Data Reduction for Efficient Probit Regression

Christian Peters

August 23, 2021

Contents

1 Introduction

Content.

2 The Probit Model

[Fahrmeir et al., 2013] [McCullagh and Nelder, 1989] [Agresti, 2015] [Gelman et al., 2013] [Albert and Chib, 1993]

2.1 Introduction as a Latent Variable Model

Suppose we are given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ containing n pairs of observations. We assume that the x_i are d-dimensional vectors, i.e. $x_i \in \mathbb{R}^d$ that contain explanatory information regarding the binary outcome $y_i \in \{0, 1\}$.

In one possible scenario perhaps, the x_i are used to represent information about a patient, such as blood pressure or weight, and the y_i are used to indicate the presence or absence of a heart disease. In such a setting, we are often interested in modeling the relationship between explanatory values of x_i and the binary outcomes y_i. The probit model is one of many approaches to accomplish this, it's core assumptions and properties are described below.

Since it is reasonable to assume that there is a degree of randomness involved in the data generating process, we model the y_i as realizations of independent random variables Y_i, which brings us to the first assumption in the probit model: We assume that the outcomes y_i were generated independently from each other, i.e. the occurence of one outcome did not affect the other outcomes.

The second assumption of the probit model is that there is a hidden random quantity Y_i^* that is associated with each Y_i such that it directly determines its outcome:

$$Y_i = \begin{cases} 1, & \text{if } Y_i^* > 0 \\ 0, & \text{if } Y_i^* \leq 0 \end{cases} \tag{1}$$

The Y_i^* are also assumed to be independent from each other and, as already noted, unobservable. This is why the Y_i^* are also called latent variables.

The third and final assumption of the probit model defines the distribution of the Y_i^* and its part of the relationship between the non-random explanatory variables x_i and the outcomes y_i. In order to describe this relationship more concisely, we put all the observations x_i inside of a matrix $X \in \mathbb{R}^{n \times d}$ in such a way that the i-th row of X corresponds to x_i. In the literature, this matrix X is often called the **model matrix**. We do the same with the Y_i^* and put them in a random vector Y^* as well, such that Y_i^* constitutes the i-th element of Y^*.

We are now ready for the third assumption of the probit model: The explanatory variables x_i influence Y_i^* in the form of a classical linear model:

$$Y^* = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{2}$$

where $\beta \in \mathbb{R}^d$ is the parameter vector of the linear model, ϵ is a normal distributed vector with independent components of mean zero and variance σ^2, and $I \in \mathbb{R}^{n \times n}$ is the $n \times n$ identity matrix. It follows directly that Y^* is also normal distributed: $Y^* \sim \mathcal{N}(X\beta, \sigma^2 I)$.

These three assumptions are already a complete specification of the probit model and are summarized in the following definition as a brief recapitulation:

Definition 1 (Probit Model)**.** *A dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with model matrix $X \in \mathbb{R}^{n \times d}$ and observed outcomes $y_i \in \{0, 1\}$, $i \in [n]$, was generated by a probit model with parameters $\beta \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}_{>0}$, if the following three assumptions are true:*

1. *The observations $y_1, ..., y_n$ are realizations of independent binary random variables $Y_1, ..., Y_n$.*
2. *The outcomes of $Y_1, ..., Y_n$ are determined by hidden continuous random variables $Y_1^*, ..., Y_n^*$ by thresholding: If $Y_i^* > 0$, then $Y_i = 1$, and if $Y_i^* \leq 0$, then $Y_i = 0$.*
3. *The vector of hidden variables Y^* follows a multivariate normal distribution: $Y^* \sim \mathcal{N}(X\beta, \sigma^2 I)$, where $\beta \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}_{>0}$ are the model parameters.*

Based on this definition, it is straight forward to determine the distribution of the response variables Y_i. We can calculate the probability $P(Y_i = 1)$ like this:

$$P(Y_i = 1) = P(Y_i^* > 0) = 1 - P(Y_i^* \leq 0) = 1 - P\left(\frac{Y_i^* - x_i\beta}{\sigma} \leq -\frac{x_i\beta}{\sigma}\right) = \Phi\left(\frac{x_i\beta}{\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz.$$

This result $P(Y_i = 1) = \Phi\left(\frac{x_i\beta}{\sigma}\right)$ leads us to an interesting observation: Both parameters β and σ are unknown model parameters and every value of σ can be compensated by a corresponding scaling of β. This means that, because we can't observe the hidden variables Y_i^*, it is impossible to determine which β and which σ generated the data without any prior knowledge. We can only draw conclusions with regard to the scaled parameter $\frac{1}{\sigma}\beta$. In this situation, we say that β and σ are *not identifiable.*

For this reason, in standard literature like [Fahrmeir et al., 2013] it is often argued that without the loss of generality we can assume that $\sigma = 1$ and arrive at

$$P(Y_i = 1) = \Phi(x_i\beta). \tag{3}$$

Since Y_i is binary, it follows that

$$P(Y_i = 0) = 1 - P(Y_i = 1) = 1 - \Phi(x_i\beta) = \Phi(-x_i\beta),$$

which immediately leads us to the model equations

$$Y_i \sim Bin(1, \pi_i), \quad \pi_i = \Phi(x_i\beta), \tag{4}$$

where $Bin(1, \pi_i)$ is a bernoulli distribution with success probability $\pi_i = \Phi(x_i\beta)$.

2.2 A Special Case of the Generalized Linear Model

The final equations of the probit model that we arrived at in equation 4 are a special case of a more general model concept, the generalized linear model (GLM), that we briefly touch on below.

Generalized linear models consist of three components. The first one is the so called **random component**, a set of $n \in \mathbb{N}$ independent random variables $\{Y_i\}_{i=1}^n$. In GLMs, the distribution of these random variables is assumed to be a member of the *exponential family*, a broad family of probability distributions that encompasses the normal distribution, the binomial distribution and many others. It is characterized in more detail in [Agresti, 2015].

The second component of a GLM is the **linear predictor**. Just like in the probit model, we also assume that we are presented with some fixed observations $\{x_i \in \mathbb{R}^d\}_{i=1}^n$, that are assumed to have some explanatory power with regard to the Y_i. We thus call these observations the explanatory variables. The linear predictor will be used to relate the explanatory variables to the distribution of the Y_i by linearly combining them as follows:

$$\eta_i = x_i \beta,$$

where $\eta_i \in \mathbb{R}$ denotes the linear predictor related to observation x_i and $\beta \in \mathbb{R}^d$ is the unknown parameter vector of the GLM that has to be estimated when fitting the model.

The third component of a GLM is the so called **link function**. This is a monotonic and differentiable function g that connects the linear predictor η_i to the distribution of the Y_i like this:

$$g(E[Y_i]) = \eta_i.$$

We are thus using the link function g to transform the expected value $E[Y_i]$ in such a way that it can be predicted by a linear model, hence the name *generalized linear models*.

Equivalently, we can also characterize this relationsthip by using the inverse function $h = g^{-1}$, also called the **response function**:

$$E[Y_i] = h(\eta_i).$$

Due to the requirement that g is monotonic and differentiable, this function h always exists.

We are now ready to establish the connection between the probit model and the generalized linear model. As we saw in equation 4, the assumptions of the probit model imply that the Y_i follow independent binomial distributions with a success probability of $\pi_i = \Phi(x_i\beta)$. The binomial distribution is a member of the exponential family, so we can also think of the Y_i as the random component of a GLM.

It also follows directly from the binomial distribution that $E[Y_i] = \pi_i$, thus we have from the probit model equations that $\pi_i = E[Y_i] = \Phi(x_i\beta)$, and equivalently $\Phi^{-1}(E[Y_i]) = x_i\beta$. Thus, we can think of Φ as the response function of a GLM and Φ^{-1} as the link function, which completes the specification of the probit model as a special case of GLMs.

