# QUALIFYING EXAM: MATH 281A

PROFESSOR JELENA BRADIC

Consider the following suppervised classification problem, where we aim to classify a vector $X \in \mathbb{R}^d$ as belonging to class $Y = 0$ or $Y = 1$. We have data drawn i.i.d. from a standard multivariate normal model, where

$$X|Y = 0 \sim \mathcal{N}\left(-\frac{1}{2}\theta_0, \mathbb{I}_{d\times d}\right) \text{ and } X|Y = 1 \sim \mathcal{N}\left(\frac{1}{2}\theta_0, \mathbb{I}_{d\times d}\right). \tag{1}$$

In the above $\mathbb{I}_{d\times d}$ denotes a matrix whose diagonal elements are all 1 and the rest are zero. We assume that $\theta_0 \in \mathbb{R}^d$ is the unknown mean vector for each of the two classes and we also assume

$$P(Y = 0) = P(Y = 1) = 1/2$$

or in other terms that the two classes are well balanced.

Assume that you are given an i.i.d. sample $(X_i, Y_i)$, $i = 1, \cdots, n$ from the model (**??**) and that your would like to estimate $\theta_0$.

(a) To estimate $\theta_0$ a natural strategy is to utilize the Gaussianity (normality) of the model (1) and define an estimate as

$$\hat{\theta}_1 = N_1^{-1} \sum_{i:Y_i=1} X_i - N_0^{-1} \sum_{i:Y_i=0} X_i. \tag{2}$$

In the above $N_1$ and $N_0$ denote the number of samples in class 1 and 0, respectively. Show that

$$\sqrt{n}(\hat{\theta}_1 - \theta_0) \to \mathcal{N}(0, \Sigma_1)$$

in distribution and provide (specify) the matrix $\Sigma_1$.

(b) Another approach to estimating $\theta_0$ is to consider a logistic regression model, defining the logistic loss for a pair $(x, y)$ by

$$l(\theta; x, y) = \log(1 + e^{x^\top \theta}) - yx^\top \theta.$$

Define the empirical risk estimator

$$\hat{R}_n(\theta) = n^{-1} \sum_{i=1}^{n} l(\theta; X_i, Y_i)$$

*Date*: May 17th.

and

$$\hat{\theta}_2 = \arg\min_{\theta} \hat{R}_n(\theta).$$

Show that

$$\sqrt{n}(\hat{\theta}_2 - \theta_0) \to \mathcal{N}(0, \Sigma_2)$$

in distribution and specify $\Sigma_2$.

(c) Now consider the actual classification risk $R(\theta) = E[l(\theta; X, Y)]$. Show that

$$n(R(\hat{\theta}_1) - R(\theta_0)) \to W_1 \text{ and } n(R(\hat{\theta}_2) - R(\theta_0)) \to W_2$$

in distribution, for some random variables $W_1, W_2$ and specify their distribution.

(d) Which of the two estimators is more efficient ?