1 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

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

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 $w \in \mathbb{R}_{>0}^n$. 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 1. *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 2. *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 [2]. □

2 Coresets

Definition 1. *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 $g(z) = -\log \Phi(-z)$.

2.1 Lower Bounds

Theorem 1. *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 [3]), we can use it to derive a lower bound on the coreset size. The same technique was also used in [5] 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 [3], 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 $p_j = \left(\cos\left(\frac{j}{n}\right), \sin\left(\frac{j}{n}\right)\right)$ 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 a new point $p_i = (1-\delta)\left(\cos\left(\frac{j}{n}\right), \sin\left(\frac{j}{n}\right)\right)$ for some small $\delta > 0$ to the set and labels it with -1. Next, he uses his point p_i 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 $x_j = 1$, his new point p_i is 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 point p_i 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 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.

Let us now see how large the size k of Alice's coreset must be for this protocol to work with constant probability. In [3] it was shown, that the minimum length of a message that Alice can 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. $\square$

3 Sensitivity Sampling

Definition 2. *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 $\mu_w(Z) \leq \mu$.

Definition 3 ([1, 4])**.** *Let $F = \{g_1, ..., g_n\}$ be a set of functions, $g_i : \mathbb{R} \to \mathbb{R}^{\geq 0}$, $i = 1, ..., n$ 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 4 ([1])**.** *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 5 ([1])**.** *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*

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

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

Definition 6 ([1])**.** *Let F be a finite set of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}^{\geq 0}$. For every $\beta \in \mathbb{R}^d$ and $r \geq 0$, let*

$$\text{range}(F, \beta, r) = \{f \in F \mid f(\beta) \geq r\}$$

and let

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

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

Theorem 2 ([1, 5])**.** *Let $\mathcal{F} = \{f_1, ..., f_n\}$ be a set of functions, $f_i : \mathbb{R} \to \mathbb{R}^{\geq 0}$, $i = 1, ..., n$ weighted by $w \in \mathbb{R}^n_{>0}$. Let $\epsilon, \delta \in (0, \frac{1}{2})$. Let $s_i \geq \varsigma_i$. Let $S = \sum_{i=1}^n s_i \geq \mathfrak{S}$. 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

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

where each element of $\mathcal{R}$ is sampled independently with probability $p_j = \frac{s_j}{S}$ from $\mathcal{F}$, $u_i = \frac{S w_j}{s_j |\mathcal{R}|}$ 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 $\frac{S w_j}{s_j |\mathcal{R}|}$.

Lemma 3. *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 $\Delta(\mathfrak{R}_{\mathcal{F}^c_{probit}}) \le d+1$.

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

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

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

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

Note, that $\left\{g_i \in G \mid x_i\beta \ge g^{-1}\left(\frac{r}{c}\right)\right\}$ 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

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

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

Lemma 4. *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 $\Delta(\mathfrak{R}_{\mathcal{F}_{probit}}) \le t\cdot(d+1)$.

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 3, 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 $|F_j \cap G| \ge \frac{|G|}{t} > \frac{t\cdot(d+1)}{t} = d+1$. This is a contradiction to lemma 3, which concludes the proof. □

Lemma 5. *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 3 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 $2 \leq z_i\beta$ and from the monotonicity of g that

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

□

Lemma 6. *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 3 satisfies $z_i\beta \leq 2$, then $w_i g(z_i\beta) \leq \frac{w_i}{\mathcal{W}}(80 + 16\mu)f_w(\beta)$.*

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

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

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 7. *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 8. *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.

□

4 Notes

TODO: Find an upper bound for the VC dimension of the range space induced by a set of weighted functions independent of the weights.

Lemma 9. *Let $Z \in \mathbb{R}^{n \times d}$, $w \in \mathbb{R}^n_{>0}$. The range space induced by*

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

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

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

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

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

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

□

References

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

[2] R. D. Gordon. 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, 1941.

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

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

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