2.3 Parameter Estimation

Generalized linear models and therefore the probit model are usually estimated by using the **maximum likelihood** method. This method seeks to maximize the probability that some observed data $y_1, ..., y_n$, $n \in \mathbb{N}$ was generated under the assumption of the model and given some parameter vector $\beta \in \mathbb{R}^d$.

To make notation a little easier, we also put the outcomes y_i in a vector $y \in \{0, 1\}^n$ such that y_i is the i-th component of y. In the same way, we also put the random variables Y_i inside of a random vector Y.

In the probit model, the likelihood function is given as

$$\mathcal{L}(\beta) = P(Y = y; \beta) = \prod_{i=1}^{n} P(Y_i = y_i; \beta), \tag{5}$$

because the Y_i are independent. By using a little trick, we can write $P(Y_i = y_i; \beta)$ like this:

$$P(Y_i = y_i; \beta) = \Phi[(2y_i - 1)x_i\beta],$$

which enables us to arrive at the likelihood

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} P(Y_i = y_i; \beta) = \prod_{i=1}^{n} \Phi[(2y_i - 1)x_i\beta] = \prod_{i=1}^{n} \Phi(-z_i\beta). \tag{6}$$

Here, we introduced the new vector $z_i = -(2y_i - 1)x_i$ to simplify the notation. The maximum likelihood estimate for β is then given by

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmax}}\ \mathcal{L}(\beta), \tag{7}$$

and for $n \to \infty$ it holds that $E[\hat{\beta}] = \beta$ [Fahrmeir et al., 2013].

However, for finite sample sizes, the existence of $\hat{\beta}$ cannot be guaranteed and is dependent on the observed data. An overview of the conditions for the existence and uniqueness of $\hat{\beta}$ is given in [Demidenko, 2001]. In particular, there is one important condition shown in [Lesaffre and Kaufmann, 1992], that is related to the concept of linear separability which we introduce in the following definition.

Definition 2 (Linear separability). *Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be a dataset with $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ for all $i \in [n]$. Let $S_0 = \{i \in [n] : \ y_i = 0\}$ and $S_0 = \{i \in [n] : \ y_i = 1\}$. If there exists a $\beta \in \mathbb{R}^d$ such that*

$$\forall i \in S_0 : \ x_i\beta \leq 0 \quad \textit{and} \quad \forall i \in S_1 : \ x_i\beta > 0,$$

then we call $\mathcal{D}$ linearly separable.

Intuitively speaking, a dataset is linearly separable if there exists a hyperplane that perfectly separates the datapoints labeled with 1 from the datapoints labeled with 0. This property of a dataset can be used to infer the existence of the maximum likelihood estimate $\hat{\beta}$ as stated in the following theorem.

Theorem 1 ([Lesaffre and Kaufmann, 1992]). *Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a dataset like in definition 2 and let $X \in \mathbb{R}^{n \times d}$ be the corresponding model matrix. Under the condition that $rank(X) = d$, i.e. X has full column-rank, the maximum likelihood estimate $\hat{\beta}$ for the parameter β of the probit model exists if and only if $\mathcal{D}$ is not linearly separable.*

Theorem 1 states, that the maximum likelihood estimate always exists if the dataset is not linearly separable, provided the model matrix X has full column-rank.
In [Wedderburn, 1976], it was further shown, that if the maximum likelihood estimate exists, then it is also unique. It now remains to explore, how the maximum likelihood optimization problem can be solved in such a case.

2.3.1 Finding the Maximum Likelihood Estimate

For the reason that the likelihood function $\mathcal{L}(\beta)$ is numerically inconvenient to maximize, the natural logarithm is often applied as a transformation to simplify the optimization problem:

$$\ell(\beta) = \ln \mathcal{L}(\beta) = \sum_{i=1}^{n} \ln \Phi(-z_i\beta). \tag{8}$$

Since we later wish to interpret ℓ as a loss function, we prefer to minimize the negative value of ℓ rather than maximizing:

$$f(\beta) = -\ell(\beta) = \sum_{i=1}^{n} \ln\left(\frac{1}{1-\Phi(z_i\beta)}\right) = \sum_{i=1}^{n} g(z_i\beta). \tag{9}$$

Here, we define $g(x) = \ln\left(\frac{1}{1-\Phi(x)}\right)$ and call it the **probit loss**, i.e. the loss-function that determines how much each z_i contributes to the total loss $f(\beta)$ for a given value of β. We call $f(\beta)$ the objective function of the probit model.

The optimization of f is usually done by applying the Newton-Raphson algorithm, an iterative procedure that starts at some initial guess $\beta^{(0)}$ and successively updates it like this:

$$\beta^{(t)} = \beta^{(t-1)} - \left(\frac{\partial^2 f(\beta^{(t-1)})}{\partial\beta\partial\beta^T}\right)^{-1} \cdot \frac{\partial f(\beta^{(t-1)})}{\partial\beta}, \tag{10}$$

where $\left(\frac{\partial^2 f(\beta^{(t-1)})}{\partial\beta\partial\beta^T}\right)^{-1}$ refers to the inverse of the hessian matrix of f, evaluated at $\beta^{(t-1)}$. Likewise, $\frac{\partial f(\beta^{(t-1)})}{\partial\beta}$ refers to the gradient of f, evaluated at $\beta^{(t-1)}$. The idea behind this procedure is, broadly speaking, to approximate f locally around $\beta^{(t)}$ as a second degree taylor-polynomial and then analytically find the minimum of this polynomial. The minimum of this local polynomial approximation of f is then iteratively used as the basis for the next step of the Newton-Raphson algorithm.

It remains to find the gradient as well as the hessian matrix of f. Because f is a sum of the function g evaluated at different points, it makes sense to first determine the derivative of g. This can be accomplished by using the chain rule as follows:

$$\begin{aligned}\frac{d}{dx}g(x) &= \frac{d}{dx}\ln\left(\frac{1}{1-\Phi(x)}\right) \\ &= (1-\Phi(x))\cdot\frac{d}{dx}\left(\frac{1}{1-\Phi(x)}\right) \\ &= (1-\Phi(x))\cdot\frac{(-1)}{(1-\Phi(x))^2}\cdot\frac{d}{dx}(1-\Phi(x)) \\ &= \frac{(-1)}{1-\Phi(x)}\cdot(-1)\cdot\phi(x) \\ &= \frac{\phi(x)}{1-\Phi(x)},\end{aligned} \tag{11}$$

where $\phi(x)$ is the density function of the standard normal distribution function:

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}.$$

We can use this result to calculate the gradient of f:

$$\begin{aligned}\frac{\partial}{\partial\beta}f(\beta) &= \frac{\partial}{\partial\beta}\sum_{i=1}^{n} g(z_i\beta) \\ &= \sum_{i=1}^{n} z_i\cdot\frac{\partial}{\partial\beta}g(z_i\beta) \\ &= \sum_{i=1}^{n} z_i\cdot\frac{\phi(z_i\beta)}{1-\Phi(z_i\beta)}\end{aligned} \tag{12}$$

Next, we need to determine the hessian matrix of f. In order to do this, we again start by finding the second derivative of g, this time using the quotient rule:

$$\begin{aligned}\frac{d^2}{dx^2}g(x) &= \frac{d}{dx}\frac{\phi(x)}{1-\Phi(x)} \\ &= \frac{\phi'(x)(1-\Phi(x))-\phi(x)\cdot(-1)\cdot\phi(x)}{(1-\Phi(x))^2} \\ &= \frac{(-1)\cdot x\cdot\phi(x)(1-\Phi(x))-\phi(x)\cdot(-1)\cdot\phi(x)}{(1-\Phi(x))^2} \\ &= \frac{[\phi(x)]^2-x\cdot\phi(x)\cdot(1-\Phi(x))}{(1-\Phi(x))^2} \\ &= \left(\frac{\phi(x)}{1-\Phi(x)}\right)^2 - x\cdot\frac{\phi(x)}{1-\Phi(x)} \\ &= \frac{\phi(x)}{1-\Phi(x)}\left(\frac{\phi(x)}{1-\Phi(x)}-x\right) \\ &= g'(x)\cdot(g'(x)-x)\end{aligned} \tag{13}$$

