal decision is to invest approximately α times their wealth in the Kelly allocation, while allocating approximately $1-\alpha$ of their wealth in cash [26, 24] (so, for example, one can approximate a "half" Kelly strategy by maximizing $U_\delta(x)$ for $\delta = -1$). Thus, we have a utility theory interpretation not only of Kelly betting, but fractional Kelly betting as well (at least for the games in which the approximation is valid [22]).

Another alternative to the Kelly betting system is that which is mentioned in Ferguson [9], where one considers the bettor's objective to be that of maximizing the utility function $V_\theta(W) = -\exp(-\theta W)$, $\theta > 0$. This objective is used in the Billie Sol Estes model, which is a slight deviation from our original game from 8.2 in that the gambler is allowed to both wager amounts that exceed their wealth, and still play if their wealth falls below zero (however, they still can only bet positive amounts) [9]. Corresponding to this problem (supposing $p > 1/2$), the betting system that maximizes their utility is given by:

$$b = \frac{1}{2\theta}\log\left(\frac{p}{1-p}\right)$$

where b is the amount of dollars bet at any trial.

To prove this, note that

$$EV_\theta(W) = -p\exp(-\theta(W+b)) - (1-p)\exp(-\theta(W-b)) \tag{68}$$

By taking partial derivative of expression 68 with respect to b and setting this to zero we have that:

[22] For the coin tossing game previously discussed, the approximation is poor; however, for when the random variable representing the unknown gain/loss has a lognormal distribution, we in fact have equality [26]

$$0 = p\theta \exp(-\theta(W + b)) - (1 - p)\theta \exp(-\theta(W - b))$$

which, after rearranging and dividing by $\theta \exp(-\theta W)$ implies that

$$\exp(-\theta b) = \frac{(1 - p)}{p} \exp(\theta b))$$

Then, we simply take the logarithm of both sides and rearrange to yield the desired result

A noticable advantage of using this particular betting system is that the prescribed bet is independent of the total amount of resources the bettor possesses. So, the gambler need not keep track of their own wealth as they gamble. Furthermore, it is a quick calculation to check that the absolute risk aversion, given by the previously mentioned Arrow-Pratt formula $x^{-1}R(x) = -V_\theta''(x)/V_\theta'(x)$, equals $\theta > 0$ [38]. Since the log utility function exhibits decreasing absolute risk aversion, it is clear that given a utility function of the above form there will exist some level of wealth, w_0, such that this utility function will have a greater level of absolute risk aversion than $\log(x)$. Consequently, we can think of these utility functions as, asymptotically ($w \to \infty$), being reflective of a more risk averse individual than one that possesses log utility (however, for low levels of wealth, betting according to b is more risky than the Kelly strategy, in the sense that the size of bets will be larger and, in the event of an unfavorable outcome, may result in negative levels of wealth).

12.7 Concluding Remarks

Breiman's extension of the results we have already introduced in 8.2 are especially pertinent to our ultimate objective of constructing a method of betting on horse races, where we almost always have more than two choices on which to bet, and which have different odds. If our previous work estimating the probability of a horse j winning race i implies that closing odds do not accurately reflect the probability of a horse's victory, then it is possible that such information will allow us to find bets with positive expected return [42]. As such, we can employ the Kelly betting strategy (or an alternative) to exploit this bias in our favour.

13 Final Remarks

We have discussed the theory necessary to extend a model for probability of horses' victory in a given race via a combination of using a discrete choice model with smooth functions in the score function. The key task ahead is to extend its application to considering a sequence of races where the number of horses is not constant. Then, one would be able to provide a more realistic model for actual betting markets.

Consequently, one could go further and 'bet' according to the Kelly system on historical data (as was done in Johnson [18]) so to surmise whether positive could indeed be obtained by simply using a statistical model to estimate probabilities and an accompanying betting system to dictate how much to bet on what horse. A further extension of the model would be to consider smooths that are not univariate, so that we may capture the influence of combined effects of predictors (bettors are thought not to take into account such combined effects in their wagers, thus there is a possibility that such a discrepancy could be exploited [18].

References

[1] Robert M Bell and Thomas M Cover. Competitive optimality of logarithmic investment. *Mathematics of Operations Research*, 5(2):161–166, 1980.

[2] Michael J Boskin. A conditional logit model of occupational choice. *The Journal of Political Economy*, 82(2):389–398, 1974.

[3] Leo Breiman. Optimal gambling systems for favorable games, 1961.

[4] Leo Breiman. *Probability*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[5] Sid Browne. Can you do better than kelly in the short run. *Finding the Edge, Mathematical Analysis of Casino Games*, pages 215–231, 2000.

[6] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

[7] Gerard Debreu. Review of rd luce, individual choice behavior: A theoretical analysis. *American Economic Review*, 50(1):186–188, 1960.

[8] SN Ethier and S Tavare. The proportional bettor's return on investment. *Journal of Applied Probability*, pages 563–573, 1983.

