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

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 $w \in \mathbb{R}^n_{>0}$. 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 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 $z \geq 1$:

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

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 $0 \leq z \leq 1$. For $z = 0$ we have $g(0) > \frac{1}{2} > \frac{1}{2}0^2 = 0$ and for $z = 1$ we have $g(1) > 1 > \frac{1}{2}1^2 = \frac{1}{2}$. Since both $g(z)$ and $\frac{1}{2}z^2$ are continuous and monotonically increasing functions for $0 \leq z \leq 1$, it follows that $g(z) \geq \frac{1}{2}z^2$ for all $0 \leq z \leq 1$. □

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. [1] $\square$

Definition 1. *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 $\mu_w(Z) \leq \mu$.

Lemma 3. *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 (TODO) satisfies $2 \leq z_i\beta$, then $w_i g(z_i\beta) \leq 2\|U_i\|_2^2(1+\mu) f_w(\beta)$.*

Proof. Let $\sqrt{D_w}Z = UR$, where U is an orthonormal basis for the columnspace of

[1]The proofs of both lemmas were heavily inspired by `https://www.johndcook.com/blog/norm-dist-bounds/`.

$\sqrt{D_w}Z$. It follows from $2 \leq z_i\beta$ and from the monotonicity of g that

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

□

Lemma 4. *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 (TODO) 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 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 $\sqrt{D_w}Z$. 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 6. *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