We can now use this result to find the hessian matrix of f:

$$\begin{aligned}\frac{\partial^2}{\partial\beta\partial\beta^T}f(\beta) &= \sum_{i=1}^{n}\frac{\partial^2}{\partial\beta\partial\beta^T}g(z_i\beta)\\ &= \sum_{i=1}^{n} z_i z_i^T g'(z_i\beta)(g'(z_i\beta) - z_i\beta) \\ &= \sum_{i=1}^{n} z_i z_i^T \frac{\phi(z_i\beta)}{1-\Phi(z_i\beta)}\left(\frac{\phi(z_i\beta)}{1-\Phi(z_i\beta)} - z_i\beta\right).\end{aligned} \tag{14}$$

Because it can be shown that, if the maximum likelihood estimate exists, $f(\beta)$ is a convex function [Wedderburn, 1976], and that the newton raphson algorithm converges to the global optimum when applied to a convex function [Nocedal and Wright, 2006], the optimization procedure converges to the maximum likelihood estimate $\hat{\beta}$ under the conditions that X has full column rank and that the data is not linearly separable.

2.4 The Bayesian Perspective

The most fundamental difference between the bayesian approach to the probit model and the frequentist approach that was discussed above, is the assumption, that the model parameter β is not a fixed value, but a random variable with a probability distribution. The goal of bayesian data analysis is to draw conclusions about the distribution of the model parameter and to update these conclusions after observing more and more data.

To characterize the prior uncertainty about the model parameter β, the first step of bayesian data analysis is to specify a so called **prior distribution**. In the probit model, one common choice that is also described in [Fahrmeir et al., 2013] is to assume that

$$\beta \sim \mathcal{N}(\mu_\beta, \Sigma_\beta), \tag{15}$$

i.e. β follows a normal distribution with mean $\mu_\beta \in \mathbb{R}^d$ and covariance matrix $\Sigma_\beta \in \mathbb{R}^{d\times d}$.

The next step is to determine how we should update this initial assumption about β after we observed some new data represented by the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. Ultimately, the goal is to determine the **posterior distribution** of β given the new data, i.e. to determine $p(\beta|Y = y)$, where y is the vector of observations and Y is the random vector that we assumed to have generated these observations in the probit model.

The way to do this is to make use of the bayes rule:

$$p(\beta|Y = y) = \frac{p(Y = y|\beta)p(\beta)}{p(Y = y)}. \tag{16}$$

This relationship tells us the different parts that we have to combine in order to arrive at the posterior distribution $p(\beta|Y = y)$.

The first part is the likelihood function $p(Y = y|\beta)$, which we already dealt with in section 2.3 and the second component is the prior distribution $p(\beta)$, that we assumed to be normal.

The most challenging part to compute is the quantity $p(Y = y)$. We can see why it is so challenging by writing it out:

$$\begin{aligned} p(Y = y) &= \int p(Y = y|\beta)p(\beta)d\beta \\ &= \int \prod_{i=1}^{n} \Phi(-z_i\beta)\frac{1}{\sqrt{(2\pi)^d \det \Sigma_\beta}} \exp\left(-\frac{1}{2}(\beta - \mu_\beta)^T\Sigma_\beta^{-1}(\beta - \mu_\beta)\right)d\beta \end{aligned} \tag{17}$$

This infinite integral over all possible values of β is impossible to solve analytically, which means that it's impossible to exactly compute the posterior distribution for the probit model. But luckily, there are workarounds to this problem.

The first consideration is, that we could also analyze the posterior distribution if we had a large enough sample available instead. When the sample size is big enough, the Glivenko-Cantelli theorem tells us that the empirical posterior distribution converges to the true posterior distribution [Vaart, 1998]. This allows us to analyze the posterior distribution by analyzing a large enough sample of it, but the problem of how to obtain such a sample still remains.

In practice, instead of directly sampling from $p(\beta|Y = y)$, the posterior distribution can be approximated by so called Markov chain Monte Carlo (MCMC) methods. One such method that works particularly well for the probit model is the Gibbs sampler, which is described in the next section.

2.4.1 Gibbs Sampling in the Probit Model

Gibbs sampling is an iterative tool for drawing samples from probability distributions, that was first applied in the context of bayesian inference by [Gelfand and Smith, 1990] and has been adapted to the probit model by [Albert and Chib, 1993] using the idea of *data augmentation*, which was first introduced in [Tanner and Wong, 1987]. We describe this idea in the following section, as it yields an efficient algorithm for sampling from the posterior distribution of the probit model.

Remember that the probit model has two components: The vector of latent variables Y^* that follows a linear model $Y^* \mid \beta \sim \mathcal{N}(X\beta, 1)$, where we assume that $\sigma = 1$ for reasons of identifiability (see section 2.1) and the random vector Y that produces the observed outcomes y by thresholding: If $Y_i^* > 0$, then $Y_i = 1$ and $Y_i = 0$ otherwise. We also assumed a normal prior distribution: $\beta \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$.

Now, imagine that we knew the outcomes of the latent variable vector Y^*. The conditional distribution of β given the realization y^* of the latent variables can be shown to be normal [Albert and Chib, 1993]:

$$\beta \mid Y^* = y^* \;\sim\; \mathcal{N}(b, B), \tag{18}$$

where $b = (\Sigma_\beta^{-1} + X^TX)^{-1}(\Sigma_\beta^{-1}\mu_\beta + X^Ty^*)$ and $B = (\Sigma_\beta^{-1} + X^TX)^{-1}$. As we can see, the result is a normal distribution from which we can sample efficiently.

The problem is, that in reality we can't observe the latent variables and therefore we don't know the realizations y^*. Here, an important finding by [Albert and Chib, 1993] comes into play: If we could observe β and see the realization $\tilde{\beta}$, then we could determine the conditional distribution of the latent variable vector Y^*:

$$Y_i^* \mid \beta = \tilde{\beta},\ Y_i = y_i \ \sim \begin{cases} \mathcal{N}(x_i^T\tilde{\beta}, 1) \text{ truncated at the left by } 0, & \text{if } y_i = 1 \\ \mathcal{N}(x_i^T\tilde{\beta}, 1) \text{ truncated at the right by } 0, & \text{if } y_i = 0 \end{cases} \tag{19}$$

This means, that given a realization $\tilde{\beta}$ and the observed values in y, the latent variables follow a truncated normal distribution, from which it is also possible to sample efficiently.

These two observations bring us directly to the Gibbs sampling algorithm for the probit model. The first thing, that this algorithm does is to start at some value $\tilde{\beta}^{(0)}$. [Albert and Chib, 1993] suggest that this could for example be the maximum likelihood estimate, that we already discussed in section 2.3.

The next step of the Gibbs sampling algorithm is to use this value $\tilde{\beta}^{(0)}$ to sample a realization $y^{*(1)}$ from the latent variable vector Y^*, by using the conditional distribution in equation 19. Given $y^{*(1)}$, it is then possible to sample a new value $\tilde{\beta}^{(1)}$ from the normal distribution in equation 18, which starts a new cycle. These two sampling steps, which can both be carried out efficiently, are repeated until the desired amount of samples is reached.

To sum up, we augmented the observed data by incorporating the hidden variables Y^* to arrive at a two-stage procedure that draws alternating samples from the conditional distributions of β and Y^*, hence the name data augmentation.

3 Coresets and Sensitivity Sampling

Content.