[9] Thomas S Ferguson. Betting systems which minimize the probability of ruin. *Journal of the Society for Industrial & Applied Mathematics*, 13(3):795–818, 1965.

[10] Thomas S Ferguson and Costis Melolidakis. Last round betting. *Journal of Applied Probability*, pages 974–987, 1997.

[11] Milton Friedman and Leonard J Savage. The utility analysis of choices involving risk. *The Journal of Political Economy*, 56(4):279–304, 1948.

[12] Peter J Green, Bernard W Silverman, Bernard W Silverman, and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman & Hall London, 1994.

[13] Chong Gu. *Smoothing spline ANOVA models*. Springer, 2002.

[14] Nils H Hakansson and William T Ziemba. Capital growth theory. *Handbooks in Operations Research and Management Science*, 9:65–86, 1995.

[15] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical science*, pages 297–310, 1986.

[16] Donald B Hausch, William T Ziemba, and Mark Rubinstein. Efficiency of the market for racetrack betting. *Management science*, 27(12):1435–1452, 1981.

[17] Carolyn J Heinrich and Jeffrey B Wenger. The economic contributions of james j. heckman and daniel l. mcfadden. *Review of Political Economy*, 14(1):69–89, 2002.

[18] Johnnie EV Johnson, Owen Jones, and Leilei Tang. Exploring decision makers' use of price information in a speculative market. *Management Science*, 52(6):897–908, 2006.

[19] J Kelly Jr. A new interpretation of information rate. *Information Theory, IRE Transactions on*, 2(3):185–189, 1956.

[20] Jaromir Joseph Koliha. *Metrics, norms and integrals: an introduction to contemporary analysis.* World Scientific, 2008.

[21] Erwin Kreyszig. *Introductory functional analysis with applications.* Wiley. com, 2007.

[22] Henry Allen Latane. Criteria for choice among risky ventures. *The Journal of Political Economy*, 67(2):144–155, 1959.

[23] Haim Levy. *Stochastic dominance [electronic resource]: investment decision making under uncertainty*, volume 12. Springer, 2006.

[24] LC MacLean and WT Ziemba. Li (2005). time to wealth goals in capital accumulation and the optimal trade-off of growth versus security. *Quantitative Finance*, 5(4):343–357.

[25] Leonard C MacLean, Edward O Thorp, Yonggan Zhao, and William T Ziemba. How does the fortune's formula kelly capital growth model perform? *Journal of Portfolio Management*, 37(4):96, 2011.

[26] Leonard C MacLean, Edward O Thorp, and William T Ziemba. Good and bad properties of the kelly criterion. *Risk*, 20(2):1, 2010.

[27] Leonard C MacLean and William T Ziemba. Capital growth: Theory and practice. *Handbook of Asset and Liability Management*, 1:429–474, 2006.

[28] Gangadharrao S Maddala. *Limited-dependent and qualitative variables in econometrics.* Number 3. Cambridge University Press, 1983.

[29] Burton G Malkiel and Eugene F Fama. Efficient capital markets: A review of theory and empirical work*. *The journal of Finance*, 25(2):383–417, 1970.

[30] Colin L Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973.

[31] P McCullough and JA Nelder. Generalized linear models chapman and hall. *New York*, 1989.

[32] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1973.

[33] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques, and tools.* Princeton university press, 2010.

[34] Jan Mossin. Optimal multiperiod portfolio policies. *The Journal of Business*, 41(2):215–229, 1968.

[35] Efe A Ok. *Real analysis with economic applications*, volume 10. Princeton University Press, 2007.

[36] Tsvi Ophir. The geometric-mean principle revisited. *Journal of Banking & Finance*, 2(1):103–107, 1978.

[37] Timothy J Pleskac. Decision and choice: Luce's choice axiom. 2012.

[38] John W Pratt. Risk aversion in the small and in the large. *Econometrica: Journal of the Econometric Society*, pages 122–136, 1964.

[39] Christian H Reinsch. Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183, 1967.

[40] Paul A Samuelson. St. petersburg paradoxes: Defanged, dissected, and historically described. *Journal of Economic Literature*, 15(1):24–55, 1977.

[41] Paul A Samuelson. Why we should not make mean log of wealth big though years to act are long. *Journal of Banking & Finance*, 3(4):305–307, 1979.

[42] Raymond D Sauer. The economics of wagering markets. *Journal of economic Literature*, 36(4):2021–2064, 1998.

[43] Peter Schmidt and Robert P Strauss. The prediction of occupation using multiple logit models. *International Economic Review*, 16(2):471–486, 1975.

[44] Edward O Thorp. Optimal gambling systems for favorable games. *Revue de l'Institut International de Statistique*, pages 273–293, 1969.

[45] Edward O Thorp. The kelly criterion in blackjack, sports betting, and the stock market. *Finding the Edge: Mathematical Analysis of Casino Games*, pages 163–213, 2000.

[46] Leighton Vaughan Williams. *Information efficiency in financial and betting markets.* Cambridge University Press, 2005.

[47] Simon Wood. *Generalized additive models: an introduction with R.* CRC Press, 2006.