4 Data Streams

Content.

5 Experiments

Content.

6 Concluding Remarks

Content.

7 Notes

7.1 VC Dimension

An alternative approach is to write down the VC dimension by using an instance space and a concept class as given in [Kearns and Vazirani, 1994].

Lemma 1. *Let $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d \times \mathbb{R}_{>0}$ be the instance space consisting of n points with their last coordinate being positive. The concept class of interest, $\mathcal{C}$ over X, is given as follows:*

$$\mathcal{C} = \left\{\{x \in X : \ f_{\beta,r}(x) \geq 0\} \mid \beta \in \mathbb{R}^d, r \geq 0\right\},$$

with

$$f_{\beta,r}(x) = x_{d+1} \cdot g\left(\sum_{i=1}^{d} x_i \beta_i\right) - r$$

and

$$g(x) = -\log \Phi(-x).$$

The VC dimension of $\mathcal{C}$ is equal to the VC dimension of the range space induced by $\mathcal{F}^w_{probit} = \{w_i g(z_i \beta) \mid i \in [n]\}$, $Z \in \mathbb{R}^{n \times d}$, $w \in \mathbb{R}^n_{>0}$.

There a few different strategies that can be used to find an upper bound on the VC dimension of $\mathcal{C}$, as shown by the following lemmas. The first one is a simple upper bound for finite concept classes:

Lemma 2. *Let X be an instance space and $\mathcal{C}$ be a concept class over X. If the cardinality of $\mathcal{C}$ can be bounded by m, i.e. $|\mathcal{C}| \leq m$, then $VCdim(\mathcal{C}) \leq log(m)$.*

The next lemma partitions the concept class into smaller classes, for each of which the VC dimension can be bounded:

Lemma 3. *Let X be an instance space and $\mathcal{C}$ be a concept class over X. Let $\mathcal{C}_1, ..., \mathcal{C}_k$ be a partition of $\mathcal{C}$ into k disjoint subsets, i.e. $\mathcal{C} = \bigcup_{i=1}^k \mathcal{C}_i$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset \ \forall i \neq j$. Then, $VCdim(\mathcal{C}) \leq \sum_{i=1}^k VCdim(\mathcal{C}_i)$.*

Proof. For the sake of contradiction, assume there was a set $S \subseteq X$ of size $|S| > \sum_{i=1}^k VCdim(\mathcal{C}_i)$ that is shattered by $\mathcal{C}$. If S is shattered by C, every subset of S must also be shattered by C. Consider the intersections $T_i = \bigcup_{c \in \mathcal{C}_i} S \cap c$. Every T_i is a subset of S and $S = \bigcup_{i=1}^k T_i$. Since S is shattered by $\mathcal{C}$, every T_i must be shattered by $\mathcal{C}_i$. We assumed that $|S| > \sum_{i=1}^k VCdim(\mathcal{C}_i)$. It follows that there exists a T_j with $T_j > VCdim(\mathcal{C}_j)$. Since T_j is also shattered by $\mathcal{C}_j$, this is a contradiction, which concludes the proof. □

A result in [Linial et al., 1991] suggests an even smaller upper bound:

Lemma 4 ([Linial et al., 1991])**.** *Let X be an instance space and $\mathcal{C}$ be a concept class over X. Let $\mathcal{C} = \bigcup_{i=1}^k \mathcal{C}_i$ and $VCdim(\mathcal{C}_i) \leq m$. If k is bounded by a polynomial function of m, then $VCdim(\mathcal{C}) \leq 3m$.*

Instead of partitioning the concept class, we could also partition the instance space and obtain a similar bound:

Lemma 5. *Let X be an instance space and $\mathcal{C}$ be a concept class over X. Let $X_1, ..., X_k$ be a partition of X into k disjoint subsets, i.e. $X = \bigcup_{i=1}^{k} X_i$ and $X_i \cap X_j = \emptyset \ \forall i \neq j$. Let $\mathcal{C}_i = \{X_i \cap c \mid c \in \mathcal{C}\}$ be a concept class over X_i for all $i \in [k]$.*
Then, $VCdim(\mathcal{C}) \leq \sum_{i=1}^{k} VCdim(\mathcal{C}_i)$.

Proof. Again, assume there existed a set $S \subseteq X$ of size $|S| > \sum_{i=1}^{k} VCdim(\mathcal{C}_i)$ that is shattered by $\mathcal{C}$. S can be partitioned into disjoined subsets $T_i = S \cap X_i$, with $\bigcup_{i=1}^{k} T_i = S$. Every T_i must be shattered by $\mathcal{C}_i$. Since we assumed that $|S| > \sum_{i=1}^{k} VCdim(\mathcal{C}_i)$, there exists a T_j with $|T_j| > VCdim(C_j)$ which is also shattered by C_j. This contradiction concludes the proof. □

7.2 New idea for VC dimension proof

Lemma 6. *Let*

$$h_{\beta,r}(x) = \begin{cases} 1 \text{ if } & x_{d+1} \cdot g\left(\sum_{i=1}^{d} x_i \beta_i\right) - r \geq 0 \\ 0 \text{ else} \end{cases}$$

Be a function from $\mathbb{R}^{d+1}$ to $\{0,1\}$ with parameters $\beta \in \mathbb{R}^d$ and $r \in \mathbb{R}_{\geq 0}$ with

$$g(x) = \log\left(\frac{1}{1-\Phi(x)}\right),$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}z^2} dz \ .$$

Let

$$H = \left\{x \mapsto h_{\beta,r}(x) \mid \beta \in \mathbb{R}^d, \ r \in \mathbb{R}_{\geq 0}\right\}$$

be the hypothesis class determined by h. Then, the VC dimension of H is . . .

Proof. Let $S = \sum_{i=1}^{d} x_i \beta_i$. We show that h can be computed in t steps as follows:

$$\begin{aligned} & x_{d+1} \cdot g(S) - r \geq 0 \\ \iff & \log\left(\frac{1}{1-\Phi(S)}\right) \geq \frac{r}{x_{d+1}} \\ \iff & \frac{1}{1-\Phi(S)} \geq \exp\left(\frac{r}{x_{d+1}}\right) \\ \iff & 1-\Phi(S) \leq \exp\left(-\frac{r}{x_{d+1}}\right) \\ \iff & \Phi(S) \geq 1-\exp\left(-\frac{r}{x_{d+1}}\right) \\ \iff & \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{S} e^{-\frac{1}{2}z^2} dz \geq 1-\exp\left(-\frac{r}{x_{d+1}}\right) \\ \iff & \int_{-\infty}^{S} e^{-\frac{1}{2}z^2} dz \geq \sqrt{2\pi}\left(1-\exp\left(-\frac{r}{x_{d+1}}\right)\right) \end{aligned}$$

□

7.3 Online Leverage Scores

The leverage scores of a matrix $A \in \mathbb{R}^{n\times d}$ are given by $l_i = a_i^T(A^TA)^{-1}a_i$ [Cohen et al., 2020]. According to [Cohen et al., 2020], we can obtain overestimates of these scores by using only a subset of the rows in A to compute them.

Let A_j be a matrix that contains only the first j rows of A. It follows that the the estimated leverage score $\tilde{l}_j = a_j^T(A_j^TA_j)^{-1}a_j$ is an overestimate of l_j. In a recent paper by [Chhaya et al., 2020], it was shown that the sum of these overestimates can be bounded regardless of how the rows in A are ordered:

Lemma 7 ([Chhaya et al., 2020]).

$$\sum_{i=1}^{n} \tilde{l}_j \in O(d + d\log\|A\| - \min_{i\in[n]}\|a_i\|)$$

Next, we show how a simple algorithm that computes $\tilde{l}_j$ in an online manner (passing row by row over the data stream) can be constructed requiring only $\mathcal{O}(d^2)$ of working memory. The idea is to only keep the matrix $A_j^TA_j \in \mathbb{R}^{d\times d}$ in memory and update it for every new row a_{j+1} using a rank one update $A_{j+1}^TA_{j+1} = A_j^TA_j + a_{j+1}\cdot a_{j+1}^T$. See [Golub and van Loan, 2013] for more on matrix multiplication using outer products. The algorithm is given in algorithm 1.

```
Algorithm 1: Online Leverage Scores
  Input: Matrix $A \in \mathbb{R}^{n\times d}$
  Output: Online leverage scores $\tilde{l}_i$ for all $i \in [n]$
Initialize $M_0 = 0^{d\times d}$
foreach $a_i := i$'th row vector of $A$, $a_i \in \mathbb{R}^d$ do
    $M_i = M_{i-1} + a_i \cdot a_i^T$
    $\tilde{l}_i = a_i^T M_i^\dagger a_i$
return $\tilde{l}_i,\ i \in [n]$
```

7.4 Probit Regression

Situation: We have n data points $(x_i, y_i),\ i = 1, ..., n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$.

Probit Model: y_i is a realization of the random variable Y_i. $Y_1, ..., Y_n$ are independent. The distribution of Y_i is as follows:

$$P(Y_i = 1|x_i;\beta) = \Phi(x_i^T\beta)$$
$$P(Y_i = -1|x_i;\beta) = 1 - \Phi(x_i^T\beta) = \Phi(-x_i^T\beta)$$

where $\beta \in \mathbb{R}^d$. It follows that

$$P(Y_i = y_i|x_i;\beta) = \Phi(y_ix_i^T\beta)$$

Likelihood: The likelihood of a parameter vector β is given as follows:

$$L(\beta) = \prod_{i=1}^{n} P(Y_i = y_i | x_i; \beta) = \prod_{i=1}^{n} \Phi(y_i x_i^T \beta)$$

The negative log-likelihood that we wish to minimize is:

$$\mathcal{L}(\beta) = -\sum_{i=1}^{n} \log \Phi(y_i x_i^T \beta)$$

The weighted case: We introduce sample weights $w_i \in \mathbb{R}_{>0}$ comprising a weight vector $w \in \mathbb{R}^n_{>0}$. Further, let $g(z) = -\log \Phi(-z)$. The objective function now becomes:

$$f_w(\beta) = \sum_{i=1}^{n} w_i g(-y_i x_i^T \beta)$$

To make the notation easier, we define $z_i = -y_i x_i^T$ and introduce the matrix $Z \in \mathbb{R}^{n \times d}$ with row vectors $Z_i = z_i$. This gives us:

$$f_w(\beta) = \sum_{i=1}^{n} w_i g(z_i \beta)$$

Gradient: The gradient of the objective function is needed during optimization. To derive it, we first need the derivative of $g(z)$:

$$g'(z) = \frac{d}{dz} - \log \Phi(-z) = \frac{\phi(z)}{\Phi(-z)}$$

Now we can calculate the gradient of the objective function as follows:

$$\frac{\partial f_w(\beta)}{\partial \beta} = \sum_{i=1}^{n} w_i \frac{\partial g(z_i \beta)}{\partial \beta} = \sum_{i=1}^{n} w_i z_i g'(z_i \beta)$$

Lemma 8. *Let $g(z) = -\log \Phi(-z)$. Then it holds for all $z \geq 0$ that:*

$$\frac{1}{2} z^2 \leq g(z)$$

Proof. The following relationship holds for all $z \geq 1$:

$$\begin{aligned}
\Phi(-z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-z} \exp\left(-\frac{1}{2}x^2\right) dx \\
&\leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-z} -x \exp\left(-\frac{1}{2}x^2\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \\
&\leq \exp\left(-\frac{1}{2}z^2\right)
\end{aligned}$$

We therefore have for $z \geq 1$:

$$e^{g(z)} = e^{-\log \Phi(-z)} = \frac{1}{\Phi(-z)} \geq e^{\frac{1}{2}z^2}$$

Since $\exp(\cdot)$ is a monotonically increasing function, it follows that $g(z) \geq \frac{1}{2}z^2$ for all $z \geq 1$.
Let us now turn to the case when $0 \leq z \leq 1$. Both $g(z)$ and $\frac{1}{2}z^2$ are monotonically increasing and continuous functions for $0 \leq z \leq 1$. Together with the fact that $g(0) > \frac{1}{2}$ it follows for all $0 \leq z \leq 1$ that

$$g(z) \geq g(0) > \frac{1}{2} = \max_{0 \leq z \leq 1} \frac{1}{2}z^2 \geq \frac{1}{2}z^2$$

which concludes the proof. □

Lemma 9. *Let $g(z) = -\log \Phi(-z)$. Then it holds for all $z \geq 2$ that:*

$$g(z) \leq z^2$$

Proof. We first show that $\Phi(-z) \geq \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-\frac{1}{2}z^2}$ for all $z \geq 0$. In order to prove this lower bound, we define $h(z) = \Phi(-z) - \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-\frac{1}{2}z^2}$ and show that $h(z)$ is positive for all $z \geq 0$. The derivative $h'(z) = -\sqrt{\frac{2}{\pi}} \frac{e^{-\frac{1}{2}z^2}}{(z^2+1)^2}$ is negative for all z, so $h(z)$ is a monotonically decreasing function. Also, it clearly holds that $h(0) > 0$ and $\lim_{z \to \infty} h(z) = 0$. It follows that $h(z) \geq 0$ for all $z > 0$ which proves the lower bound.

In the next step, we use this result to show that $e^{z^2} \cdot \Phi(-z) \geq 1$ for all $z \geq 2$:

$$\begin{aligned} e^{z^2} \cdot \Phi(-z) &\geq e^{z^2} \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-\frac{1}{2}z^2} \\ &= e^{\frac{1}{2}z^2} \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} \\ &= e^{\frac{1}{2}z^2} \frac{1}{\frac{4}{3}(z^2+1)} \frac{\frac{4}{3}z}{\sqrt{2\pi}} \\ &\geq \frac{e^{\frac{1}{2}z^2}}{\frac{4}{3}(z^2+1)} \\ &\geq \frac{e^{\frac{1}{2}z^2}}{e^{\frac{1}{2}z^2}} \\ &= 1 \end{aligned}$$

From this it follows directly that $\frac{1}{\Phi(-z)} \leq e^{z^2}$ and thus we have for all $z \geq 2$:

$$e^{g(z)} = e^{-\log \Phi(-z)} = \frac{1}{\Phi(-z)} \leq e^{z^2}$$

Since $\exp(\cdot)$ is monotonically increasing, the claim that $g(z) \leq z^2$ for all $z \geq 2$ follows as a direct consequence.

The ideas for these proofs are based on the work in [Gordon, 1941]. □

7.5 Coresets and Sensitivity Sampling

Definition 3. *Let $X \in \mathbb{R}^{n\times d}$, $y \in \{-1,1\}^n$ be an instance of probit regression with sample weights $w \in \mathbb{R}^n_{>0}$ and let $z_i = -y_i x_i^T$, $i = 1, ..., n$. Then $C \in \mathbb{R}^{k\times d}$ weighted by $u \in \mathbb{R}^k_{>0}$ is a $(1 \pm \epsilon)$-coreset of X, y for probit regression if*

$$(1-\epsilon) f_{w,Z}(\beta) \leq f_{u,C}(\beta) \leq (1+\epsilon) f_{w,Z}(\beta) \quad \forall \beta \in \mathbb{R}^d,$$

where $f_{w,Z}(\beta) = \sum_{i=1}^n w_i g(z_i\beta)$, $f_{u,C}(\beta) = \sum_{i=1}^k u_i g(c_i\beta)$ and $g(z) = -\log \Phi(-z)$.

7.6 Lower Bounds

Theorem 2. *Let $X \in \mathbb{R}^{n\times 2}$, $y \in \{-1,1\}^n$ be an instance of probit regression. Any coreset $C \in \mathbb{R}^{k\times 2}$ of X, y for probit regression consists of at least $k \in \Omega\left(\frac{n}{\log n}\right)$ points.*

Proof. We first show how such a coreset could be used in a communication protocol for the INDEX communication game to encode a message. Since there exists a lower bound on the minimum message length of the INDEX game (see [Kremer et al., 1999]), we can use it to derive a lower bound on the coreset size. The same technique was also used in [Munteanu et al., 2018] to find lower bounds for coresets of logistic regression and is here slightly adapted for probit regression.

The INDEX game consists of two players, Alice and Bob. Alice is given a random binary string $x \in \{0,1\}^n$ of n bits and Bob is given an index $i \in [n]$. The goal is for Alice to send a message to Bob that allows Bob to obtain the value x_i of Alice's binary string x. It was shown in [Kremer et al., 1999], that the minimum length of a message sent by Alice that still allows Bob to obtain x_i with constant probability is in $\Omega(n)$ bits. We will now see how a coreset for probit regression can be used to encode such a message.

The first step is for Alice to convert her binary string x into a set P of two-dimensional points as follows: For each entry x_j of her binary string where $x_j = 1$, she adds a point $p_j = \left(\cos\left(2\pi\frac{j}{n}\right), \sin\left(2\pi\frac{j}{n}\right)\right)$ to her set P and labels it with 1. As we can see, all of these points are on the unit circle and all of them are labeled with 1. Next, she uses these points to construct a coreset for probit regression $C \in \mathbb{R}^{k\times 2}$ of P and sends it to Bob. We will later see, how large the size k of this coreset must be, so that Bob can still obtain x_i with constant probability.

As soon as Alice's coreset C arrives at Bob, Bob can use it to obtain the value of x_i. To do this, Bob first adds two new points $q_1 = \left(\cos\left(2\pi\frac{i-0.5}{n}\right), \sin\left(2\pi\frac{i-0.5}{n}\right)\right)$ and $q_2 = \left(\cos\left(2\pi\frac{i+0.5}{n}\right), \sin\left(2\pi\frac{i+0.5}{n}\right)\right)$ to the set and labels both points with -1 (see figure 1). Next, he uses his points q_1 and q_2 together with the coreset C to obtain a solution for the corresponding probit regression problem. He can then use the value of the cost function to determine the value of x_i like this:

Since Alice only added a point p_j to her set if $x_j = 1$, both of his points q_1 and q_2 are linearly seperable from Alice's points if the value of $x_i = 0$, i.e. Alice didn't add a point for x_i. In this case, the value of the cost function tends to zero. If, on the other hand, Bob's new points q_1 and q_2 can't be linearly seperated from the other points, it means that Alice added a point for $x_i = 1$. In this case, there must be at least

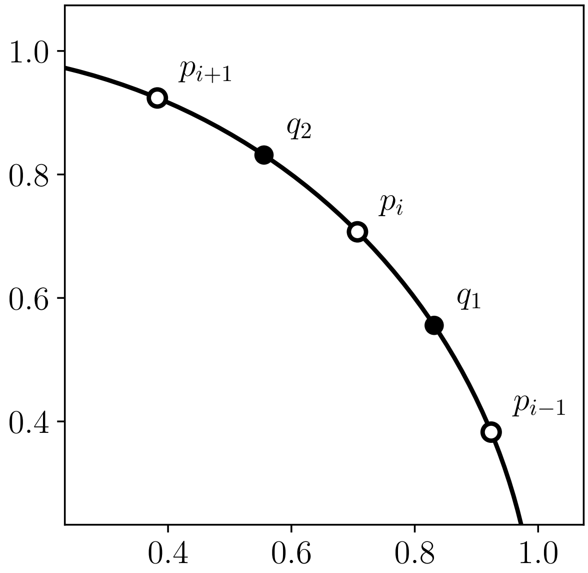

Figure 1: Bob places two points q_1 and q_2 in such a way on the unit circle, that they can be linearly seperated from the other points if and only if Alice didn't place a point at p_i.

one misclassification and the value of the cost function is at least $g(0) = \log(2)$. Since coresets can be used to obtain $(1+\epsilon)$-approximation of the objective function, Bob can use this case distinction to determine the value of x_i.

There is one special case that has to be dealt with in order for this protocol to work. If Alice's coreset only consists of the single point p_i, Bob's points q_1 and q_2 could still be linearly seperated although Alice added p_i. The workaround to this is simple though: Bob can always just add two more points at the locations of p_{i-1} and p_{i+1} and label them with 1. Now, q_1 and q_2 can only be linearly seperated from the other points if and only if Alice didn't add a point p_i.

Let us now see how large the size k of Alice's coreset must be for this protocol to work with constant probability. In [Kremer et al., 1999] it was shown, that the minimum length of a message that Alice must send is in $\Omega(n)$ bits. Since each of the points that Alice created can be encoded in $\log(n)$ space, it follows from the lower bound that $\Omega(n) \subseteq \Omega(k \log(n))$, so k must be in $\Omega\left(\frac{n}{\log(n)}\right)$.

We can conclude that if there existed a $(1+\epsilon)$-coreset for probit regression with size $k \in o\left(\frac{n}{\log(n)}\right)$, it would contradict the minimum message length of INDEX, which proves the claim. □

7.7 Sensitivity Sampling

Definition 4. *Let $Z \in \mathbb{R}^{n\times d}$. Then we define*

$$\mu_w(Z) = \sup_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{\left\|(\sqrt{D_w}Z\beta)^+\right\|_2^2}{\left\|(\sqrt{D_w}Z\beta)^-\right\|_2^2} = \sup_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{\left\|(\sqrt{D_w}Z\beta)^-\right\|_2^2}{\left\|(\sqrt{D_w}Z\beta)^+\right\|_2^2}$$

Z weighted by w is called μ-complex if $\mu_w(Z) \le \mu$.

Definition 5 ([Feldman et al., 2020, Langberg and Schulman, 2010])**.** *Let $F = \{g_1, ..., g_n\}$ be a set of functions, $g_i : \mathbb{R} \to \mathbb{R}^{\ge 0}$, $i = 1, ..., n$ weighted by $w \in \mathbb{R}^n_{>0}$. The sensitivity of g_i for $f_w(\beta) = \sum_{i=1}^n w_i g_i(\beta)$ is defined as*

$$\varsigma_i = \sup_{\beta\in\mathbb{R}^d,\ f_w(\beta)>0} \frac{w_i g_i(\beta)}{f_w(\beta)}.$$

The total sensitivity, i.e. the sum of the sensitivities is $\mathfrak{S} = \sum_{i=1}^n \varsigma_i$.

Definition 6 ([Feldman et al., 2020])**.** *A range space is a pair $\mathfrak{R} = (F, \mathcal{R})$, where F is a set and $\mathcal{R}$ is a family (set) of subsets of F, called ranges.*

Definition 7 ([Feldman et al., 2020])**.** *The VC-dimension $\Delta(\mathfrak{R})$ of a range space $\mathfrak{R} = (F, \mathcal{R})$ is the size $|G|$ of the largest subset $G \subseteq F$ such that*

$$|\{G \cap \mathrm{range} \mid \mathrm{range} \in \mathcal{R}\}| = 2^{|G|},$$

i.e. G is shattered by $\mathcal{R}$.

Definition 8 ([Feldman et al., 2020])**.** *Let F be a finite set of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}^{\ge 0}$. For every $\beta \in \mathbb{R}^d$ and $r \ge 0$, let*

$$\mathrm{range}(F, \beta, r) = \{f \in F \mid f(\beta) \ge r\}$$

and let

$$\mathcal{R}(F) = \left\{\mathrm{range}(F, \beta, r) \mid \beta \in \mathbb{R}^d,\ r \ge 0\right\}.$$

Then we call $\mathfrak{R}_F := (F, \mathcal{R}(F))$ the range space induced by F.

Theorem 3 ([Feldman et al., 2020, Munteanu et al., 2018])**.** *Let $\mathcal{F} = \{f_1, ..., f_n\}$ be a set of functions, $f_i : \mathbb{R} \to \mathbb{R}^{\ge 0}$, $i = 1, ..., n$ weighted by $w \in \mathbb{R}^n_{>0}$. Let $\epsilon, \delta \in (0, \frac{1}{2})$. Let $s_i \ge \varsigma_i$. Let $S = \sum_{i=1}^n s_i \ge \mathfrak{S}$. Given s_i, one can compute in time $O(|\mathcal{F}|)$ a set $\mathcal{R} \subseteq \mathcal{F}$ of*

$$O\left(\frac{S}{\epsilon^2}\left(\Delta \log S + \log\left(\frac{1}{\delta}\right)\right)\right)$$

weighted functions such that with probability $1 - \delta$ we have for all $\beta \in \mathbb{R}^d$ simultaneously

$$\left|\sum_{f\in\mathcal{F}} w_i f_i(\beta) - \sum_{f\in\mathcal{R}} u_i f_i(\beta)\right| \le \epsilon \sum_{f\in\mathcal{F}} w_i f_i(\beta)$$

where each element of $\mathcal{R}$ is sampled independently with probability $p_j = \frac{s_j}{S}$ from $\mathcal{F}$, $u_i = \frac{Sw_j}{s_j|\mathcal{R}|}$ denotes the weight of a function $f_i \in \mathcal{R}$ that corresponds to $f_j \in \mathcal{F}$, and where Δ is an upper bound on the VC-dimension of the range space $\mathfrak{R}_{\mathcal{F}^*}$ induced by $\mathcal{F}^*$. $\mathcal{F}^*$ is the set of functions $f_j \in \mathcal{F}$ scaled by $\frac{Sw_j}{s_j|\mathcal{R}|}$.

Lemma 10. *Let $Z \in \mathbb{R}^{n\times d}$, $c \in \mathbb{R}_{>0}$. The range space induced by*

$$\mathcal{F}^c_{probit} = \{cg(z_i\beta) \mid i \in [n]\}$$

satisfies $\Delta(\mathfrak{R}_{\mathcal{F}^c_{probit}}) \leq d+1$.

Proof. For all $G \subseteq \mathcal{F}^c_{probit}$ we have

$$\left|\left\{G \cap \text{range} \mid \text{range} \in \mathcal{R}(\mathcal{F}^c_{probit})\right\}\right| = \left|\left\{\text{range}(G,\beta,r) \mid \beta \in \mathbb{R}^d,\ r \geq 0\right\}\right|.$$

Since g is invertible and monotone, we have for all $\beta \in \mathbb{R}^d$ and $r \geq 0$ that

$$\begin{aligned}\text{range}(G,\beta,r) &= \{g_i \in G \mid g_i(\beta) \geq r\}\\ &= \{g_i \in G \mid cg(x_i\beta) \geq r\}\\ &= \left\{g_i \in G \mid x_i\beta \geq g^{-1}\left(\frac{r}{c}\right)\right\}.\end{aligned}$$

Note, that $\left\{g_i \in G \mid x_i\beta \geq g^{-1}\left(\frac{r}{c}\right)\right\}$ corresponds to the positively classified points of the affine hyperplane classifier $x \mapsto \text{sign}\left(x\beta - g^{-1}\left(\frac{r}{c}\right)\right)$. We thus have for all $G \subseteq \mathcal{F}^c_{probit}$, that

$$\left|\left\{G \cap \text{range} \mid \text{range} \in \mathcal{R}(\mathcal{F}^c_{probit})\right\}\right| = \left|\left\{\{g_i \in G \mid x_i\beta - s \geq 0\} \ \mid \beta \in \mathbb{R}^d,\ s \in \mathbb{R}\right\}\right|.$$

Since the VC dimension of the set of affine hyperplane classifiers is $d+1$, it follows that $\Delta(\mathfrak{R}_{\mathcal{F}^c_{probit}}) \leq d+1$, which concludes our proof. □

Lemma 11. *Let $Z \in \mathbb{R}^{n\times d}$ be weighted by $w \in \mathbb{R}^n_{>0}$ where $w_i \in \{v_1, ..., v_t\}$ for all $i \in [n]$. The range space induced by*

$$\mathcal{F}_{probit} = \{w_i g(z_i\beta) \mid i \in [n]\}$$

satisfies $\Delta(\mathfrak{R}_{\mathcal{F}_{probit}}) \leq t \cdot (d+1)$.

Proof. We partition the functions of $\mathcal{F}_{probit}$ into t disjoint classes

$$F_j = \{w_i g(z_i\beta) \in \mathcal{F}_{probit} \mid w_i = v_j\}, \quad j \in [t].$$

The functions in each of these classes have an equal weight, wich means that by lemma 10, each of their induced range spaces has a VC-dimension of at most $d+1$.

For the sake of contradiction, assume that $\Delta(\mathfrak{R}_{\mathcal{F}_{probit}}) > t\cdot(d+1)$ and let G be the corresponding set of size $|G| > t\cdot(d+1)$ that is shattered by $\mathcal{R}(\mathcal{F}_{probit})$. Since the sets F_j are disjoint, each intersection $F_j \cap G$ must be shattered by $\mathcal{R}(F_j)$. Further, at least one of the intersections must have at minimum $\frac{|G|}{t}$ elements, which means that for at least one $j \in [t]$ it holds that $|F_j \cap G| \geq \frac{|G|}{t} > \frac{t\cdot(d+1)}{t} = d+1$. This is a contradiction to lemma 10, which concludes the proof. □

Lemma 12. *Let $Z \in \mathbb{R}^{n\times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ be μ-complex. Let U be an orthonormal basis for the columnspace of $\sqrt{D_w}Z$. If for index i, the supreme β in definition 5 satisfies $2 \le z_i\beta$, then $w_i g(z_i\beta) \le 2\|U_i\|_2^2(1+\mu)f_w(\beta)$.*

Proof. Let $\sqrt{D_w}Z = UR$, where U is an orthonormal basis for the columnspace of $\sqrt{D_w}Z$. It follows from $2 \le z_i\beta$ and from the monotonicity of g that

$$\begin{aligned}
w_i g(z_i\beta) &= w_i g\left(\frac{\sqrt{w_i}z_i\beta}{\sqrt{w_i}}\right) = w_i g\left(\frac{U_iR\beta}{\sqrt{w_i}}\right) \le w_i g\left(\frac{\|U_i\|_2\|R\beta\|_2}{\sqrt{w_i}}\right)\\
&= w_i g\left(\frac{\|U_i\|_2\|UR\beta\|_2}{\sqrt{w_i}}\right) = w_i g\left(\frac{\|U_i\|_2\|\sqrt{D_w}Z\beta\|_2}{\sqrt{w_i}}\right)\\
&\le \|U_i\|_2^2\|\sqrt{D_w}Z\beta\|_2^2 \le \|U_i\|_2^2(1+\mu)\|(\sqrt{D_w}Z\beta)^+\|_2^2\\
&= \|U_i\|_2^2(1+\mu)\sum_{j:\ \sqrt{w_j}z_j\beta\ge 0} w_j(z_j\beta)^2\\
&\le 2\|U_i\|_2^2(1+\mu)\sum_{j:\ \sqrt{w_j}z_j\beta\ge 0} w_j g(z_j\beta)\\
&\le 2\|U_i\|_2^2(1+\mu)\sum_{j=1}^{n} w_j g(z_j\beta)\\
&= 2\|U_i\|_2^2(1+\mu)f_w(\beta)
\end{aligned}$$

□

Lemma 13. *Let $Z \in \mathbb{R}^{n\times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ be μ-complex. If for index i, the supreme β in definition 5 satisfies $z_i\beta \le 2$, then $w_i g(z_i\beta) \le \frac{w_i}{\mathcal{W}}(80+16\mu)f_w(\beta)$.*

Proof. Let $K^- = \{j \in [n] \mid z_j\beta \le -1\}$ and $K^+ = \{j \in [n] \mid z_j\beta > -1\}$. Note that $g(-1) > \frac{1}{10}$ and $g(z_i\beta) \le g(2) < 4$. Also, $\sum_{j\in K^+} w_j + \sum_{j\in K^-} w_j = \mathcal{W}$.
Thus, if $\sum_{j\in K^+} w_j \ge \frac{1}{2}\mathcal{W}$ then

$$f_w(\beta) = \sum_{j=1}^{n} w_j g(z_j\beta) \ge \sum_{j\in K^+} w_j g(z_j\beta) \ge \frac{\sum_{j\in K^+} w_j}{10} \ge \frac{\mathcal{W}}{20} = \frac{\mathcal{W}}{20w_i}w_i \ge \frac{\mathcal{W}}{80w_i}w_i g(z_i\beta)$$

If on the other hand $\sum_{j \in K^+} w_j < \frac{1}{2}\mathcal{W}$, then $\sum_{j \in K^-} w_j \geq \frac{1}{2}\mathcal{W}$. Thus

$$\begin{aligned}
f_w(\beta) &= \sum_{j=1}^{n} w_j g(z_j\beta) \geq \sum_{j:\ z_j\beta > 0} w_j g(z_j\beta) \geq \frac{1}{2} \sum_{j:\ z_j\beta>0} w_j (z_j\beta)^2 \\
&= \frac{1}{2}\|(\sqrt{D_w}Z\beta)^+\|_2^2 \geq \frac{1}{2\mu}\|(\sqrt{D_w}Z\beta)^-\|_2^2 \\
&= \frac{1}{2\mu} \sum_{j:\ z_j\beta<0} w_j(z_j\beta)^2 \\
&\geq \frac{1}{2\mu} \sum_{j \in K^-} w_j(z_j\beta)^2 \\
&\geq \frac{1}{2\mu} \sum_{j \in K^-} w_j \\
&\geq \frac{\mathcal{W}}{4\mu} \\
&\geq \frac{\mathcal{W}}{16\mu w_i} w_i g(z_i\beta)
\end{aligned}$$

Adding both bounds, we get that for $z_i\beta \leq 2$:

$$w_i g(z_i\beta) \leq f_w(\beta)\frac{80 w_i}{\mathcal{W}} + f_w(\beta)\frac{16\mu w_i}{\mathcal{W}} = \frac{w_i}{\mathcal{W}}(80 + 16\mu) f_w(\beta)$$

□

Lemma 14. *Let $Z \in \mathbb{R}^{n \times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ be μ-complex. Let U be an orthonormal basis for the columnspace of $\sqrt{D_w}Z$. For each $i \in [n]$, the sensitivity of $g_i(\beta) = g(z_i\beta)$ is bounded by $\varsigma_i \leq s_i = (80 + 16\mu)(\|U_i\|_2^2 + \frac{w_i}{\mathcal{W}})$. The total sensitivity is bounded by $\mathfrak{S} \leq 192\mu d$.*

Proof.

$$\begin{aligned}
\varsigma_i = \sup_{\beta} \frac{w_i g(z_i\beta)}{f_w(\beta)} &\leq \sup_{\beta} \frac{2\|U_i\|_2^2(1+\mu)f_w(\beta) + \frac{w_i}{\mathcal{W}}(80 + 16\mu)f_w(\beta)}{f_w(\beta)} \\
&= 2\|U_i\|_2^2(1+\mu) + \frac{w_i}{\mathcal{W}}(80 + 16\mu) \\
&\leq \|U_i\|_2^2(80 + 16\mu) + \frac{w_i}{\mathcal{W}}(80 + 16\mu) \\
&= (80 + 16\mu)(\|U_i\|_2^2 + \frac{w_i}{\mathcal{W}})
\end{aligned}$$

$$
\begin{aligned}
\mathfrak{S} = \sum_{i=1}^{n} \varsigma_i &\leq (80+16\mu)\sum_{i=1}^{n} \|U_i\|_2^2 + \frac{w_i}{\mathcal{W}} \\
&= (80+16\mu)(\|U\|_F^2 + 1) \\
&= (80+16\mu)(d+1) \\
&\leq 96\mu(d+1) \\
&\leq 192\mu d
\end{aligned}
$$

□

Lemma 15. *Let $U \in \mathbb{R}^{n\times d}$ be an orthonormal matrix. Then $\|U\|_F^2 = d$.*

Proof.

$$
\begin{aligned}
\|U\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{d} |u_{ij}|^2 \\
&= \sum_{j=1}^{d}\sum_{i=1}^{n} |u_{ij}|^2 \\
&\overset{(1)}{=} \sum_{j=1}^{d} 1 \\
&= d
\end{aligned}
$$

(1) follows from the fact that the columns of U have unit norm due to its orthonormality.

□

References

[Agresti, 2015] Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* Wiley.

[Albert and Chib, 1993] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

[Chhaya et al., 2020] Chhaya, R., Choudhari, J., Dasgupta, A., and Shit, S. (2020). Streaming coresets for symmetric tensor factorization. *CoRR*, abs/2006.01225.

[Cohen et al., 2020] Cohen, M. B., Musco, C., and Pachocki, J. (2020). Online row sampling. *Theory of Computing*, 16(15):1–25.

[Demidenko, 2001] Demidenko, E. (2001). Computational aspects of probit model. *Mathematical Communications*, 6:233–247.

[Fahrmeir et al., 2013] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2013). *Regression.* Springer-Verlag Berlin Heidelberg.

[Feldman et al., 2020] Feldman, D., Schmidt, M., and Sohler, C. (2020). Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657.

[Gelfand and Smith, 1990] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

[Gelman et al., 2013] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis.* Chapman and Hall/CRC, 3 edition.

[Golub and van Loan, 2013] Golub, G. H. and van Loan, C. F. (2013). *Matrix Computations.* JHU Press, fourth edition.

[Gordon, 1941] Gordon, R. D. (1941). Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366.

[Kearns and Vazirani, 1994] Kearns, M. J. and Vazirani, U. V. (1994). *An Introduction to Computational Learning Theory.* MIT Press.

[Kremer et al., 1999] Kremer, I., Nisan, N., and Ron, D. (1999). On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49.

[Langberg and Schulman, 2010] Langberg, M. and Schulman, L. J. (2010). Universal ϵ-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 598–607.

[Lesaffre and Kaufmann, 1992] Lesaffre, E. and Kaufmann, H. (1992). Existence and uniqueness of the maximum likelihood estimator for a multivariate probit model. *Journal of the American Statistical Association*, 87:805–811.

[Linial et al., 1991] Linial, N., Mansour, Y., and Rivest, R. L. (1991). Results on learnability and the vapnik-chervonenkis dimension. *Information and Computation*, 90(1):33–49.

[McCullagh and Nelder, 1989] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC.

[Munteanu et al., 2018] Munteanu, A., Schwiegelshohn, C., Sohler, C., and Woodruff, D. P. (2018). On coresets for logistic regression. In *Advances in Neural Information Processing Systems 31, (NeurIPS)*, pages 6562–6571.

[Nocedal and Wright, 2006] Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer-Verlag New York, 2 edition.

[Tanner and Wong, 1987] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

[Vaart, 1998] Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[Wedderburn, 1976] Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(1):27–32.